

Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, Volkan Cevher

▶ **To cite this version:**

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. Applied and Computational Harmonic Analysis, 2018. hal-01958890

HAL Id: hal-01958890
https://inria.hal.science/hal-01958890

Submitted on 19 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Optimal Rates for Spectral Algorithms with Least-Squares Regression over Hilbert Spaces

Junhong Lin[1], Alessandro Rudi[2,3], Lorenzo Rosasco[4,5], Volkan Cevher[1]

[1]Laboratory for Information and Inference Systems, École Polytechnique Fédérale de Lausanne, CH1015-Lausanne, Switzerland.

[2] INRIA, Sierra project-team, 75012 Paris, France.

[3] Département Informatique - École Normale Supérieure, Paris, France.

[4] University of Genova, 16146 Genova, Italy.

[5] LCSL, Massachusetts Institute of Technology and Italian Institute of Technology.

December 18, 2018

Abstract

In this paper, we study regression problems over a separable Hilbert space with the square loss, covering non-parametric regression over a reproducing kernel Hilbert space. We investigate a class of spectral/regularized algorithms, including ridge regression, principal component regression, and gradient methods. We prove optimal, high-probability convergence results in terms of variants of norms for the studied algorithms, considering a capacity assumption on the hypothesis space and a general source condition on the target function. Consequently, we obtain almost sure convergence results with optimal rates. Our results improve and generalize previous results, filling a theoretical gap for the non-attainable cases.
Keywords Learning theory, Reproducing kernel Hilbert space, Sampling operator, Regularization scheme, Regression.

1 Introduction

Let the input space H be a separable Hilbert space with inner product denoted by $\langle \cdot, \cdot \rangle_H$ and the output space $\mathbb{R}$. Let ρ be an unknown probability measure on $H \times \mathbb{R}$, $\rho_X(\cdot)$ the induced marginal measure on H, and $\rho(\cdot|x)$ the conditional probability measure on $\mathbb{R}$ with respect to $x \in H$ and ρ. Let the hypothesis space $H_\rho = \{f : H \to \mathbb{R} | \exists \omega \in H \text{ with } f(x) = \langle \omega, x \rangle_H, \rho_X\text{-almost surely}\}$. The goal of least-squares regression is to approximately solve the following expected risk minimization,

$$\inf_{f \in H_\rho} \mathcal{E}(f), \quad \mathcal{E}(f) = \int_{H \times \mathbb{R}} (f(x) - y)^2 d\rho(x, y), \tag{1}$$

where the measure ρ is known only through a sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^n$ of size $n \in \mathbb{N}$, independently and identically distributed according to ρ. Let $L^2_{\rho_X}$ be the Hilbert space of square integral functions from H to $\mathbb{R}$ with respect to ρ_X, with its norm given by $\|f\|_\rho = \left(\int_H |f(x)|^2 d\rho_X\right)^{1/2}$. The function that minimizes the expected risk over all measurable functions is the regression function [6, 27], defined as

$$f_\rho(x) = \int_{\mathbb{R}} y d\rho(y|x), \qquad x \in H, \rho_X\text{-almost every}. \tag{2}$$

Throughout this paper, we assume that the support of ρ_X is compact and there exists a constant $\kappa \in [1, \infty[$, such that

$$\langle x, x' \rangle_H \leq \kappa^2, \quad \forall x, x' \in H, \rho_X\text{-almost every.} \tag{3}$$

Under this assumption, H_ρ is a subspace of $L^2_{\rho_X}$, and a solution f_H for (1) is the projection of the regression function $f_\rho(x)$ onto the closure of H_ρ in $L^2_{\rho_X}$. See e.g., [14, 1], or Section 2 for further details.

The above problem was raised for non-parametric regression with kernel methods [6, 27] and it is closely related to functional regression [20]. A common and classic approach for the above problem is based on spectral algorithms. It amounts to solving an empirical linear equation, where to avoid over-fitting and to ensure good performance, a filter function for regularization is involved, see e.g., [1, 10]. Such approaches include ridge regression, principal component regression, gradient methods and iterated ridge regression.

A large amount of research has been carried out for spectral algorithms within the setting of learning with kernel methods, see e.g., [26, 5] for Tikhonov regularization, [33, 31] for gradient methods, and [4, 1] for general spectral algorithms. Statistical results have been developed in these references, but still, they are not satisfactory. For example, most of the previous results either restrict to the case that the space H_ρ is universal consistency (i.e., H_ρ is dense in $L^2_{\rho_X}$) [26, 31, 4] or the attainable case (i.e., $f_H \in H_\rho$) [5, 1]. Also, some of these results require an unnatural assumption that the sample size is large enough and the derived convergence rates tend to be (capacity-dependently) suboptimal in the non-attainable cases. Finally, it is still unclear whether one can derive capacity-dependently optimal convergence rates for spectral algorithms under a general source assumption.

In this paper, we study statistical results for spectral algorithms. Considering a capacity assumption of the space H [32, 5], and a general source condition [1] of the target function f_H, we show high-probability, optimal convergence results in terms of variants of norms for spectral algorithms. As a corollary, we obtain almost sure convergence results with optimal rates. The general source condition is used to characterize the regularity/smoothness of the target function f_H in $L^2_{\rho_X}$, rather than in H_ρ as those in [5, 1]. The derived convergence rates are optimal in a minimax sense. Our results, not only resolve the issues mentioned in the last paragraph but also generalize previous results to convergence results with different norms and consider a more general source condition.

2 Learning with Kernel Methods and Notations

In this section, we first introduce supervised learning with kernel methods, which is a special instance of the learning setting considered in this paper. We then introduce some useful notations and auxiliary operators.

Learning with Kernel Methods. Let Ξ be a closed subset of Euclidean space $\mathbb{R}^d$. Let μ be an unknown but fixed Borel probability measure on $\Xi \times Y$. Assume that $\{(\xi_i, y_i)\}_{i=1}^n$ are i.i.d. from the distribution μ. A reproducing kernel K is a symmetric function $K : \Xi \times \Xi \to \mathbb{R}$ such that $(K(u_i, u_j))_{i,j=1}^\ell$ is positive semidefinite for any finite set of points $\{u_i\}_{i=1}^\ell$ in Ξ. The kernel K defines a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ as the completion of the linear span of the set $\{K_\xi(\cdot) := K(\xi, \cdot) : \xi \in \Xi\}$ with respect to the inner product

$\langle K_\xi, K_u\rangle_K := K(\xi, u)$. For any $f \in \mathcal{H}_K$, the reproducing property holds: $f(\xi) = \langle K_\xi, f\rangle_K$. In learning with kernel methods, one considers the following minimization problem

$$\inf_{f\in\mathcal{H}_K} \int_{\Xi\times\mathbb{R}} (f(\xi) - y)^2 d\mu(\xi, y).$$

Since $f(\xi) = \langle K_\xi, f\rangle_K$ by the reproducing property, the above can be rewritten as

$$\inf_{f\in\mathcal{H}_K} \int_{\Xi\times\mathbb{R}} (\langle f, K_\xi\rangle_K - y)^2 d\mu(\xi, y).$$

Defining another probability measure $\rho(K_\xi, y) = \mu(\xi, y)$, the above reduces to (1).

Notations and Auxiliary Operators. We next introduce some notations and auxiliary operators which will be useful in the following. For a given bounded operator $L : L^2_{\rho_X} \to H$, $\|L\|$ denotes the operator norm of L, i.e., $\|L\| = \sup_{f\in L^2_{\rho_X}, \|f\|_\rho=1} \|Lf\|_H$.

Let $\mathcal{S}_\rho : H \to L^2_{\rho_X}$ be the linear map $\omega \to \langle\omega, \cdot\rangle_H$, which is bounded by κ under Assumption (3). Furthermore, we consider the adjoint operator $\mathcal{S}^*_\rho : L^2_{\rho_X} \to H$, the covariance operator $\mathcal{T} : H \to H$ given by $\mathcal{T} = \mathcal{S}^*_\rho\mathcal{S}_\rho$, and the operator $\mathcal{L} : L^2_{\rho_X} \to L^2_{\rho_X}$ given by $\mathcal{S}_\rho\mathcal{S}^*_\rho$. It can be easily proved that $\mathcal{S}^*_\rho g = \int_H xg(x)d\rho_X(x)$, $\mathcal{L}f = \int_H f(x)\langle x, \cdot\rangle_H d\rho_X(x)$ and $\mathcal{T} = \int_H \langle\cdot, x\rangle_H x d\rho_X(x)$. Under Assumption (3), the operators $\mathcal{T}$ and $\mathcal{L}$ can be proved to be positive trace class operators (and hence compact):

$$\|\mathcal{L}\| = \|\mathcal{T}\| \le \mathrm{tr}(\mathcal{T}) = \int_H \mathrm{tr}(x\otimes x)d\rho_X(x) = \int_H \|x\|^2_H d\rho_X(x) \le \kappa^2. \tag{4}$$

For any $\omega \in H$, it is easy to prove the following isometry property [27]

$$\|\mathcal{S}_\rho\omega\|_\rho = \|\sqrt{\mathcal{T}}\omega\|_H. \tag{5}$$

Moreover, according to the spectral theorem,

$$\|\mathcal{L}^{-\frac{1}{2}}\mathcal{S}_\rho\omega\|_\rho \le \|\omega\|_H \tag{6}$$

We define the sampling operator $\mathcal{S}_\mathbf{x} : H \to \mathbb{R}^n$ by $(\mathcal{S}_\mathbf{x}\omega)_i = \langle\omega, x_i\rangle_H$, $i \in [n]$, where the norm $\|\cdot\|_{\mathbb{R}^n}$ in $\mathbb{R}^n$ is the Euclidean norm times $1/\sqrt{n}$. Its adjoint operator $\mathcal{S}^*_\mathbf{x} : \mathbb{R}^n \to H$, defined by $\langle\mathcal{S}^*_\mathbf{x}\mathbf{y}, \omega\rangle_H = \langle\mathbf{y}, \mathcal{S}_\mathbf{x}\omega\rangle_{\mathbb{R}^n}$ for $\mathbf{y} \in \mathbb{R}^n$ is thus given by $\mathcal{S}^*_\mathbf{x}\mathbf{y} = \frac{1}{n}\sum_{i=1}^n y_i x_i$. Moreover, we can define the empirical covariance operator $\mathcal{T}_\mathbf{x} : H \to H$ such that $\mathcal{T}_\mathbf{x} = \mathcal{S}^*_\mathbf{x}\mathcal{S}_\mathbf{x}$. Obviously,

$$\mathcal{T}_\mathbf{x} = \frac{1}{n}\sum_{i=1}^n \langle\cdot, x_i\rangle_H x_i.$$

By Assumption (3), similar to (4), we have

$$\|\mathcal{T}_\mathbf{x}\| \le \mathrm{tr}(\mathcal{T}_\mathbf{x}) \le \kappa^2. \tag{7}$$

A simple calculation shows that [6, 27] for all $f \in L^2_{\rho_X}$,

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_\rho.$$

Then it is easy to see that (1) is equivalent to $\inf_{f\in H_\rho} \|f - f_\rho\|^2_\rho$. Using the projection theorem, one can prove that a solution f_H for the above problem is the projection of the regression function f_ρ onto the closure of H_ρ in $L^2_{\rho_X}$, and moreover, for all $f \in H_\rho$, (see e.g., [14]),

$$\mathcal{S}^*_\rho f_\rho = \mathcal{S}^*_\rho f_H, \tag{8}$$

and

$$\mathcal{E}(f) - \mathcal{E}(f_H) = \|f - f_H\|^2_\rho. \tag{9}$$

3 Spectral/Regularized Algorithms

In this section, we demonstrate and introduce spectral algorithms.

The search for an approximate solution in H_ρ for Problem (1) is equivalent to the search of an approximated solution in H for

$$\inf_{\omega\in H}\widetilde{\mathcal{E}}(\omega),\quad \widetilde{\mathcal{E}}(\omega)=\int_{H\times\mathbb{R}}(\langle\omega,x\rangle_H-y)^2d\rho(x,y). \tag{10}$$

As the expected risk $\widetilde{\mathcal{E}}(\omega)$ can not be computed exactly and that it can be only approximated through the empirical risk $\widetilde{\mathcal{E}}_{\mathbf{z}}(\omega)$, defined as

$$\widetilde{\mathcal{E}}_{\mathbf{z}}(\omega)=\frac{1}{n}\sum_{i=1}^{n}(\langle\omega,x_i\rangle_H-y_i)^2,$$

a first idea to deal with the problem is to replace the objective function in (10) with the empirical risk, which leads to an estimator $\hat{\omega}$ satisfying the empirical, linear equation

$$\mathcal{T}_{\mathbf{x}}\hat{\omega}=\mathcal{S}_{\mathbf{x}}^*\mathbf{y}.$$

However, solving the empirical, linear equation directly may lead to a solution that fits the sample points very well but has a large expected risk. This is called as overfitting phenomenon in statistical learning theory. Moreover, the inverse of the empirical covariance operator $\mathcal{T}_{\mathbf{x}}$ does not exist in general. To tackle with this issue, a common approach in statistical learning theory and inverse problems, is to replace $\mathcal{T}_{\mathbf{x}}^{-1}$ with an alternative, regularized one, which leads to spectral algorithms [8, 4, 1].

A spectral algorithm is generated by a specific choice of filter function. Recall that the definition of filter functions is given as follows.

Definition 3.1 (Filter functions)**.** *Let Λ be a subset of $\mathbb{R}_+$. A class of functions $\{\mathcal{G}_\lambda:[0,\kappa^2]\to[0,\infty[,\lambda\in\Lambda\}$ is said to be filter functions with qualification τ ($\tau\geq 1$) if there exist some positive constants $E,F_\tau<\infty$ such that*

$$\sup_{\alpha\in[0,1]}\sup_{\lambda\in\Lambda}\sup_{u\in]0,\kappa^2]}|u^\alpha\mathcal{G}_\lambda(u)|\lambda^{1-\alpha}\leq E. \tag{11}$$

and

$$\sup_{\alpha\in[0,\tau]}\sup_{\lambda\in\Lambda}\sup_{u\in]0,\kappa^2]}|(1-\mathcal{G}_\lambda(u)u)|u^\alpha\lambda^{-\alpha}\leq F_\tau. \tag{12}$$

Given a filter function $\mathcal{G}_\lambda$, the spectral algorithm is defined as follows.

Algorithm 1. *Let $\mathcal{G}_\lambda$ be a filter function indexed with $\lambda>0$. The spectral algorithm over the samples $\mathbf{z}$ is given by*[1]

$$\omega_\lambda^{\mathbf{z}}=\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{S}_{\mathbf{x}}^*\mathbf{y}, \tag{13}$$

and

$$f_\lambda^{\mathbf{z}}=\mathcal{S}_\rho\omega_\lambda^{\mathbf{z}}. \tag{14}$$

[1]Let L be a self-adjoint, compact operator over a separable Hilbert space H. $\mathcal{G}_\lambda(L)$ is an operator on L defined by spectral calculus: suppose that $\{(\sigma_i,\psi_i)\}_i$ is a set of normalized eigenpairs of L with the eigenfunctions $\{\psi_i\}_i$ forming an orthonormal basis of H, then $\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})=\sum_i\mathcal{G}_\lambda(\sigma_i)\psi_i\otimes\psi_i$.

Different filter functions correspond to different regularization algorithms. The following examples provide several specific choices on filter functions, which leads to different types of regularization methods, see e.g. [10, 1, 26].

Example 3.1 (Spectral cut-off)**.** *Consider the spectral cut-off or truncated singular value decomposition (TSVD) defined by*

$$\mathcal{G}_\lambda(u) = \begin{cases} u^{-1}, & \text{if } u \geq \lambda, \\ 0, & \text{if } u < \lambda. \end{cases}$$

Then the qualification τ could be any positive number and $E = F_\tau = 1$.

Example 3.2 (Gradient methods)**.** *The choice $\mathcal{G}_\lambda(u) = \sum_{k=1}^{t} \eta(1-\eta u)^{t-k}$ with $\eta \in]0, \kappa^2]$ where we identify $\lambda = (\eta t)^{-1}$, corresponds to gradient methods or Landweber iteration algorithm. The qualification τ could be any positive number, $E = 1$, and $F_\tau = (\tau/\mathrm{e})^\tau$.*

Example 3.3 ((Iterated) ridge regression)**.** *Let $l \in \mathbb{N}$. Consider the function*

$$\mathcal{G}_\lambda(u) = \sum_{i=1}^{l} \lambda^{i-1}(\lambda + u)^{-i} = \frac{1}{u}\left(1 - \frac{\lambda^l}{(\lambda+u)^l}\right).$$

It is easy to show that the qualification $\tau = l$, $E = l$ and $F_\tau = 1$. In the case that $l = 1$, the algorithm is ridge regression.

The performance of spectral algorithms can be measured in terms of the excess risk, $\mathcal{E}(f_\lambda^{\mathbf{z}}) - \inf_{H_\rho} \mathcal{E}$, which is exactly $\|f_\lambda^{\mathbf{z}} - f_H\|_\rho^2$ according to (9). Assuming that $f_H \in H_\rho$, which implies that there exists some ω_* such that $f_H = \mathcal{S}_\rho \omega_*$ (in this case, the solution with minimal H-norm for $f_H = \mathcal{S}_\rho \omega$ is denoted by ω_H), it can be measured in terms of H-norm, $\|\omega_\lambda^{\mathbf{z}} - \omega_H\|_H$, which is closely related to $\|\mathcal{L}^{-\frac{1}{2}}\mathcal{S}_\rho(\omega_\lambda^{\mathbf{z}} - \omega_H)\|_H = \|\mathcal{L}^{-\frac{1}{2}}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho$ according to (6). In what follows, we will measure the performance of spectral algorithms in terms of a broader class of norms, $\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho$, where $a \in [0, \frac{1}{2}]$ is such that $\mathcal{L}^{-a} f_H$ is well defined. Throughout this paper, we assume that $1/n \leq \lambda \leq 1$.

4 Convergence Results

In this section, we first introduce some basic assumptions and then present convergence results for spectral algorithms.

4.1 Assumptions

The first assumption relates to a moment condition on the output value y.

Assumption 1. *There exists positive constants Q and M such that for all $l \geq 2$ with $l \in \mathbb{N}$,*

$$\int_{\mathbb{R}} |y|^l d\rho(y|x) \leq \frac{1}{2} l! M^{l-2} Q^2, \tag{15}$$

ρ_X-almost surely.

The above assumption is very standard in statistical learning theory. It is satisfied if y is bounded almost surely, or if $y = \langle \omega_*, x \rangle_H + \epsilon$, where ϵ is a Gaussian random variable with zero mean and it is independent from x. Obviously, Assumption 1 implies that the regression function f_ρ is bounded almost surely, as

$$|f_\rho(x)| \leq \int_{\mathbb{R}} |y| d\rho(y|x) \leq \left(\int_{\mathbb{R}} |y|^2 d\rho(y|x) \right)^{\frac{1}{2}} \leq Q. \tag{16}$$

The next assumption relates to the regularity/smoothness of the target function f_H. As $f_H \in \overline{Range(\mathcal{S}_\rho)}$ and $\mathcal{L} = \mathcal{S}_\rho \mathcal{S}_\rho^*$, it is natural to assume a general source condition on f_H as follows.

Assumption 2. *f_H satisfies*

$$\int_H (f_H(x) - f_\rho(x))^2 x \otimes x d\rho_X(x) \preceq B^2 \mathcal{T}, \tag{17}$$

and the following source condition

$$f_H = \phi(\mathcal{L}) g_0, \quad \textit{with} \quad \|g_0\|_\rho \leq R. \tag{18}$$

Here, $B, R \geq 0$ and $\phi : [0, \kappa^2] \to \mathbb{R}^+$ is a non-decreasing index function such that $\phi(0) = 0$ and $\phi(\kappa^2) < \infty$. Moreover, for some $\zeta \in [0, \tau]$, $\phi(u) u^{-\zeta}$ is non-decreasing, and the qualification τ of $\mathcal{G}_\lambda$ covers the index function ϕ.

Recall that the qualification τ of $\mathcal{G}_\lambda$ covers the index function ϕ is defined as follows [1].

Definition 4.1. *We say that the qualification τ covers the index function ϕ if there exists a $c > 0$ such that for all $0 < \lambda \leq \kappa^2$,*

$$c \frac{\lambda^\tau}{\phi(\lambda)} \leq \inf_{\lambda \leq u \leq \kappa^2} \frac{u^\tau}{\phi(u)}. \tag{19}$$

Condition (17) is trivially satisfied if f_H is bounded almost surely. Moreover, when making a consistency assumption, i.e., $\inf_{H_\rho} \mathcal{E} = \mathcal{E}(f_\rho)$, as that in [26, 4, 5, 28], for kernel-based non-parametric regression, it is satisfied with $B = 0$. Condition (18) is a more general source condition that characterizes the "regularity/smoothness" of the target function. It is trivially satisfied with $\phi(u) = 1$ as $f_H \in \overline{H_\rho} \subseteq L^2_{\rho_X}$. In non-parametric regression with kernel methods, one typically considers Hölders condition (corresponding to $\phi(u) = u^\alpha, \alpha \geq 0$) [26, 5, 4] . [1, 18, 21] considers a general source condition but only with an index function $\phi(u)\sqrt{u}$, where ϕ can be decomposed as $\psi\vartheta$ and $\psi : [0, b] \to \mathbb{R}_+$ is operator monotone with $\psi(0) = 0$ and $\psi(b) < \infty$, and $\vartheta : [0, \kappa^2] \to \mathbb{R}_+$ is Lipschitz continuous with $\vartheta(0) = 0$. In the latter case $\inf_{H_\rho} \mathcal{E}$ has a solution f_H in H_ρ as that [27, 22]

$$\mathcal{L}^{\frac{1}{2}}(L^2_{\rho_X}) \subseteq H_\rho, \tag{20}$$

In this paper, we will consider a source assumption with respect to a more general index function, $\phi = \psi\vartheta$, where $\psi : [0, b] \to \mathbb{R}_+$ is operator monotone with $\psi(0) = 0$ and $\psi(b) < \infty$, and $\vartheta : [0, \kappa^2] \to \mathbb{R}_+$ is Lipschitz continuous. Without loss of generality, we assume that the Lipschitz constant of ϑ is 1, as one can always scale both sides of the source condition (18).

Recall that the function ψ is called operator monotone on $[0, b]$, if for any pair of self-adjoint operators U, V with spectra in $[0, b]$ such that $U \preceq V$, $\phi(U) \preceq \phi(V)$.

Finally, the last assumption relates to the capacity of the hypothesis space H_ρ (induced by H).

Assumption 3. *For some $\gamma \in]0, 1]$ and $c_\gamma > 0$, $\mathcal{T}$ satisfies*

$$\mathrm{tr}(\mathcal{T}(\mathcal{T} + \lambda I)^{-1}) \leq c_\gamma \lambda^{-\gamma}, \quad \textit{for all } \lambda > 0. \tag{21}$$

The left hand-side of of (21) is called as the effective dimension [5], or the degrees of freedom [32]. It can be related to covering/entropy number conditions, see [27] for further details. Assumption 3 is always true for $\gamma = 1$ and $c_\gamma = \kappa^2$, since $\mathcal{T}$ is a trace class operator which implies the eigenvalues of $\mathcal{T}$, denoted as σ_i, satisfy $\mathrm{tr}(\mathcal{T}) = \sum_i \sigma_i \leq \kappa^2$. This is referred to as the capacity independent setting. Assumption 3 with $\gamma \in]0, 1]$ allows to derive better rates. It is satisfied, e.g., if the eigenvalues of $\mathcal{T}$ satisfy a polynomial decaying condition $\sigma_i \sim i^{-1/\gamma}$, or with $\gamma = 0$ if $\mathcal{T}$ is finite rank.

4.2 Main Results

Now we are ready to state our main results as follows.

Theorem 4.2. *Under Assumptions 1, 2 and 3, let $a \in [0, \frac{1}{2} \wedge \zeta]$, $\lambda = n^{\theta-1}$ with $\theta \in [0, 1]$, and $\delta \in]0, 1[$. The followings hold with probability at least $1 - \delta$.*
1) If $\phi : [0, b] \to \mathbb{R}_+$ is operator monotone with $b > \kappa^2$, and $\phi(b) < \infty$, or Lipschitz continuous with constant 1 over $[0, \kappa^2]$, then

$$\begin{aligned} &\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \\ \leq &\lambda^{-a} \left(\frac{\tilde{C}_1}{n\lambda^{\frac{1}{2} \vee (1-\zeta)}} + \frac{\tilde{C}_2}{\sqrt{n\lambda^\gamma}} + \tilde{C}_3 \phi(\lambda) \right) \log \frac{6}{\delta} \left(\log \frac{6}{\delta} + \gamma(\theta^{-1} \wedge \log n) \right)^{1-a}. \end{aligned} \tag{22}$$

2) If $\phi = \psi\vartheta$, where $\psi : [0, b] \to \mathbb{R}_+$ is operator monotone with $b > \kappa^2$, $\psi(0) = 0$ and $\psi(b) < \infty$, and $\vartheta : [0, \kappa^2] \to \mathbb{R}_+$ is non-decreasing, Lipschitz continuous with constant 1 and $\vartheta(0) = 0$. Furthermore, assume that the quality of $\mathcal{G}_\lambda$ covers $\vartheta(u)u^{\frac{1}{2}-a}$, then

$$\begin{aligned} \|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq &\lambda^{-a} \log \frac{6}{\delta} \left(\log \frac{6}{\delta} + \gamma(\theta^{-1} \wedge \log n) \right)^{1-a} \\ &\times \left(\frac{\tilde{C}_1}{n\lambda^{\frac{1}{2} \vee (1-\zeta)}} + \frac{\tilde{C}_4}{\sqrt{n\lambda^\gamma}} + \tilde{C}_5 \phi(\lambda) + \tilde{C}_6 \vartheta(\lambda)\psi(n^{-\frac{1}{2}}) \right), \end{aligned} \tag{23}$$

Here, $\tilde{C}_1, \tilde{C}_2, \cdots, \tilde{C}_6$ are positive constants depending only on $\kappa^2, c_\gamma, \gamma, \zeta, \phi, \tau$ B, M, Q, R, E, F_τ, b, a, c and $\|\mathcal{T}\|$ (independent from λ, n, δ, and θ, and given explicitly in the proof).

The above theorem provides convergence results with respect to variants of norms in high-probability for spectral algorithms. Balancing the different terms in the upper bounds, one has the following results with an optimal, data-dependent choice of regularization parameters. Throughout the rest of this paper, C is denoted as a positive constant that depends only on $\kappa^2, c_\gamma, \gamma, \zeta, \phi, \tau$ B, M, Q, R, E, F_τ, b, a, c and $\|\mathcal{T}\|$, and it could be different at its each appearance.

Corollary 4.3. *Under the assumptions and notations of Theorem 4.2, let* $2\zeta + \gamma > 1$ *and* $\lambda = \Theta^{-1}(n^{-1})$ *where* $\Theta(u) = (\phi(u)/\phi(1))^2 u^\gamma$. *The following holds with probability at least* $1-\delta$.
1) Let ϕ *be as in Part 1) of Theorem 4.2, then*

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq C \frac{\phi(\Theta^{-1}(n^{-1}))}{(\Theta^{-1}(n^{-1}))^a} \log^{2-a} \frac{6}{\delta}. \tag{24}$$

2) Let ϕ *be as in Part 2) of Theorem 4.2 and* $\lambda \geq n^{-\frac{1}{2}}$, *then* (24) *holds.*

The error bounds in the above corollary are optimal as they match the minimax rates from [21] (considering only the case $\zeta \geq 1/2$ and $a = 0$). The assumption that the quality of $\mathcal{G}_\lambda$ covers $\vartheta(u)u^{\frac{1}{2}}$ in Part 2) of Corollary 4.3 is also implicitly required in [1, 18, 21], and it is always satisfied for principle component analysis and gradient methods. The condition $\lambda \geq n^{-1/2}$ will be satisfied in most cases when the index function has a Lipschitz continuous part, and moreover, it is trivially satisfied when $\zeta \geq 1$, as will be seen from the proof.

As a direct corollary of Theorem 4.2, we have the following results considering Hölder source conditions.

Corollary 4.4. *Under the assumptions and notations of Theorem 4.2, we let* $\phi(u) = \kappa^{-2(\zeta-1)_+} u^\zeta$ *in Assumption 2 and* $\lambda = n^{-\frac{1}{1\vee(2\zeta+\gamma)}}$, *then with probability at least* $1-\delta$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq C \begin{cases} n^{-\frac{\zeta-a}{2\zeta+\gamma}} \log^{2-a} \frac{6}{\delta} & \text{if } 2\zeta+\gamma > 1, \\ n^{-(\zeta-a)} \log \frac{6}{\delta} \left(\log \frac{6}{\delta} + \log n^\gamma\right)^{1-a} & \text{if } 2\zeta+\gamma \leq 1. \end{cases} \tag{25}$$

The error bounds in (25) are optimal as the convergence rates match the minimax rates shown in [5, 3] with $\zeta \geq 1/2$. The above result asserts that spectral algorithms with an appropriate regularization parameter converge optimally.

Corollary 4.4 provides convergence results in high-probability for the studied algorithms. It implies convergence in expectation and almost sure convergence shown in the follows. Moreover, when $\zeta \geq 1/2$, it can be translated into convergence results with respect to norms related to H.

Corollary 4.5. *Under the assumptions of Corollary 4.4, the following holds.*
1) For any $q \in \mathbb{N}_+$, *we have*

$$\mathbb{E}\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho^q \leq C \begin{cases} n^{-\frac{q(\zeta-a)}{2\zeta+\gamma}} & \text{if } 2\zeta+\gamma > 1, \\ n^{-q(\zeta-a)} (1 \vee \log n^\gamma)^{q(1-a)} & \text{if } 2\zeta+\gamma \leq 1. \end{cases} \tag{26}$$

2) For any $0 < \epsilon < \zeta - a$,

$$\lim_{n\to\infty} \|\mathcal{L}^{-a}(\mathcal{S}_\rho f_\lambda^{\mathbf{z}} - f_H)\|_\rho n^{\frac{\zeta-a-\epsilon}{1\vee(2\zeta+\gamma)}} = 0, \quad \text{almost surely.}$$

3) If $\zeta \geq 1/2$, *then for some* $\omega_H \in H$, $\mathcal{S}_\rho \omega_H = f_H$ *almost surely, and with probability at least* $1-\delta$,

$$\|\mathcal{T}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \omega_H)\|_H \leq C n^{-\frac{\zeta-a}{2\zeta+\gamma}} \log^{2-a} \frac{6}{\delta}. \tag{27}$$

Remark 4.6. *If* $H = \mathbb{R}^d$, *then Assumption 3 is trivially satisfied with* $c_\gamma = \kappa^2(d \wedge \sigma_{\min}^{-1}), \gamma = 0$, *and Assumption 2 could be satisfied* [2] *with any* $\zeta > 1/2$. *Here* $\sigma_{\min}$ *denotes the smallest*

[2]Note that this is not true in general if H is a general Hilbert space, and the proof for the finite-dimensional cases could be simplified, leading to some smaller constants in the error bounds.

eigenvalue of $\mathcal{T}$. Thus, following from the proof of Theorem 4.2, we have that with probability at least $1-\delta$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq C\sqrt{\frac{c_\gamma}{n}}\log\frac{6}{\delta}\left(\log\frac{6}{\delta}\log c_\gamma\right)^{1-a}.$$

The proof for all the results stated in this subsection are postponed in the next section.

4.3 Discussions

There is a large amount of research on theoretical results for non-parametric regression with kernel methods in the literature, see e.g., [30, 23, 29, 15, 7, 18, 25, 13] and references therein. As noted in Section 2, our results apply to non-parametric regression with kernel methods. In what follows, we will translate some of the results for kernel-based regression into results for regression over a general Hilbert space and compare our results with these results.

We first compare Corollary 4.4 with some of these results in the literature for spectral algorithms with Hölder source conditions. Making a source assumption as

$$f_\rho = \mathcal{L}^\zeta g_0 \quad \text{with } \|g_0\|_\rho \leq R, \tag{28}$$

$1/2 \leq \zeta \leq \tau$, and with $\gamma > 0$, [11] shows that with probability at least $1-\delta$,

$$\|f_\lambda^{\mathbf{z}} - f_\rho\|_\rho \leq Cn^{-\frac{\zeta}{2\zeta+\gamma}}\log^4\frac{1}{\delta}.$$

Condition (28) implies that $f_\rho \in \overline{H_\rho}$ as $H_\rho = range(\mathcal{S}_\rho)$ and $\mathcal{L} = \mathcal{S}_\rho\mathcal{S}_\rho^*$. Thus $f_H = f_\rho$ almost surely.[3] In comparison, Corollary 4.4 is more general. It provides convergence results in terms of different norms for a more general Hölder source condition, allowing $0 < \zeta \leq 1/2$ and $\gamma = 0$. Besides, it does not require the extra assumption $f_H = f_\rho$ and the derived error bound in (25) has a smaller depending order on $\log\frac{1}{\delta}$. For the assumption (28) with $0 \leq \zeta < 1/2$, certain results are derived in [26] for Tikhonov regularization and in [31] for gradient methods, but the rates are suboptimal and capacity-independent. Recently, [13] shows that under the assumption (28), with $\zeta \in]0, \tau]$ and $\gamma \in [0, 1]$, spectral algorithm has the following error bounds in expectation,

$$\mathbb{E}\|f_\lambda^{\mathbf{z}} - f_\rho\|_\rho^2 \leq C\begin{cases} n^{-\frac{2\zeta}{2\zeta+\gamma}} & \text{if } 2\zeta+\gamma > 1, \\ n^{-2\zeta}(1 \vee \log n^\gamma) & \text{if } 2\zeta+\gamma \leq 1. \end{cases}$$

Note also that [7] provides the same optimal error bounds as the above, but only restricts to the cases $1/2 \leq \zeta \leq \tau$ and $n \geq n_0$. In comparison, Corollary 4.4 is more general. It provides convergence results with different norms and it does not require the universal consistency assumption. The derived error bound in (25) is more meaningful as it holds with high probability. However, it has an extra logarithmic factor in the upper bound for the case $2\zeta + \gamma \leq 1$, which is worser than that from [13]. [1, 3] study statistical results for spectral algorithms, under a Hölder source condition, $f_H \in \mathcal{L}^\zeta g_0$ with $1/2 \leq \zeta \leq \tau$. Particularly, [3] shows that if

$$n \geq C\lambda^{-2}\log^2\frac{1}{\delta}, \tag{29}$$

[3] Such a assumption is satisfied if $\inf_{H_\rho}\widetilde{\mathcal{E}} = \widetilde{\mathcal{E}}(f_\rho)$ and it is supported by many function classes and reproducing kernel Hilbert space in learning with kernel methods [27].

then with probability at least $1-\delta$, with $1/2 < \zeta \le \tau$ and $0 \le a \le 1/2$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \le C n^{-\frac{\zeta-a}{2\zeta+\gamma}} \log \frac{6}{\delta}.$$

In comparison, Corollary 4.4 provides optimal convergence rates even in the case that $0 \le \zeta \le 1/2$, while it does not require the extra condition (29). Note that we do not pursue an error bound that depends both on R and the noise level as those in [3, 7], but it should be easy to modify our proof to derive such error bounds (at least in the case that $\zeta \ge 1/2$). The only results by now for the non-attainable cases with a general Hölder condition with respect to f_H (rather than f_ρ) are from [14], where convergence rates of order $O(n^{-\frac{\zeta}{1\vee(2\zeta+\gamma)}} \log^2 n)$ are derived (but only) for gradient methods assuming n is large enough.

We next compare Theorem 4.2 with results from [1, 21] for spectral algorithms considering general source conditions. Assuming that $f_H \in \phi(\mathcal{L})\sqrt{\mathcal{L}} g_0$ with $\|g_0\|_\rho \le R$ (which implies $f_H = \mathcal{S}_\rho \omega_H$ for some $\omega_H \in H$,) where ϕ is as in Part 2) of Theorem 4.2, [1] shows that if the qualification of $\mathcal{G}_\lambda$ covers $\phi(u)\sqrt{u}$ and (29) holds, then with probability at least $1-\delta$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \le C\lambda^{-a}\left(\phi(\lambda)\sqrt{\lambda} + \frac{1}{\sqrt{\lambda n}}\right)\log\frac{6}{\delta}, \quad a = 0, \frac{1}{2}.$$

The error bound is capacity independent, i.e., with $\gamma = 1$. Involving the capacity assumption[4], the error bound is further improved in [21], to

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \le C\lambda^{-a}\left(\phi(\lambda)\sqrt{\lambda} + \frac{1}{\sqrt{n\lambda^\gamma}}\right)\log\frac{6}{\delta}, \quad a = 0, \frac{1}{2}.$$

As noted in [11, Discussion], these results lead to the following estimates in expectation

$$\mathbb{E}\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho^2 \le C\lambda^{-2a}\left(\phi(\lambda)\sqrt{\lambda} + \frac{1}{\sqrt{n\lambda^\gamma}}\right)^2 \log n, \quad a = 0, \frac{1}{2}$$

In comparison with these results, Theorem 4.2 is more general, considering a general source assumption and covering the general case that f_H may not be in H_ρ. Furthermore, it provides convergence results with respect to a broader class of norms, and it does not require the condition (29). Finally, it leads to convergence results in expectation with a better rate (without the logarithmic factor) when the index function is $\phi(u)\sqrt{u}$, and it can infer almost-sure convergence results.

5 Proofs

In this section, we prove the results stated in Section 4. We first give some basic lemmas, and then give the proof of the main results.

5.1 Lemmas

Deterministic Estimates

We first introduce the following lemma, which is a generalization of [1, Proposition 7]. For notational simplicity, we denote

$$\mathcal{R}_\lambda(u) = 1 - \mathcal{G}_\lambda(u)u, \tag{30}$$

[4]Note that from the proof from [21], we can see the results from [21] also require (29).

and

$$\mathcal{N}(\lambda) = \mathrm{tr}(\mathcal{T}(\mathcal{T}+\lambda)^{-1}).$$

Lemma 5.1. *Let $\phi : [0, \kappa^2] \to \mathbb{R}^+$ be a non-decreasing index function and the qualification τ of the filter function $\mathcal{G}_\lambda$ covers the index function ϕ, and for some $\zeta \in [0, \tau]$, $\phi(u)u^{-\zeta}$ is non-decreasing. Then for all $a \in [0, \zeta]$,*

$$\sup_{0<u\le\kappa^2} |\mathcal{R}_\lambda(u)|\phi(u)u^{-a} \le c_g\phi(\lambda)\lambda^{-a}, \quad c_g = \frac{F_\tau}{c \wedge 1}, \tag{31}$$

where c is from Definition 4.1.

Proof. When $\lambda \le u \le \kappa^2$, by (19), we have

$$\frac{\phi(u)}{u^\tau} \le \frac{1}{c}\frac{\phi(\lambda)}{\lambda^\tau}.$$

Thus,

$$|\mathcal{R}_\lambda(u)|\phi(u)u^{-a} = |\mathcal{R}_\lambda(u)|u^{\tau-a}\phi(u)u^{-\tau} \le |\mathcal{R}_\lambda(u)|u^{\tau-a}c^{-1}\phi(\lambda)\lambda^{-\tau} \le F_\tau c^{-1}\lambda^{-a}\phi(\lambda),$$

where for the last inequality, we used (12). When $0 < u \le \lambda$, since $\phi(u)u^{-\zeta}$ is non-decreasing,

$$|\mathcal{R}_\lambda(u)|\phi(u)u^{-a} = |\mathcal{R}_\lambda(u)|u^{\zeta-a}\phi(u)u^{-\zeta} \le |\mathcal{R}_\lambda(u)|u^{\zeta-a}\phi(\lambda)\lambda^{-\zeta} \le F_\tau\phi(\lambda)\lambda^{-a},$$

where we used (12) for the last inequality. From the above analysis, one can finish the proof. □

Using the above lemma, we have the following results for the deterministic vector ω_λ, defined by

$$\omega_\lambda = \mathcal{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^* f_H. \tag{32}$$

Lemma 5.2. *Under Assumption 2, we have for all $a \in [0, \zeta]$,*

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda - f_H)\|_\rho \le c_g R\phi(\lambda)\lambda^{-a}, \tag{33}$$

and

$$\|\omega_\lambda\|_H \le E\phi(\kappa^2)\kappa^{-(2\zeta\wedge 1)}\lambda^{-(\frac{1}{2}-\zeta)_+}. \tag{34}$$

The left hand-side of (33) is often called as the true bias.

Proof. Following from the definition of ω_λ in (32), we have

$$\mathcal{S}_\rho\omega_\lambda - f_H = \mathcal{S}_\rho\mathcal{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^* f_H - f_H = (\mathcal{L}\mathcal{G}_\lambda(\mathcal{L}) - I)f_H.$$

Introducing with (18), with the notation $\mathcal{R}_\lambda(u) = 1 - \mathcal{G}_\lambda(u)u$, we get

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda - f_H)\|_\rho = \|\mathcal{L}^{-a}\mathcal{R}_\lambda(\mathcal{L})\phi(\mathcal{L})g_0\|_\rho \le \|\mathcal{L}^{-a}\mathcal{R}_\lambda(\mathcal{L})\phi(\mathcal{L})\|R.$$

Applying the spectral theorem with (4) and Lemma 5.1 which leads to

$$\|\mathcal{L}^{-a}\mathcal{R}_\lambda(\mathcal{L})\phi(\mathcal{L})\| \le \sup_{u\in[0,\kappa^2]} |\mathcal{R}_\lambda(u)|u^{-a}\phi(u) \le c_g\phi(\lambda)\lambda^{-a},$$

one can get (33).
From the definition of ω_λ in (32) and applying (18), we have

$$\|\omega_\lambda\|_H = \|\mathcal{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^*\phi(\mathcal{L})g_0\|_H \leq \|\mathcal{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^*\phi(\mathcal{L})\|R.$$

According to the spectral theorem, with (4), one has

$$\|\mathcal{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^*\phi(\mathcal{L})\| = \sqrt{\|\phi(\mathcal{L})\mathcal{S}_\rho\mathcal{G}_\lambda(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^*\phi(\mathcal{L})\|} = \|\mathcal{G}_\lambda(\mathcal{L})\mathcal{L}^{\frac{1}{2}}\phi(\mathcal{L})\| \leq \sup_{u\in[0,\kappa^2]}|\mathcal{G}_\lambda(u)u^{\frac{1}{2}}\phi(u)|.$$

Since both $\phi(u)$ and $\phi(u)u^{-\zeta}$ are non-decreasing and non-negative over $[0,\kappa^2]$, thus $\phi(u)u^{-\zeta'}$ is also non-decreasing for any $\zeta' \in [0,\zeta]$. If $\zeta \geq 1/2$, then

$$\sup_{u\in[0,\kappa^2]}|\mathcal{G}_\lambda(u)|u^{\frac{1}{2}}\phi(u) = \sup_{u\in[0,\kappa^2]}|\mathcal{G}_\lambda(u)|u\phi(u)u^{-\frac{1}{2}} \leq E\phi(\kappa^2)\kappa^{-1},$$

where for the last inequality, we used (11) and that $\phi(u)u^{-\frac{1}{2}}$ is non-decreasing. If $\zeta < 1/2$, similarly, we have

$$\sup_{u\in[0,\kappa^2]}|\mathcal{G}_\lambda(u)|u^{\frac{1}{2}}\phi(u) = \sup_{u\in[0,\kappa^2]}|\mathcal{G}_\lambda(u)|u^{\frac{1}{2}+\zeta}\phi(u)u^{-\zeta} \leq E\lambda^{\zeta-\frac{1}{2}}\phi(\kappa^2)\kappa^{-2\zeta}.$$

From the above analysis, one can prove (34). □

Probabilistic Estimates
We next introduce the following lemma, whose prove can be found in [13]. Note that the lemma improves those from [12] for the matrix cases and Lemma 7.2 in [24] for the operator cases , as it does not need the assumption that the sample size is large enough while considering the influence of γ for the logarithmic factor.

Lemma 5.3. *Under Assumption 3, let $\delta \in (0,1)$, $\lambda = n^{-\theta}$ for some $\theta \geq 0$, and*

$$a_{n,\delta,\gamma}(\theta) = 8\kappa^2\left(\log\frac{4\kappa^2(c_\gamma+1)}{\delta\|\mathcal{T}\|} + \theta\gamma\min\left(\frac{1}{\mathrm{e}(1-\theta)_+}, \log n\right)\right). \tag{35}$$

We have with probability at least $1-\delta$,

$$\|(\mathcal{T}+\lambda)^{1/2}(\mathcal{T}_{\mathbf{x}}+\lambda)^{-1/2}\|^2 \leq 3a_{n,\delta,\gamma}(\theta)(1\vee n^{\theta-1}),$$

and

$$\|(\mathcal{T}+\lambda)^{-1/2}(\mathcal{T}_{\mathbf{x}}+\lambda)^{1/2}\|^2 \leq \frac{4}{3}a_{n,\delta,\gamma}(\theta)(1\vee n^{\theta-1}),$$

To proceed the proof of our next lemmas, we need the following concentration result for Hilbert space valued random variable used in [5] and based on the results in [19].

Lemma 5.4. *Let $w_1,\cdots,w_m$ be i.i.d random variables in a Hilbert space with norm $\|\cdot\|$. Suppose that there are two positive constants B and σ^2 such that*

$$\mathbb{E}[\|w_1 - \mathbb{E}[w_1]\|^l] \leq \frac{1}{2}l!B^{l-2}\sigma^2, \quad \forall l \geq 2. \tag{36}$$

Then for any $0<\delta<1/2$, the following holds with probability at least $1-\delta$,

$$\left\|\frac{1}{m}\sum_{k=1}^m w_m - \mathbb{E}[w_1]\right\| \leq 2\left(\frac{B}{m} + \frac{\sigma}{\sqrt{m}}\right)\log\frac{2}{\delta}.$$

The following lemma is a consequence of the lemma above (see e.g., [26] for a proof).

Lemma 5.5. *Let $0 < \delta < 1/2$. It holds with probability at least $1 - \delta$:*

$$\|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\| \leq \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|_{HS} \leq \frac{6\kappa^2}{\sqrt{n}} \log \frac{2}{\delta}.$$

Here, $\|\cdot\|_{HS}$ denotes the Hilbert-Schmidt norm.

One novelty of this paper is the following new lemma, which provides a probabilistic estimate on the terms caused by both the variance and approximation error. The lemma is mainly motivated by [26, 5, 14, 13]. Note that the condition (17) is slightly weaker than the condition $\|f_H\|_\infty < \infty$ required in [14] for analyzing gradient methods.

Lemma 5.6. *Under Assumptions 1, 2 and 3, let ω_λ be given by (32). For all $\delta \in]0, 1/2[$, the following holds with probability at least $1 - \delta$:*

$$\|\mathcal{T}_\lambda^{-1/2}[(\mathcal{T}_{\mathbf{x}}\omega_\lambda - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}) - (\mathcal{T}\omega_\lambda - \mathcal{S}_\rho^* f_\rho)]\|_H \leq \left(\frac{C_1}{n\lambda^{\frac{1}{2}\vee(1-\zeta)}} + \sqrt{\frac{C_2(\phi(\lambda))^2}{n\lambda} + \frac{C_3}{n\lambda^\gamma}} \right) \log \frac{2}{\delta}. \quad (37)$$

Here, $C_1 = 8\kappa(M + E\phi(\kappa^2)\kappa^{(1-2\zeta)_+}), C_2 = 96c_g^2R^2\kappa^2$ and $C_3 = 32(3B^2 + 4Q^2)c_\gamma$.

Proof. Let $\xi_i = \mathcal{T}_\lambda^{-\frac{1}{2}}(\langle\omega_\lambda, x\rangle_H - y_i)x_i$ for all $i \in [n]$. From the definition of the regression function f_ρ in (2) and (8), a simple calculation shows that

$$\mathbb{E}[\xi] = \mathbb{E}[\mathcal{T}_\lambda^{-\frac{1}{2}}(\langle\omega_\lambda, x\rangle_H - f_\rho(x))x] = \mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}\omega_\lambda - \mathcal{S}_\rho^* f_\rho) = \mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}\omega_\lambda - \mathcal{S}_\rho^* f_H). \quad (38)$$

In order to apply Lemma 5.4, we need to estimate $\mathbb{E}[\|\xi - \mathbb{E}[\xi]\|_H^l]$ for any $l \in \mathbb{N}$ with $l \geq 2$. In fact, using Hölder's inequality twice,

$$\mathbb{E}\|\xi - \mathbb{E}[\xi]\|_H^l \leq \mathbb{E}\left(\|\xi\|_H + \mathbb{E}\|\xi\|_H\right)^l \leq 2^{l-1}(\mathbb{E}\|\xi\|_H^l + (\mathbb{E}\|[\xi]\|_H)^l) \leq 2^l\mathbb{E}\|\xi\|_H^l. \quad (39)$$

We now estimate $\mathbb{E}\|\xi\|_H^l$. By Hölder's inequality,

$$\mathbb{E}\|\xi\|_H^l = \mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^l(y - \langle\omega_\lambda, x\rangle_H)^l] \leq 2^{l-1}\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^l(|y|^l + |\langle\omega_\lambda, x\rangle_H|^l)].$$

According to (3), one has

$$\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H \leq \|\mathcal{T}_\lambda^{-\frac{1}{2}}\|\|x\|_H \leq \frac{1}{\sqrt{\lambda}}\kappa. \quad (40)$$

Moreover, by Cauchy-Schwarz inequality and (3), $|\langle\omega_\lambda, x\rangle_H| \leq \|\omega_\lambda\|_H\|x\|_H \leq \kappa\|\omega_\lambda\|_H$. Thus, we get

$$\mathbb{E}\|\xi\|_H^l \leq 2^{l-1}\left(\frac{\kappa}{\sqrt{\lambda}}\right)^{l-2}\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2(|y|^l + (\kappa\|\omega_\lambda\|_H)^{l-2}|\langle\omega_\lambda, x\rangle_H|^2). \quad (41)$$

Note that by (15),

$$\begin{aligned}\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2|y|^l] &= \int_H \|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2 \int_{\mathbb{R}} |y|^l d\rho(y|x)d\rho_X(x) \\ &\leq \frac{1}{2}l!M^{l-2}Q^2\int_H \|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2 d\rho_X(x).\end{aligned}$$

Using $\|w\|_H^2 = \mathrm{tr}(w \otimes w)$ which implies

$$\int_H \|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 d\rho_X(x) = \int_H \mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}} x \otimes x \mathcal{T}_\lambda^{-\frac{1}{2}}) d\rho_X(x) = \mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}} \mathcal{T} \mathcal{T}_\lambda^{-\frac{1}{2}}) = \mathcal{N}(\lambda), \tag{42}$$

we get

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 |y|^l] \leq \frac{1}{2} l! M^{l-2} Q^2 \mathcal{N}(\lambda). \tag{43}$$

Besides, by Cauchy-Schwarz inequality,

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 |\langle \omega_\lambda, x\rangle_H|^2] \leq 3\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 (|\langle \omega_\lambda, x\rangle_H - f_H(x)|^2 + |f_H(x) - f_\rho(x)|^2 + |f_\rho(x)|^2)].$$

By (40) and (33),

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 (|\langle \omega_\lambda, x\rangle_H - f_H(x)|^2] \leq \frac{\kappa^2}{\lambda} \mathbb{E}[|\langle \omega_\lambda, x\rangle_H - f_H(x)|^2] = \frac{\kappa^2}{\lambda} \|\mathcal{S}_\rho \omega_\lambda - f_H\|_\rho^2 \leq c_g^2 R^2 \kappa^2 \frac{(\phi(\lambda))^2}{\lambda},$$

and by (16) and (42),

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 |f_\rho(x)|^2] \leq Q^2 \mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2] = Q^2 \mathcal{N}(\lambda).$$

Therefore,

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 |\langle \omega_\lambda, x\rangle_H|^2] \leq 3 \left(c_g^2 R^2 \kappa^2 \phi^2(\lambda) \lambda^{-1} + \mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 |f_H(x) - f_\rho(x)|^2] + Q^2 \mathcal{N}(\lambda) \right).$$

Using $\|w\|_H^2 = \mathrm{tr}(w \otimes w)$ and (17), we have

$$\begin{aligned}
\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 |f_H(x) - f_\rho(x)|^2] =& \mathbb{E}[|f_H(x) - f_\rho(x)|^2 \,\mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}} x \otimes x \mathcal{T}_\lambda^{-\frac{1}{2}})] \\
=& \,\mathrm{tr}(\mathcal{T}_\lambda^{-1} \mathbb{E}[(f_H(x) - f_\rho(x))^2 x \otimes x]) \\
\leq& B^2 \,\mathrm{tr}(\mathcal{T}_\lambda^{-1} \mathcal{T}) = B^2 \mathcal{N}(\lambda),
\end{aligned}$$

and therefore,

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H^2 |\langle \omega_\lambda, x\rangle_H|^2] \leq 3 \left(c_g^2 R^2 \kappa^2 (\phi(\lambda))^2 \lambda^{-1} + (B^2 + Q^2) \mathcal{N}(\lambda) \right).$$

Introducing the above estimate and (43) into (41), we derive

$$\begin{aligned}
\mathbb{E}\|\xi\|_H^l \leq& 2^{l-1} \left(\frac{\kappa}{\sqrt{\lambda}} \right)^{l-2} \left(\frac{1}{2} l! M^{l-2} Q^2 \mathcal{N}(\lambda) + 3(\kappa \|\omega_\lambda\|_H)^{l-2} (c_g^2 R^2 \kappa^2 (\phi(\lambda))^2 \lambda^{-1} + (B^2 + Q^2) \mathcal{N}(\lambda)) \right) \\
\leq& 2^{l-1} \left(\frac{\kappa M + \kappa^2 \|\omega_\lambda\|_H}{\sqrt{\lambda}} \right)^{l-2} \frac{1}{2} l! \left(Q^2 \mathcal{N}(\lambda) + 3(c_g^2 R^2 \kappa^2 (\phi(\lambda))^2 \lambda^{-1} + (B^2 + Q^2) \mathcal{N}(\lambda)) \right), \\
\leq& 2^{l-1} \left(\frac{\kappa M + \kappa^2 \|\omega_\lambda\|_H}{\sqrt{\lambda}} \right)^{l-2} \frac{1}{2} l! \left(3 c_g^2 R^2 \kappa^2 (\phi(\lambda))^2 \lambda^{-1} + (3B^2 + 4Q^2) c_\gamma \lambda^{-\gamma} \right),
\end{aligned}$$

where for the last inequality, we used Assumption 3. Introducing the above estimate into (39), and then substituting with (34) and noting that $\lambda \leq 1$, we get

$$\mathbb{E}[\|\xi - \mathbb{E}[\xi]\|_H^l] \leq \frac{1}{2} l! \left(\frac{4\kappa(M + E\phi(\kappa^2)\kappa^{(1-2\zeta)_+})}{\lambda^{\frac{1}{2} \vee (1-\zeta)}} \right)^{l-2} 8 \left(3 c_g^2 R^2 \kappa^2 (\phi(\lambda))^2 \lambda^{-1} + (3B^2 + 4Q^2) c_\gamma \lambda^{-\gamma} \right).$$

Applying Lemma 5.4, one can get the desired result. □

Basic Operator Inequalities

Lemma 5.7. *[9, Cordes inequality] Let A and B be two positive bounded linear operators on a separable Hilbert space. Then*

$$\|A^s B^s\| \leq \|AB\|^s, \quad \textit{when } 0 \leq s \leq 1.$$

Lemma 5.8. *[17, 16] Suppose ψ is an operator monotone index function on $[0, b]$, with $b > 1$. Then there is a constant $c_\psi < \infty$ depending on $b-a$, such that for any pair B_1, B_2, $\|B_1\|, \|B_2\| \leq a$, of non-negative self-adjoint operators on some Hilbert space, it holds,*

$$\|\psi(B_1) - \psi(B_2)\| \leq c_\psi \psi(\|B_1 - B_2\|).$$

Moreover, there is $c'_\psi > 0$ such that

$$c'_\psi \frac{\lambda}{\psi(\lambda)} \leq \frac{\sigma}{\psi(\sigma)},$$

whenever $0 < \lambda < \sigma \leq a < b$.

Lemma 5.9. *Let $\vartheta : [0, a] \to \mathbb{R}_+$ be Lipschitz continuous with constant 1 and $\vartheta(0) = 0$. Then for any pair B_1, B_2, $\|B_1\|, \|B_2\| \leq a$, of non-negative self-adjoint operators on some Hilbert space, it holds,*

$$\|\vartheta(B_1) - \vartheta(B_2)\|_{HS} \leq \|B_1 - B_2\|_{HS}.$$

Proof. The result follows from [2, Subsection 8.2]. □

5.2 Proof of Main Results

Now we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. Following from Lemmas 5.3, 5.5 and 5.6, and by a simple calculation, with $\lambda = n^{\theta-1}$ and $\theta \in [0, 1]$, we get that with probability at least $1 - \delta$, the following holds:

$$\|\mathcal{T}_\lambda^{\frac{1}{2}} \mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\|^2 \vee \|\mathcal{T}_\lambda^{-\frac{1}{2}} \mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\|^2 \leq \Delta_1, \quad \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\| \leq \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|_{HS} \leq \Delta_3, \tag{44}$$

$$\text{and} \quad \|\mathcal{T}_\lambda^{-1/2}[(\mathcal{T}_{\mathbf{x}}\omega_\lambda - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}) - (\mathcal{T}\omega_\lambda - \mathcal{S}_\rho^* f_\rho)]\|_H \leq \Delta_2, \tag{45}$$

where

$$\Delta_1 = C_4(\log\frac{6}{\delta} + \gamma(\frac{1}{\theta} \wedge \log n)), \quad C_4 = 24\kappa^2 \log\frac{2\mathrm{e}\kappa^2(c_\gamma + 1)}{\|\mathcal{T}\|},$$

$$\Delta_2 = \left(\frac{C_1}{n\lambda^{\frac{1}{2}\vee(1-\zeta)}} + \sqrt{C_2}\phi(\lambda) + \frac{\sqrt{C_3}}{\sqrt{n\lambda^\gamma}}\right)\log\frac{6}{\delta},$$

$$\Delta_3 = \frac{6\kappa^2}{\sqrt{n}}\log\frac{6}{\delta}.$$

Obviously, we have $\Delta_1 \geq 1$ since $\log\frac{6}{\delta} > 1$ and by (4), $C_4 \geq 1$.

We now begin with the following inequality:

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho = \|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq \|\mathcal{L}^{-a}\mathcal{S}_\rho(\omega_\lambda^{\mathbf{z}} - \omega_\lambda)\|_\rho + \|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda - f_H)\|_\rho.$$

Introducing with (33), we get

$$\begin{aligned}\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq & \|\mathcal{L}^{-a}\mathcal{S}_\rho(\omega_\lambda^{\mathbf{z}} - \omega_\lambda)\|_\rho + c_g R\phi(\lambda)\lambda^{-a} \\ \leq & \|\mathcal{L}^{-a}\mathcal{S}_\rho\mathcal{T}^{a-\frac{1}{2}}\|\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{T}_{\mathbf{x}\lambda}^{a-\frac{1}{2}}\|\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \omega_\lambda)\|_H + c_g R\phi(\lambda)\lambda^{-a}.\end{aligned}$$

By the spectral theorem, $\mathcal{L} = \mathcal{S}_\rho\mathcal{S}_\rho^*$, $\mathcal{T} = \mathcal{S}_\rho^*\mathcal{S}_\rho$, and (4), we have $\|\mathcal{L}^{-a}\mathcal{S}_\rho\mathcal{T}^{a-\frac{1}{2}}\| \leq 1$. Moreover, by Lemma 5.7 and $0 \leq a \leq \zeta \wedge \frac{1}{2}$,

$$\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{T}_{\mathbf{x}\lambda}^{a-\frac{1}{2}}\| = \|\mathcal{T}_\lambda^{\frac{1}{2}(1-2a)}\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}(1-2a)}\| \leq \|\mathcal{T}_\lambda^{\frac{1}{2}}\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\|^{1-2a} \leq \Delta_1^{\frac{1}{2}-a}.$$

We thus get

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq \Delta_1^{\frac{1}{2}-a}\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \omega_\lambda)\|_H + c_g R\phi(\lambda)\lambda^{-a}.$$

Subtracting and adding with the same term, using the triangle inequality and recalling the notation $\mathcal{R}_\lambda(u)$ defined in (30), we get

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq \Delta_1^{\frac{1}{2}-a}\left(\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\omega_\lambda\|_H + \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H\right) + c_g R\phi(\lambda)\lambda^{-a}.$$

Introducing with (13),

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq \Delta_1^{\frac{1}{2}-a}\left(\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\omega_\lambda\|_H + \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H\right) + c_g R\phi(\lambda)\lambda^{-a}. \tag{46}$$

Estimating $\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H$:
We first have

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H \leq \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\|\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\mathcal{T}_\lambda^{\frac{1}{2}}\|\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H.$$

With (11) and (7), we have

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\| \leq \sup_{u\in[0,\kappa^2]}|(u+\lambda)^{1-a}\mathcal{G}_\lambda(u)| \leq \sup_{u\in[0,\kappa^2]}|(u^{1-a}+\lambda^{1-a})\mathcal{G}_\lambda(u)| \leq 2E\lambda^{-a},$$

and thus

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H \leq 2E\lambda^{-a}\Delta_1^{1/2}\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H$$

Since by (33) and (8),

$$\begin{aligned}&\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H \\ \leq & \|\mathcal{T}_\lambda^{-\frac{1}{2}}[(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda) - (\mathcal{T}\omega_\lambda - \mathcal{S}_\rho^* f_\rho)]\|_H + \|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}\omega_\lambda - \mathcal{S}_\rho^* f_\rho)\|_H \\ \leq & \|\mathcal{T}_\lambda^{-\frac{1}{2}}[(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda) - (\mathcal{T}\omega_\lambda - \mathcal{S}_\rho^* f_\rho)]\|_H + \|\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{S}_\rho^*\|\|\mathcal{S}_\rho\omega_\lambda - f_H\|_\rho \\ \leq & \Delta_2 + c_g R\phi(\lambda),\end{aligned}$$

we thus have

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T}_{\mathbf{x}})(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}\omega_\lambda)\|_H \leq 2E\lambda^{-a}\Delta_1^{1/2}(\Delta_2 + c_g R\phi(\lambda)). \tag{47}$$

Estimating $\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\omega_\lambda\|_H$:
Note that from the definition of ω_λ in (32), (18), $\mathcal{L}=\mathcal{S}_\rho\mathcal{S}_\rho^*$, and $\mathcal{T}=\mathcal{S}_\rho^*\mathcal{S}_\rho$,

$$\omega_\lambda=\mathcal{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^*\phi(\mathcal{L})g_0=\mathcal{G}_\lambda(\mathcal{T})\phi(\mathcal{T})\mathcal{S}_\rho^*g_0,$$

and thus,

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\omega_\lambda\|_H\leq\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{G}_\lambda(\mathcal{T})\phi(\mathcal{T})\mathcal{S}_\rho^*\|R=\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{G}_\lambda(\mathcal{T})\phi(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|R. \quad (48)$$

In what follows, we will estimate $\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{G}_\lambda(\mathcal{T})\phi(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|$, considering three different cases.
Case 1: $\phi(\cdot)$ is operator monotone.
We first have

$$\begin{aligned}\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\leq&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\|\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\mathcal{T}_\lambda^{\frac{1}{2}}\|\|\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{T}^{\frac{1}{2}}\|\|\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\|\\\leq&\|\mathcal{T}_{\mathbf{x}\lambda}^{1-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\|\Delta_1^{\frac{1}{2}}\|\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\|\end{aligned}$$

By the spectral theorem and (12), with (7),

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{1-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\|\leq\sup_{u\in[0,\kappa^2]}|(u+\lambda)^{1-a}\mathcal{R}_\lambda(u)|\leq\sup_{u\in[0,\kappa^2]}|(u^{1-a}+\lambda^{1-a})\mathcal{R}_\lambda(u)|\leq 2F\lambda^{1-a},$$

(where we write $F_\tau=F$ throughout) and it thus follows that

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\leq 2F\Delta_1^{\frac{1}{2}}\lambda^{1-a}\|\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\|.$$

Using the spectral theorem, with (4), we get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\leq 2F\Delta_1^{\frac{1}{2}}\lambda^{1-a}\sup_{u\in[0,\kappa^2]}|\mathcal{G}_\lambda(u)\phi(u)|.$$

When $0<u\leq\lambda$, as $\phi(u)$ is non-decreasing, $\phi(u)\leq\phi(\lambda)$. Applying (11), we have

$$\mathcal{G}_\lambda(u)\phi(u)\leq E\phi(\lambda)\lambda^{-1}.$$

When $\lambda<u\leq\kappa^2$, following from Lemma 5.8, we have that there is a $c'_\phi\geq 1$, which depends only on ϕ, κ^2 and b, such that

$$\phi(u)u^{-1}\leq c'_\phi\phi(\lambda)\lambda^{-1}.$$

Then, combing with (11),

$$\mathcal{G}_\lambda(u)\phi(u)=\mathcal{G}_\lambda(u)u\phi(u)u^{-1}\leq Ec'_\phi\phi(\lambda)\lambda^{-1}.$$

Therefore, for all $0<u\leq\kappa^2$, $\mathcal{G}_\lambda(u)\phi(u)\leq Ec'_\phi\phi(\lambda)\lambda^{-1}$ and consequently,

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\leq 2Ec'_\phi F\Delta_1^{\frac{1}{2}}\lambda^{-a}\phi(\lambda).$$

Introducing the above into (48), we get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\omega_\lambda\|_H\leq 2Ec'_\phi FR\Delta_1^{\frac{1}{2}}\lambda^{-a}\phi(\lambda). \quad (49)$$

Case 2: $\phi(\cdot)$ is Lipschitz continuous with constant 1.
By the triangle inequality, we have

$$\begin{aligned}
&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \\
\leq&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T}_{\mathbf{x}})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| + \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})(\phi(\mathcal{T})-\phi(\mathcal{T}_{\mathbf{x}}))\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \\
\leq&\|\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T}_{\mathbf{x}})\|\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{T}_\lambda^{a-\frac{1}{2}}\|\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| + \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\|\|\phi(\mathcal{T})-\phi(\mathcal{T}_{\mathbf{x}})\|_{HS}\|\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|.
\end{aligned}$$

Since $\phi(u)$ is Lipschitz continuous with constant 1 and $\phi(0)=0$, then according to Lemma 5.9, $\|\phi(\mathcal{T})-\phi(\mathcal{T}_{\mathbf{x}})\|_{HS}\leq\|\mathcal{T}-\mathcal{T}_{\mathbf{x}}\|_{HS}$. It thus follows that

$$\begin{aligned}
&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \\
\leq&\|\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T}_{\mathbf{x}})\|\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{T}_\lambda^{a-\frac{1}{2}}\|\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| + \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\|\|\mathcal{T}-\mathcal{T}_{\mathbf{x}}\|_{HS}\|\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \\
\leq&c_g\phi(\lambda)\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{T}_\lambda^{a-\frac{1}{2}}\|\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| + \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\|\Delta_3\|\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|,
\end{aligned}$$

where for the last inequality, we used (31) to bound $\|\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T}_{\mathbf{x}})\|$:

$$\|\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T}_{\mathbf{x}})\|\leq\sup_{u\in[0,\kappa^2]}|\mathcal{R}_\lambda(u)\phi(u)|\leq c_g\phi(\lambda).$$

Applying Lemma 5.7 which implies

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{T}_\lambda^{a-\frac{1}{2}}\| = \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}(1-2a)}\mathcal{T}_\lambda^{-\frac{1}{2}(1-2a)}\|\leq\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\mathcal{T}_\lambda^{-\frac{1}{2}}\|^{1-2a}\leq\Delta_1^{\frac{1}{2}-a},$$

we get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\leq c_g\phi(\lambda)\Delta_1^{\frac{1}{2}-a}\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| + \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\|\Delta_3\|\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|. \tag{50}$$

By the spectral theorem and (11), with (4) and $0\leq a\leq\frac{1}{2}$, we have

$$\|\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\leq\sup_{u\in[0,\kappa^2]}|u^{\frac{1}{2}}\mathcal{G}_\lambda(u)|\leq E\lambda^{-\frac{1}{2}}\quad\text{and} \tag{51}$$

$$\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\leq\sup_{u\in[0,\kappa^2]}(u^{\frac{1}{2}-a}+\lambda^{\frac{1}{2}-a})|\mathcal{G}_\lambda(u)|u^{\frac{1}{2}}\leq 2E\lambda^{-a}.$$

Similarly, by (12), with (7),

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\|\leq\sup_{u\in[0,\kappa^2]}(u^{\frac{1}{2}-a}+\lambda^{\frac{1}{2}-a})|\mathcal{R}_\lambda(u)|\leq 2F\lambda^{\frac{1}{2}-a}.$$

Therefore, following from the above three estimates and (50), we get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\leq 2c_g\phi(\lambda)\Delta_1^{\frac{1}{2}-a}E\lambda^{-a}+2EF\lambda^{-a}\Delta_3. \tag{52}$$

Introducing the above into (48), we get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\omega_\lambda\|_H\leq 2ER\lambda^{-a}(c_g\phi(\lambda)\Delta_1^{\frac{1}{2}-a}+F\Delta_3). \tag{53}$$

Applying (53) (or (49)) and (47) into (46), by a direct calculation, we get

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}}-f_H)\|_\rho\leq\Delta_1^{\frac{1}{2}-a}2E\lambda^{-a}(\Delta_1^{\frac{1}{2}}\Delta_2+\Delta_1^{\frac{1}{2}}R(c_\phi''+c_g)\phi(\lambda)+FR\Delta_3)+c_gR\phi(\lambda)\lambda^{-a}.$$

Here, $c''_\phi = c'_\phi F$ if ϕ is operator monotone or $c''_\phi = c_g$ if ϕ is Lipschitz continuous with constant 1. Introducing with Δ_1, Δ_2 and Δ_3, by a direct calculation and $\lambda \leq 1$, one can prove the first part of the theorem with

$$\tilde{C}_1 = 2EC_1C_4^{1-a}, \quad \tilde{C}_2 = 2E\sqrt{C}_3C_4^{1-a} + 12\kappa^2 EFRC_4^{\frac{1}{2}-a}, \quad \text{and}$$
$$\tilde{C}_3 = 2E\sqrt{C}_2C_4^{1-a} + c_gR + 2ERC_4^{1-a}(c''_\phi + c_g).$$

Case 3: $\phi = \psi\vartheta$, where ψ is operator monotone and ϑ is Lipschitz continuous with constant 1. Since $\phi = \vartheta\psi$, we can rewrite $\phi(T)$ as

$$\phi(\mathcal{T}_\mathbf{x}) + (\vartheta(\mathcal{T}) - \vartheta(\mathcal{T}_\mathbf{x}))\psi(\mathcal{T}) + \vartheta(\mathcal{T}_\mathbf{x})(\psi(\mathcal{T}) - \psi(\mathcal{T}_\mathbf{x})).$$

Thus, together with the triangle inequality,

$$\begin{aligned}
&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \\
\leq&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\phi(\mathcal{T}_\mathbf{x})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| + \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})(\vartheta(\mathcal{T}) - \vartheta(\mathcal{T}_\mathbf{x}))\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\|\psi(\mathcal{T})\| \\
&+ \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\vartheta(\mathcal{T}_\mathbf{x})(\psi(\mathcal{T}) - \psi(\mathcal{T}_\mathbf{x}))\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|. \qquad (54)
\end{aligned}$$

Following the same argument as that for (52), we know that

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\phi(\mathcal{T}_\mathbf{x})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \leq 2c_gE\phi(\lambda)\Delta_1^{\frac{1}{2}-a}\lambda^{-a}, \qquad (55)$$

and

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})(\vartheta(\mathcal{T}) - \vartheta(\mathcal{T}_\mathbf{x}))\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \leq 2EF\lambda^{-a}\Delta_3. \qquad (56)$$

As the quality of $\mathcal{G}_\lambda$ covers $\vartheta(u)u^{\frac{1}{2}-a}$, applying the spectral theorem and Lemma 5.1, we get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\vartheta(\mathcal{T}_\mathbf{x})\| \leq \sup_{u\in[0,\kappa^2]}(u+\lambda)^{\frac{1}{2}-a}\mathcal{R}_\lambda(u)\vartheta(u) \leq c'_g\vartheta(\lambda)(\lambda^{\frac{1}{2}-a} + \lambda^{\frac{1}{2}-a}).$$

Since ψ is operator monotone on $[0, b]$ where $b > \kappa^2$, we know from Lemma 5.8 that there exists a positive constant $c_\psi < \infty$ depending on $b - \kappa^2$, such that

$$\|\psi(\mathcal{T}_\mathbf{x}) - \psi(\mathcal{T})\| \leq c_\psi\psi(\|\mathcal{T} - \mathcal{T}_\mathbf{x}\|).$$

If $\sqrt{n} \geq 6\log\frac{6}{\delta}$, as ψ is non-decreasing, following from (44), we have $\psi(\|\mathcal{T} - \mathcal{T}_\mathbf{x}\|) \leq \psi(\|\mathcal{T} - \mathcal{T}_\mathbf{x}\|_{HS}) \leq \psi(\Delta_3)$ and thus

$$\|\psi(\mathcal{T}_\mathbf{x}) - \psi(\mathcal{T})\| \leq c_\psi\psi(\Delta_3) \leq c_\psi c'_\psi\psi(n^{-1/2})6\kappa^2\log\frac{6}{\delta},$$

where for the last inequality, we used Lemma 5.8. If $\sqrt{n} \leq 6\log\frac{6}{\delta}$, then as $\|\mathcal{T} - \mathcal{T}_\mathbf{x}\| \leq \max(\|\mathcal{T}\|, \|\mathcal{T}_\mathbf{x}\|) \leq \kappa^2$,

$$\|\psi(\mathcal{T}_\mathbf{x}) - \psi(\mathcal{T})\| \leq c_\psi\psi(\kappa^2) \leq c_\psi\psi(\kappa^2)6\log\frac{6}{\delta}\frac{1}{\sqrt{n}} \leq c'_\psi c_\psi 6\kappa^2\log\frac{6}{\delta}\psi(n^{-1/2}),$$

where for the last inequality, we used Lemma 5.8. Therefore, following from the above analysis and (51),

$$\begin{aligned}
&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\vartheta(\mathcal{T}_\mathbf{x})(\psi(\mathcal{T}) - \psi(\mathcal{T}_\mathbf{x}))\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \\
\leq&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_\mathbf{x})\vartheta(\mathcal{T}_\mathbf{x})\|\|\psi(\mathcal{T}) - \psi(\mathcal{T}_\mathbf{x})\|\|\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\| \\
\leq&12c'_gc_\psi c'_\psi E\kappa^2\log\frac{6}{\delta}\lambda^{-a}\vartheta(\lambda)\psi(n^{-1/2}).
\end{aligned}$$

Introducing the above estimate, (55) and (56) into (54), with $\|\psi(\mathcal{T})\| \leq \psi(\kappa^2)$ (since ψ is operator monotone and (4)), we conclude that

$$\begin{aligned}&\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\phi(\mathcal{T})\mathcal{G}_\lambda(\mathcal{T})\mathcal{T}^{\frac{1}{2}}\|\\ \leq&2\lambda^{-a}\left(c_gE\phi(\lambda)\Delta_1^{\frac{1}{2}-a}+EF\Delta_3\psi(\kappa^2)+6\kappa^2c_g'c_\psi c_\psi'E\vartheta(\lambda)\psi(n^{-\frac{1}{2}})\log\frac{6}{\delta}\right).\end{aligned}$$

Introducing the above into (48), we get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{T}_{\mathbf{x}})\omega_\lambda\| \leq 2\lambda^{-a}\left(c_gE\phi(\lambda)\Delta_1^{\frac{1}{2}-a}+EF\psi(\kappa^2)\Delta_3+6\kappa^2c_g'c_\psi c_\psi'E\vartheta(\lambda)\psi(n^{-\frac{1}{2}})\log\frac{6}{\delta}\right)R.$$

Combining the above and (47) with (46), by a direct calculation, we get

$$\begin{aligned}&\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}}-f_H)\|_\rho\\ \leq&\lambda^{-a}\Big(2E\Delta_1^{1-a}(\Delta_2+2c_gR\phi(\lambda))+c_gR\phi(\lambda)+2EFR\psi(\kappa^2)\Delta_1^{\frac{1}{2}-a}\Delta_3\\ &+12\kappa^2c_g'ERc_\psi c_\psi'\Delta_1^{\frac{1}{2}-a}\vartheta(\lambda)\psi(n^{-\frac{1}{2}})\log\frac{6}{\delta}\Big).\end{aligned}$$

Introducing with Δ_1, Δ_2 and Δ_3, by a simple calculation, with $\lambda \leq 1$, we can prove the second part of the theorem with

$$\tilde{C}_4 = 2EC_4^{1-a}\sqrt{C_3}+12EFR\psi(\kappa^2)\kappa^2C_4^{\frac{1}{2}-a},\quad \tilde{C}_5 = 2EC_4^{1-a}(3c_gR+\sqrt{C_2}),$$

$$\text{and}\quad \tilde{C}_6 = 12\kappa^2c_g'c_\psi c_\psi'ERC_4^{\frac{1}{2}-a}.$$

□

Proof of Corollary 4.3. Let θ be such that $\lambda = n^{\theta-1}$. As $\Theta(u)$ is non-decreasing, $\Theta(0) = 0$, $\Theta(1) = 1$ and that λ satisfies $\Theta(\lambda) = n^{-1}$, then $0 \leq \lambda \leq 1$. Moreover, as that $\phi(\lambda)\lambda^{-\zeta}$ is non-decreasing which implies

$$\left(\frac{\phi(1)}{\sqrt{n}\phi(\lambda)\lambda^{-\zeta}}\right)^2 \geq \frac{1}{n}, \tag{57}$$

and that

$$\lambda^{\gamma+2\zeta} = \left(\frac{\phi(1)}{\sqrt{n}\phi(\lambda)\lambda^{-\zeta}}\right)^2,$$

then $\lambda \geq n^{-\frac{1}{2\zeta+\gamma}}$. Thus, with $2\zeta+\gamma > 1$, $\theta = \log_n\lambda + 1 \geq -\frac{1}{2\zeta+\gamma}+1 > 0$. Also, $\theta \leq 1$ as $\lambda \leq 1$. Applying Part 1) of Theorem 4.2, and noting that by $2\zeta+\gamma > 1$, $1 \geq \lambda \geq n^{-\frac{1}{2\zeta+\gamma}}$ and (57),

$$\frac{1}{n\lambda^{\frac{1}{2}}} \leq \frac{1}{\sqrt{n}} \leq \frac{1}{\sqrt{n\lambda^\gamma}} = \frac{\phi(\lambda)}{\phi(1)},\quad \frac{1}{n\lambda^{1-\zeta}} \leq \frac{1}{\sqrt{n\lambda^\gamma}}\ (\text{if } 2\zeta \leq 1).$$

we can prove the first desired result. The second desired result can be proved by using Part 2) of Theorem 4.2, the above estimates, as well as $\psi(n^{-1/2}) \leq \psi(\lambda)$ (since ψ is non-decreasing). □

Proof of Corollary 4.4. If $\zeta \leq 1$, then ϕ is operator monotone [17, Theorem 1 and Example 1]. If $\zeta \geq 1$, then ϕ is Lipschitz continuous with constant 1 over $[0, \kappa^2]$. Applying Part 1) of Theorem 4.2, one can prove the desired results. □

Proof of Corollary 4.5. The proof can be done by using Corollary 4.4 with simple arguments. For notional simplicity, we let

$$\Lambda_n = \begin{cases} n^{-\frac{(\zeta-a)}{2\zeta+\gamma}} & \text{if } 2\zeta+\gamma > 1, \\ n^{-(\zeta-a)} \left(1 \vee \log n^{\gamma}\right)^{(1-a)} & \text{if } 2\zeta+\gamma \leq 1. \end{cases}$$

1) Using the fact that for any non-negative random variable ξ, $\mathbb{E}[\xi] = \int_{t\geq 0} \Pr(\xi \geq t)dt$, and Corollary 4.4, for any $q \in \mathbb{N}_+$

$$\mathbb{E}\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho^q \leq C \int_{t\geq 0} \exp\left\{\left(-\frac{t}{C\Lambda_n^q}\right)^{\frac{1}{q(2-a)}}\right\} dt \leq C\Lambda_n^q.$$

2) By Corollary 4.4, we have, with $\delta_n = n^{-2}$,

$$\sum_{n=1}^{\infty} \Pr\left(n^{\frac{\zeta-a-\epsilon}{1\vee(2\zeta+\gamma)}} \|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho > Cn^{\frac{\zeta-a-\epsilon}{1\vee(2\zeta+\gamma)}} \Lambda_n \log^{2-a} \frac{6}{\delta_n}\right) \leq \sum_{n=1}^{\infty} \delta_n^2 < \infty.$$

Note that $Cn^{\frac{\zeta-a-\epsilon}{1\vee(2\zeta+\gamma)}} \Lambda_n \log^{2-a} \frac{6}{\delta_n} \to 0$ as $n \to \infty$. Thus, applying the Borel-Cantelli lemma, one can prove Part 2).
3) Following the argument from the proof of Corollary 4.4, one can prove Part 3). □

Acknowledgment

JL and VC's work was supported in part by Office of Naval Research (ONR) under grant agreement number N62909-17-1-2111, in part by Hasler Foundation Switzerland under grant agreement number 16066, and in part by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 725594).

References

[1] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

[2] M. S. Birman and M. Solomyak. Double operator integrals in a Hilbert space. *Integral Equations and Operator Theory*, 47(2):131–168, 2003.

[3] G. Blanchard and N. Mucke. Optimal rates for regularization of statistical inverse learning problems. *arXiv preprint arXiv:1604.04054*, 2016.

[4] A. Caponnetto. Optimal learning rates for regularization operators in learning theory. *Technical report*, 2006.

[5] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[6] F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

[7] L. H. Dicker, D. P. Foster, and D. Hsu. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. *Electronic Journal of Statistics*, 11(1):1022–1047, 2017.

[8] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[9] J. Fujii, M. Fujii, T. Furuta, and R. Nakamoto. Norm inequalities equivalent to Heinz inequality. *Proceedings of the American Mathematical Society*, 118(3):827–830, 1993.

[10] L. L. Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[11] Z.-C. Guo, S.-B. Lin, and D.-X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 2017.

[12] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

[13] J. Lin and V. Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Arxiv*, 2018.

[14] J. Lin and L. Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

[15] S.-B. Lin, X. Guo, and D.-X. Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18(92):1–31, 2017.

[16] P. Mathé and S. Pereverzev. Regularization of some linear ill-posed problems with discretized random noisy data. *Mathematics of Computation*, 75(256):1913–1929, 2006.

[17] P. Mathé and S. V. Pereverzev. Moduli of continuity for operator valued functions. 2002.

[18] G. Myleiko, S. Pereverzyev Jr, and S. Solodky. Regularized Nyström subsampling in regression and ranking problems under general smoothness assumptions. 2017.

[19] I. Pinelis and A. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory of Probability & Its Applications*, 30(1):143–148, 1986.

[20] J. O. Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

[21] A. Rastogi and S. Sampath. Optimal rates for the regularized learning algorithms under general source condition. *Frontiers in Applied Mathematics and Statistics*, 3:3, 2017.

[22] L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[23] A. Rudi, R. Camoriano, and L. Rosasco. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, pages 1657–1665, 2015.

[24] A. Rudi, G. D. Canas, and L. Rosasco. On the sample complexity of subspace learning. In *Advances in Neural Information Processing Systems*, pages 2067–2075, 2013.

[25] A. Rudi and L. Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pages 3215–3225, 2017.

[26] S. Smale and D.-X. Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[27] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[28] I. Steinwart, D. R. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In *Conference On Learning Theory*, 2009.

[29] Z. Szabó, A. Gretton, B. Póczos, and B. Sriperumbudur. Two-stage sampled learning theory on distributions. In *Artificial Intelligence and Statistics*, pages 948–957, 2015.

[30] Q. Wu, Y. Ying, and D.-X. Zhou. Learning rates of least-square regularized regression. *Foundations of Computational Mathematics*, 6(2):171–192, 2006.

[31] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[32] T. Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

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

A List of Notations

Notation	Meaning
H	the input space - separable Hilbert space
ρ, ρ_X	the fixed probability measure on $H \times \mathbb{R}$, the induced marginal measure of ρ on H
$\rho(\cdot\|x)$	the conditional probability measure on $\mathbb{R}$ w.r.t. $x \in H$ and ρ
H_ρ	the hypothesis space, $\{f : H \to \mathbb{R} \| \exists \omega \in H$ with $f(x) = \langle \omega, x\rangle_H, \rho_X$-almost surely$\}$.
n	the sample size
$\mathbf{z}$	the whole samples $\{z_i\}_{i=1}^n$, where each z_i is i.i.d. according to ρ
$\mathbf{y}$	the vector of sample outputs, $(y_1, \cdots, y_n)^\top$
$\mathbf{x}$	the set of sample outputs, $\{x_1, \cdots, x_n\}$
$\mathcal{E}$	the expected risk defined by (1)
$L^2_{\rho_X}$	the Hilbert space of square integral functions from H to $\mathbb{R}$ with respect to ρ_X
f_ρ	the regression function defined (2)
κ^2	the constant from the bounded assumption (3) on the input space H
$\mathcal{S}_\rho$	the linear map from $H \to L^2_{\rho_X}$ defined by $\mathcal{S}_\rho \omega = \langle \omega, \cdot \rangle_H$
$\mathcal{S}^*_\rho$	the adjoint operator of $\mathcal{S}_\rho$: $\mathcal{S}^*_\rho f = \int_X f(x) x d\rho_X(x)$
$\mathcal{L}$	the operator from $L^2_{\rho_X}$ to $L^2_{\rho_X}$, $\mathcal{L}(f) = \mathcal{S}_\rho \mathcal{S}^*_\rho f = \int_X \langle x, \cdot \rangle_H f(x) \rho_X(x)$
$\mathcal{T}$	the covariance operator from H to H, $\mathcal{T} = \mathcal{S}^*_\rho \mathcal{S}_\rho = \int_X \langle \cdot, x\rangle x d\rho_X(x)$
$\mathcal{S}_\mathbf{x}$	the sampling operator from H to $\mathbb{R}^n$, $(\mathcal{S}_\mathbf{x}\omega)_i = \langle \omega, x_i \rangle_H, i \in \{1, \cdots, n\}$
$\mathcal{S}^*_\mathbf{x}$	the adjoint operator of $\mathcal{S}_\mathbf{x}$, $\mathcal{S}^*_\mathbf{x}\mathbf{y} = \frac{1}{n}\sum_{i=1}^n y_i x_i$
$\mathcal{T}_\mathbf{x}$	the empirical covariance operator, $\mathcal{T}_\mathbf{x} = \mathcal{S}^*_\mathbf{x}\mathcal{S}_\mathbf{x} = \frac{1}{n}\sum_{i=1}^n \langle \cdot, x_i\rangle x_i$
f_H	the projection of f_ρ onto the closure of H_ρ in $L^2_{\rho_X}$
$\mathcal{G}_\lambda(\cdot)$	the filter function of the regularized algorithm from Definition 3.1
τ	the qualification of the filter function $\mathcal{G}_\lambda$
E, F_τ	the constants related to the filter function $\mathcal{G}_\lambda$ from (11) and (12)
λ	a regularization parameter $\lambda > 0$
$\omega^\mathbf{z}_\lambda$	an estimated vector defined by (13)

$f_\lambda^{\mathbf{z}}$	an estimated function defined by (14)
M, Q	the positive constants from Assumption (15)
B	the constant from (17)
ϕ, R	the function and the parameter related to the 'regularity' of f_H (see Assumption 2)
γ, c_γ	the parameters related to the effective dimension (see Assumption 3)
$\{\sigma_i\}_i$	the sequence of eigenvalues of $\mathcal{L}$
ψ, ϑ	the functions from Part 2 of Theorem 4.2, $\phi = \psi\vartheta$
$\mathcal{T}_\lambda$,	$\mathcal{T}_\lambda = \mathcal{T} + \lambda$
$\mathcal{T}_{\mathbf{x}\lambda}$,	$\mathcal{T}_{\mathbf{x}\lambda} = \mathcal{T}_{\mathbf{x}} + \lambda$
ζ	the parameter related to the Holder source condition on f_H (see (28))
$\mathcal{R}_\lambda(u)$	$= 1 - \mathcal{G}_\lambda(u)u$
$\mathcal{N}(\lambda)$	$= \mathrm{tr}(\mathcal{T}(\mathcal{T}+\lambda)^{-1})$
c_g	the constant from Lemma 5.1
ω_λ	the population vector defined by (32)
$a_{n,\delta,\gamma}(\theta)$	the quantity defined by (35)