Convergences of Regularized Algorithms and Stochastic Gradient Methods with Random Projections

Junhong Lin
Junhong@zju.edu.cn
Center for Data Science, Zhejiang University,
310027, Hang Zhou, P. R. China.
Laboratory for Information and Inference Systems,
École Polytechnique Fédérale de Lausanne,
CH1015-Lausanne, Switzerland.

Volkan Cevher
Volkan.cevher@epfl.ch
Laboratory for Information and Inference Systems,
École Polytechnique Fédérale de Lausanne,
CH1015-Lausanne, Switzerland.

Editor: John Shawe-Taylor

Abstract

We study the least-squares regression problem over a Hilbert space, covering nonparametric regression over a reproducing kernel Hilbert space as a special case. We first investigate regularized algorithms adapted to a projection operator on a closed subspace of the Hilbert space. We prove convergence results with respect to variants of norms, under a capacity assumption on the hypothesis space and a regularity condition on the target function. As a result, we obtain optimal rates for regularized algorithms with randomized sketches, provided that the sketch dimension is proportional to the effective dimension up to a logarithmic factor. As a byproduct, we obtain similar results for Nyström regularized algorithms. Our results provide optimal, distribution-dependent rates that do not have any saturation effect for sketched/Nyström regularized algorithms, considering both the attainable and non-attainable cases, in the well-conditioned regimes. We then study stochastic gradient methods with projection over the subspace, allowing multi-pass over the data and minibatches, and we derive similar optimal statistical convergence results.

Keywords: kernel methods, regularized algorithms, stochastic gradient methods, random projection, sketching

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 $\rho$ be an unknown probability measure on $H \times \mathbb{R}$. In this paper, we study the following expected risk minimization,

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

where the measure $\rho$ is known only through a sample $z = \{z_i = (x_i, y_i)\}_{i=1}^n$ of size $n \in \mathbb{N}$, independently and identically distributed (i.i.d.) according to $\rho$. 
The above regression setting covers nonparametric regression over a reproducing kernel Hilbert space (RKHS) (Shawe-Taylor and Cristianini, 2004; Cucker and Zhou, 2007; Steinwart and Christmann, 2008), and it is close to functional regression (Ramsay, 2006) and linear inverse problems (Engl et al., 1996). A basic algorithm for the problem is ridge regression, and its generalization, spectral algorithm. Such algorithms can be viewed as solving an empirical, linear equation with the empirical covariance operator replaced by a regularized one, see (Caponnetto and Yao, 2006; Bauer et al., 2007; Gerfo et al., 2008; Lin et al., 2018) and the references therein. Here, the regularization is used to control the complexity of the solution to avoid over-fitting and to achieve the best possible generalization ability.

The function/estimator generated by classic regularized algorithm is in the subspace \( \text{span}\{\mathbf{x}\} \) of \( H \), where \( \mathbf{x} = \{x_1, \ldots, x_n\} \). More often, the search of an estimator for some specific algorithms is restricted to a different (and possibly smaller) subspace \( S \), which leads to regularized algorithms with projection. Typically, with a subsample/sketch dimension \( m < n \), \( S = \text{span}\{\tilde{x}_j : 1 \leq j \leq m\} \) where \( \tilde{x}_j \) is chosen randomly from the input set \( \mathbf{x} \), and more generally, \( S = \text{span}\{\sum_{j=1}^{n} G_{ij} \tilde{x}_j : 1 \leq i \leq m\} \) where \( G = [G_{ij}]_{1 \leq i \leq m, 1 \leq j \leq n} \) is a general randomized matrix whose rows are drawn according to a distribution. We call the resulting algorithms the Nyström regularized algorithm and the sketched-regularized algorithm, respectively. Such approaches have been shown to achieve some computational advantages for ridge regression over an RKHS, leading to solutions that use the low-rank approximation in place of the full kernel matrix and thus is faster to compute (e.g., see Williams and Seeger, 2000; Kumar et al., 2009; Mahoney, 2011; Yang et al., 2012; Gittens and Mahoney, 2016; Yang et al., 2017; Rudi et al., 2015, and references therein).

Our starting points of this paper are the contemporary papers (Bach, 2013; Alaoui and Mahoney, 2015; Yang et al., 2017; Rudi et al., 2015; Myleiko et al., 2017) which study the convergences of Nyström/sketched regularized algorithms for learning with kernel methods. Particularly, within the fixed design setting, i.e., the input set \( \mathbf{x} \) are deterministic while the output set \( \mathbf{y} = \{y_1, \ldots, y_n\} \) treated randomly, convergence results have been derived, in (Bach, 2013; Alaoui and Mahoney, 2015) for Nyström ridge regression and in (Yang et al., 2017) for sketched ridge regression. Within the random design setting, which is more meaningful (Hsu et al., 2014) in statistical learning theory, and involving a regularity/smoothness condition on the target function (Smale and Zhou, 2007), optimal statistical results on generalization error bounds (excess risks) have been obtained in (Rudi et al., 2015) for Nyström ridge regression. The latter results were further generalized in (Myleiko et al., 2017) to a general Nyström regularized algorithm.

Although results have been developed for sketched ridge regression in the fixed design setting, it is still unclear if one can get statistical results for a general sketched-regularized algorithm in the random design setting. Besides, all the derived results, either for sketched or Nyström regularized algorithms, are only for the attainable case, i.e., the expected risk minimization (1) has at least one solution in \( H \). Moreover, they saturate (Bauer et al., 2007) at a critical value, meaning that they can not lead to better convergence rates even with a

1. The Nyström subsampling scheme corresponds to a sketched scheme with the rows of the sketch matrix \( G \) randomly chosen from the rows of an identity matrix. In this paper, by abuse of terminology, we sometimes use “sketched-regularized algorithm” to mean a sketched algorithm generated by Subgaussian sketches or randomized bounded orthogonal system sketches those will be introduced in Subsection 3.3.
smoother target function. Motivated by these, in this paper, we study statistical results of projected-regularized algorithms for least-squares regression over a separable Hilbert space within the random design setting.

We first extend the analysis in (Lin and Cevher, 2018b; Lin et al., 2018) for classic-regularized algorithms to projected-regularized algorithms, and prove statistical results with respect to a broader class of norms. We then show that the same convergence rates as classic regularized algorithms can be retained for sketched-regularized algorithms, provided that the sketch dimension is proportional to the effective dimension (Zhang, 2005) up to a logarithmic factor. As a byproduct, we obtain similar results for Nyström regularized algorithms.

Interestingly, our results provide optimal, distribution-dependent rates that do not have any saturation effect for sketched/Nyström regularized algorithms in the well-conditioned regimes, considering both the attainable and non-attainable cases. In our proof, we naturally integrate proof techniques from (Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Rudi et al., 2015; Myleiko et al., 2017; Lin and Cevher, 2018b). Our novelties lie in a new estimate on the projection error for sketched-regularized algorithms, a novel analysis to conquer the saturation effect, and a refined analysis for Nyström regularized algorithms, see Section 5 for details.

Our proof techniques can be used to analyze stochastic gradient methods (SGM) adapted to the projection operator over the subspace $S$. Indeed, for classical non-projected multi-pass SGM where a minibatch of sample points are selected randomly with replacement from $z$ at each iteration, it has been shown in (Lin and Rosasco, 2017b; Lin and Cevher, 2018b) that one can approximate SGM via regularized algorithms, as the conditional expectation of SGM given $z$ is the batch gradient descent (Lin and Rosasco, 2017b), a special regularized algorithm. The regularization effect of the number of iterations and statistical results for classic multi-pass SGM have been unveiled in (Lin and Rosasco, 2017b). Besides, SGM has been successfully combined with Nyström subsampling and its computational advantage when considering mini-batches has been shown in (Lin and Rosasco, 2017a). Optimal statistical results on generalization error bounds have been shown for Nyström SGM in (Lin and Rosasco, 2017a), but only for the attainable cases.

In this paper, we provide statistical results on variants of norms with optimal rates for sketched/Nyström SGM in the well-conditioned regimes, considering both the attainable and non-attainable cases.

This paper is an extension of the conference version (Lin and Cevher, 2018a). In this paper, we provide convergence results in $H$-norm for sketched/Nyström regularized algorithms, results for sketched/Nyström SGM, and explicit constants in the error bounds depending on noise variance and bias from Proposition 4, which have not been given in (Lin and Cevher, 2018a). In (Lin and Cevher, 2018a), we give results for Nyström regularized algorithms, considering only the plain Nyström subsampling with uniform sampling regime. In this paper, we provide results for Nyström regularized algorithms, using alternative non-uniform sampling scheme—the approximate leverage scores (ALS) Nyström subsampling, see Subsection 3.4 for the details.

The rest of the paper is organized as follows. Section 2 introduces some auxiliary notations and assumptions from standard statistical learning. Section 3 presents projected-regularized algorithms and their convergence results, followed with simple discussions. Sec-
tion 4 provides projected-SGM algorithms and their convergence results. Finally, Sections 5, 6 and the appendix supplement the proofs of our main results.

2. Notations and Assumptions

In this section, we first introduce the needed notation as well as the key auxiliary operators. We then present assumptions from standard statistical learning.

2.1. Notations and Auxiliary Operators

Let $Z = H \times \mathbb{R}$, $\rho_X(\cdot)$ the induced marginal measure on $H$ of $\rho$, and let $\rho(\cdot|x)$ be the conditional probability measure on $\mathbb{R}$ with respect to $x \in H$ and $\rho$. For simplicity, we assume that the support of $\rho_X$ is compact and that there exists a constant $\kappa \in [1, \infty]$, such that

$$\langle x, x' \rangle_H \leq \kappa^2, \quad \forall x, x' \in H, \rho_X\text{-almost surely.} \quad (2)$$

Define the hypothesis space

$$H_\rho = \{ f : H \to \mathbb{R} | \exists \omega \in H \text{ with } f(x) = \langle \omega, x \rangle_H, \rho_X\text{-almost surely} \}.$$ 

Denote $L^2_{\rho_X}$ the Hilbert space of square integral functions from $H$ to $\mathbb{R}$ with respect to $\rho_X$, with its norm given by $\|f\|_\rho = \left( \int_H |f(x)|^2 d\rho_X(x) \right)^{\frac{1}{2}}$.

For a given bounded operator $L$ from a Hilbert space $H_1$ to a Hilbert space $H_2$, $\|L\|$ denotes the operator norm of $L$, i.e., $\|L\| = \sup_{f \in H_1, \|f\|_{H_1} = 1} \|Lf\|_{H_2}$. Let $r \in \mathbb{N}_+$, the set $\{1, \ldots, r\}$ is denoted by $[r]$. For any real number $a$, $a_+ = \max(a, 0)$, $a_- = \min(0, a)$.

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

$$\|L\| = \|T\| \leq \text{tr}(T) = \int_H \text{tr}(x \otimes x) d\rho_X(x) = \int_H \|x\|^2_H d\rho_X(x) \leq \kappa^2. \quad (3)$$

For any $\omega \in H$, it is easy to prove the following isometry property (Bauer et al., 2007):

$$\|S_\rho \omega\|_\rho = \|T^{\frac{1}{2}} \omega\|_H. \quad (4)$$

Moreover, according to the singular value decomposition of a compact operator, one can prove that

$$\|L^{-\frac{1}{2}} S_\rho \omega\|_\rho \leq \|\omega\|_H. \quad (5)$$

We define the (modified) sampling operator $S_X : H \to \mathbb{R}^n$ by $(S_X \omega)_i = \frac{1}{\sqrt{n}} \langle \omega, x_i \rangle_H$, $i \in [n]$, where the norm $\|\cdot\|_2$ in $\mathbb{R}^n$ is the usual Euclidean norm. Its adjoint operator $S_X^* : \mathbb{R}^n \to H$, defined by $(S_X^* \omega, \cdot)_H = \langle y, S_X \omega \rangle_2$ for $y \in \mathbb{R}^n$, is thus given by $S_X^* \omega = \frac{1}{\sqrt{n}} \sum_{i=1}^n y_i x_i$. For notational simplicity, we let $\hat{y} = \frac{1}{\sqrt{|Y|}} y$. Moreover, we can define the empirical covariance
operator $\mathcal{T}_x : H \to H$ such that $\mathcal{T}_x = S_x^* S_x$. Obviously, $\mathcal{T}_x = \frac{1}{n} \sum_{i=1}^{n} \langle \cdot, x_i \rangle_H x_i$. By Assumption (2), similar to (3), we have $\|\mathcal{T}_x\| \leq \text{tr}(\mathcal{T}_x) \leq \kappa^2$. (6)

It is easy to see that Problem (1) is equivalent to

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

(7)

The function that minimizes the expected risk over all measurable functions is the regression function (Cucker and Zhou, 2007; Steinwart and Christmann, 2008), defined as,

$$f_\rho(x) = \int_{\mathbb{R}} yd\rho(y|x), \quad x \in H, \rho_X\text{-almost surely.}$$

(8)

A simple calculation shows that the following well-known fact holds (Cucker and Zhou, 2007; Steinwart and Christmann, 2008), for all $f \in L^2_{\rho_X}$,

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

Then it is easy to see that (7) is equivalent to $\inf_{f \in H_\rho} \|f - f_\rho\|_{\rho}^2$. Under Assumption (2), $H_\rho$ is a subspace of $L^2_{\rho_X}$. Using the projection theorem, one can prove that a solution $f_H$ for the problem (7) is the projection of the regression function $f_\rho$ onto the closure of $H_\rho$ in $L^2_{\rho_X}$, and moreover, for all $f \in H_\rho$ (Lin and Rosasco, 2017b),

$$S_\rho^* f_\rho = S_\rho^* f_H,$$

(9)

and

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

(10)

Note that $f_H$ does not necessarily lie in $H_\rho$.

Throughout this paper, $S$ is a closed, finite-dimensional subspace of $H$, and $P$ is the projection operator onto $S$ or $P = I$.

2.2. Assumptions

In this subsection, we introduce three standard assumptions in statistical learning theory (Steinwart and Christmann, 2008; Cucker and Zhou, 2007). The first assumption relates to a moment condition on the noise $y - f_\rho(x)$.

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

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

(11)

and $|f_\rho(x)| \leq M$, $\rho_X$-almost surely.

Typically, the above assumption is satisfied if $y$ is bounded almost surely, or if $y = \langle \omega_s, x \rangle_H + \epsilon$, where $\epsilon$ is a Gaussian random variable with zero mean and it is independent from $x$.

The next assumption relates to the regularity/smoothness of the target function $f_H$.  


Assumption 2 $f_H$ satisfies
\[
\int_H (f_H(x) - f_\rho(x))^2 x \otimes x d\rho_X(x) \leq B^2 T,
\]
and the following Hölder source condition
\[
f_H = \mathcal{L}^\zeta g_0, \quad \text{with} \quad \|g_0\|_\rho \leq R.
\]
Here, $B, R, \zeta$ are non-negative numbers.

Condition (12) is trivially satisfied if $f_H - f_\rho$ is bounded almost surely. Moreover, when making a consistency assumption, i.e., $\inf_{H_\rho} \mathcal{E} = \mathcal{E}(f_\rho)$, as that in (Smale and Zhou, 2007, Caponnetto, 2006. Steinwart et al., 2009), for kernel-based non-parametric regression, it is satisfied\(^2\) with $B = 0$. Condition (13) characterizes the regularity of the target function $f_H$ (Smale and Zhou, 2007). A bigger $\zeta$ corresponds to a higher regularity and a stronger assumption, and it can lead to a faster convergence rate. Particularly, when $\zeta \geq 1/2$, $f_H \in H_\rho$ (Steinwart and Christmann, 2008). This means that the expected risk minimization (1) has at least one solution in $H$, which is referred to the attainable case. In this case, we let
\[
\omega_H = \mathcal{T}^{\zeta - 1} S_\rho^* g_0.
\]
Using the singular value decomposition of $S_\rho$, one can prove that $S_\rho \omega_H = f_H$.

Finally, the last assumption relates to the capacity of the space $H (H_\rho)$.

Assumption 3 For some $\gamma \in [0, 1]$ and $c_\gamma > 0$, $\mathcal{T}$ satisfies
\[
\mathcal{N}(\lambda) := \operatorname{tr}(\mathcal{T}(\mathcal{T} + \lambda I)^{-1}) \leq c_\gamma \lambda^{-\gamma}, \quad \text{for all} \ \lambda > 0.
\]
The left hand-side of (14) is called degrees of freedom (Zhang, 2005), or effective dimension (Caponnetto and De Vito, 2007). Assumption 3 is always true for $\gamma = 1$ and $c_\gamma = \kappa^2$, since $\mathcal{T}$ is a trace class operator. This is referred to 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.

3. Projected-regularized Algorithms

In this section, we first demonstrate and introduce the projected-regularized algorithms. We then present theoretical results for the projected-regularized algorithms. Finally, we give results for the sketched/Nyström regularized algorithms.

3.1. Projected-regularized Algorithms

The expected risk $\tilde{\mathcal{E}}(\omega)$ in (1) can not be computed exactly. It can be only approximated through the empirical risk $\tilde{\mathcal{E}}_\mathbf{x}(\omega)$,\(^2\)
\[
\tilde{\mathcal{E}}_\mathbf{x}(\omega) = \frac{1}{n} \sum_{i=1}^n (\langle \omega, x_i \rangle_H - y_i)^2.
\]
A first idea to deal with the problem is to replace the objective function in (1) with the empirical risk. Moreover, we restrict the solution to the subspace $S$. This leads to the projected empirical risk minimization, $\inf_{\omega \in S} \tilde{E}_{\omega}$. Using $P^2 = P$, a simple calculation shows that a solution for the above could be $\hat{\omega} = P\hat{\alpha}$, with $\hat{\alpha}$ satisfying $PT_xP\hat{\alpha} = PS_x^T\bar{y}$. The inversion of the linear operator $PT_xP$ may have a bad condition number or be unbounded.

Motivated by the classic (iterated) ridge regression, we replace the inversion of $PT_xP$ with a regularized one, which leads to the following projected (iterated) ridge regression we study throughout this paper.

**Algorithm 1** The projected (iterated) ridge regression algorithm of order $\tau$ over the sample $z$ and subspace $S$ is given by $f_\lambda^k = S_\rho^k \omega_\lambda^k$, where

$$\omega_\lambda^k = PG_\lambda(PT_xP)PS_x^T\bar{y}, \quad G_\lambda(u) = \sum_{i=1}^{\tau} \lambda^{i-1}(\lambda + u)^{-i}.$$  

**Remark 1** 1) Our results not only hold for projected ridge regression, but also hold for a general projected-regularized algorithm, in which $G_\lambda$ is a general filter function. Given $\Lambda \subset \mathbb{R}_+$, a class of functions $\{G_\lambda : [0, \kappa^2] \rightarrow [0, \infty], \lambda \in \Lambda\}$ are called filter functions with qualification $\tau$ ($\tau \geq 1$) if there exist some positive constants $E, F < \infty$ such that

$$\sup_{\lambda \in \Lambda} \sup_{u \in [0, \kappa^2]} |G_\lambda(u)(u + \lambda)| \leq E.$$  

and

$$\sup_{\alpha \in [0, \tau]} \sup_{\lambda \in \Lambda} \sup_{u \in [0, \kappa^2]} |1 - G_\lambda(u)u|(u + \lambda)^{\alpha} \lambda^{-\alpha} \leq F.$$  

The filter function $G_\lambda(u)$ is an approximation of the inverse function. It is often used in dealing with ill-posed inverse problems. We refer to (Caponnetto and Yao, 2006; Bauer et al., 2007; Gerfo et al., 2008) and references therein for further details about the filter functions.

2) A simple calculation shows that

$$G_\lambda(u) = \frac{1 - q^\tau}{u} = \frac{\sum_{i=0}^{\tau-1} q^i}{u + \lambda}, \quad q = \frac{\lambda}{\lambda + u}.$$  

Thus, $G_\lambda(u)$ is a filter function with qualification $\tau$, $E = \tau$ and $F = 1$. When $\tau = 1$, it is a filter function for the classic ridge regression and the algorithm is the projected ridge regression algorithm.

3) Another typical filter function studied in the literature is

$$G_\lambda(u) = \begin{cases} u^{-1}, & \text{if } u \geq \lambda, \\ 0, & \text{otherwise}, \end{cases}$$

which corresponds to principal component (spectral cut-off) regularization. Here, $1_{\{\cdot\}}$ denotes the indication function. In this case, $E = 2$, $F = 2^\tau$ and $\tau$ could be any positive

---

3. Let $L$ be a self-adjoint, compact operator over a separable Hilbert space $H$. $G_\lambda(L)$ is an operator on $L$ defined by spectral calculus: Suppose that $\{(\sigma_i, \psi_i)\}$ is a set of normalized eigenpairs of $L$ with the eigenfunctions $\{\psi_i\}$, forming an orthonormal basis of $H$, then we have $G_\lambda(L) = \sum_i G_\lambda(\sigma_i)\psi_i \otimes \psi_i$. 
number.

4) The choice $G_{\lambda}(u) = \sum_{k=1}^{l} \eta(1 - \eta u)^{l-k}$ with $\eta \in [0, \kappa^{-2}]$ where we identify $\lambda = (\eta t)^{-1}$, corresponds to gradient methods or the Landweber iteration algorithm. The qualification $\tau$ could be any positive number, $E = 2$, and $F = F_{\tau} = \tau^\tau \exp(1 - \tau)$.

In the above, $\lambda$ is a regularization parameter which needs to be well chosen in order to achieve the best possible performance. Throughout this paper, we assume that $1/n \leq \lambda \leq 1$.

The performance of an estimator $f^{x}_{\lambda}$ can be measured in terms of excess risk (generalization error), $E(f^{x}_{\lambda}) - \inf_{H, \rho} E = \bar{E}(\omega^{x}_{\lambda}) - \inf_{H} \bar{E}$, which is exactly $\|f^{x}_{\lambda} - f_{H}\|_{\rho}^2$ according to (10). Assuming that $f_{H} \in H_{\rho}$, i.e., $f_{H} = S_{\rho} \omega_{s}$ for some $\omega_{s} \in H$, it can be measured in terms of $H$-norm, $\|\omega^{x}_{\lambda} - \omega_{s}\|_{H}$, which is closely related to $\|\mathcal{L}^{-\frac{1}{2}} S_{\rho}(\omega^{x}_{\lambda} - \omega_{s})\|_{H} = \|\mathcal{L}^{-\frac{1}{2}} (f^{x}_{\lambda} - f_{H})\|_{\rho}$, according to (5). In what follows, we will measure the performance of an estimator $f^{x}_{\lambda}$ in terms of a broader class of norms, $\|\mathcal{L}^{-a}(f^{x}_{\lambda} - f_{H})\|_{\rho}$, where $a \in [0, 1]$ is such that $\mathcal{L}^{-a} f_{H}$ is well defined. In the attainable cases, i.e., $f_{H} \in H_{\rho}$, according to (5), $\|\mathcal{L}^{-a}(f^{x}_{\lambda} - f_{H})\|_{\rho}$ is close to $\|T^{\frac{1}{2} - a}(\omega^{x}_{\lambda} - \omega_{H})\|_{H}$. Convergence with respect to different norms is of strong interest in convex optimization, inverse problems, and statistical learning theory. Particularly, convergence with respect to target function values and the $H$-norm has been studied in convex optimization. Interestingly, the convergence in the $H$-norm can imply the convergence in target function values (although the derived rate is not optimal), while the opposite is not true in general.

3.2. General Results for Projected-regularized Algorithms

We now state our first result as follows. In the sequel, $C$ denotes a positive constant that depends only on $\kappa^{2}, c_{0}, \gamma, \lambda, B, M, Q, R, \tau$ and $\|T\|$, and it could be different at its each appearance. Moreover, we write $a_{1} \lesssim a_{2}$ to mean $a_{1} \leq C a_{2}$.

**Theorem 1** Under Assumptions 1, 2 and 3, let $\lambda = n^{-\theta}$ for some $\theta \in [0, 1)$ or $\lambda = \frac{\ln n \gamma}{n}$.

Let $a \in [0, \frac{1}{2} \wedge \zeta]$, and $\tau \geq \zeta - a$. Then the following holds with probability at least $1 - \delta$ ($0 < \delta < 1$).

1) If $\zeta \in [0, 1]$,

$$\|\mathcal{L}^{-a}(f^{x}_{\lambda} - f_{H})\|_{\rho} \lesssim \lambda^{-a} \log^{2} \frac{3}{\delta} \left( \lambda^{\zeta} + \frac{1}{\sqrt{n} \lambda^{\zeta}} + \lambda^{\zeta - 1} (\Delta_{5} + \Delta_{5}^{1-a} \lambda^{a}) \right).$$

(20)

2) If $\zeta \geq 1$ and $\lambda \geq n^{-1/2}$,

$$\|\mathcal{L}^{-a}(f^{x}_{\lambda} - f_{H})\|_{\rho} \lesssim \lambda^{-a} \log^{2} \frac{3}{\delta} \left( \lambda^{\zeta} + \frac{1}{\sqrt{n} \lambda^{\zeta}} + (\Delta_{5} + \lambda \Delta_{5}^{(\zeta-1)\lambda^{a}} + \Delta_{5}^{1-a} \lambda^{a}) \right).$$

(21)

Furthermore, if $\zeta \geq 1/2$, then the above conclusions still hold if we replace $\|\mathcal{L}^{-a}(f^{x}_{\lambda} - f_{H})\|_{\rho}$ by $\|T^{\frac{1}{2} - a}(\omega^{x}_{\lambda} - \omega_{H})\|_{H}$. Here, $\Delta_{5}$ is the projection error $\| (I - P) T^{\frac{1}{2}} \|^{2}$.

The above result provides high-probability error bounds with respect to variants of norms for projected-regularized algorithms. The upper bound consists of three terms. The first term depends on the regularity parameter $\zeta$, and it arises from estimating the bias. The second term depends on the sample size, and it arises from estimating the variance. The
third term depends on the projection error. Note that there is a trade-off among the bias term, the variance term, and the projection-error term. Ignoring the projection error, solving the trade-off between the bias and variance terms leads to the best choice on \( \lambda \) and the following result.

**Corollary 1** Under the assumptions and notations of Theorem 1, let
\[
\lambda = n^{-\frac{1}{\tau(2\zeta + \gamma)}(1 \lor \log n^{\gamma})^{1 \lor (2\zeta + \gamma) \leq 1}}.
\]
Then the following statements hold with probability at least \( 1 - \delta \).

1) If \( \zeta \leq 1 \),
\[
\| \mathcal{L}^{-a}(f_{X}^{a} - f_{H}) \|_{\rho} \lesssim \lambda^{-a} (1 + \lambda^{-1} \Delta_{5}) \log^{2} \frac{3}{\delta}.
\]

2) If \( \zeta \geq 1 \),
\[
\| \mathcal{L}^{-a}(f_{X}^{a} - f_{H}) \|_{\rho} \lesssim \lambda^{-a} \log^{2} \frac{3}{\delta} \left( \lambda^{\zeta} + \Delta_{5} \left( 1 + \left( \frac{\lambda}{\Delta_{5}} \right) \Delta_{5}^{(\zeta-1)\lambda} + \left( \frac{\lambda}{\Delta_{5}} \right)^{a} \right) \right).
\]

Furthermore, if \( \zeta \geq 1/2 \), then the above conclusions still hold if we replace \( \| \mathcal{L}^{-a}(f_{X}^{a} - f_{H}) \|_{\rho} \) by \( \| T^{\frac{1}{2} - a}(\omega_{X}^{a} - \omega_{H}) \|_{H} \).

Comparing the derived upper bound for projected-regularized algorithms with that for classic regularized algorithms in (Lin et al., 2018), we see that the former has an extra term, which is caused by projection. The above result asserts that projected-regularized algorithms perform similarly as classic regularized algorithms if the projection operator is well chosen such that the projection error is small enough.

In the special case that \( P = I \), we get the follow result.

**Corollary 2** Under the assumptions and notations of Theorem 1, let \( \lambda \) be given by (22) and \( P = I \). Then with probability at least \( 1 - \delta \),
\[
\| \mathcal{L}^{-a}(f_{X}^{a} - f_{H}) \|_{\rho} \lesssim \log^{2} \frac{3}{\delta} \left( \frac{1}{n^{\gamma}} \right)^{\frac{\zeta-a}{\zeta-a}} \left( \frac{n^{\gamma}}{n^{-\zeta-a}} \right)^{\frac{\zeta-a}{\zeta-a}} \left( \frac{n}{n^{-\zeta-a}} \right)^{\frac{\zeta-a}{\zeta-a}}, \quad \text{if } 2\zeta + \gamma \leq 1,
\]
\[
\| \mathcal{T}^{1/2 - a}(\omega_{X}^{a} - \omega_{H}) \|_{H} \lesssim \log^{2} \frac{3}{\delta} n^{-\frac{\zeta-a}{\zeta-a}}.
\]

The rate from the above with \( 2\zeta + \gamma \leq 1 \) improves the rate \( O(n^{\zeta-a} \lor (1 \lor \log n^{\gamma})^{1-a}) \) derived in (Lin et al. 2018). The convergence rates for \( 2\zeta + \gamma > 1 \) have already been given in the literature, see (Lin et al., 2018) and some of the references therein. They are optimal as they match the minimax rates summarized in Table 1. See (Caponnetto and De Vito, 2007; Steinwart et al., 2009; Blanchard and Mücke, 2018; Fischer and Steinwart, 2017) for further details about minimax rates.

**Remark 2** Corollary 2 provides convergence results in high probability for the studied algorithms. As remarked in (Lin et al., 2018), it implies convergence in expectation and almost sure convergence.
Assumptions

\[ a = 0, \zeta \in \left[\frac{1}{2}, 1\right] \]
\[ N^{-\frac{2\zeta+\gamma}{2}} \]

\[ a = 0, \zeta \in (0, \frac{1}{2}] \]
\[ N^{-\frac{2\zeta+\gamma}{2}} \]

\[ a \in [0, \frac{1}{2}], \zeta \geq \frac{1}{2} \]
\[ N^{-\frac{2(\zeta-\gamma)}{2}} \]

Table 1: Minimax Rates on \( \|L^{-a}(f_z - f_H)\|_\rho^2 \)

### 3.3. Results for Sketched-regularized Algorithms

In this subsection, we state results for sketched-regularized algorithms.

In sketched-regularized algorithms, the range of the projection operator \( P \) is the subspace \( \text{range}\{S^*_xG^*\} \), where \( G \in \mathbb{R}^{m \times n} \) is a sketch matrix with \( m < n \) satisfying the following concentration inequality: For any finite subset \( E \) in \( \mathbb{R}^n \) and for any \( t \in (0, 1) \),

\[
P\left(\|Ga\|_2^2 - \|a\|_2^2 \geq t\|a\|_2^2 : \exists a \in E\right) \leq 2|E|e^{-\frac{t^2m}{c_0' \log^3 n}}. \tag{26}
\]

Here, \( c_0' \) is a universal positive constant and \( \beta \) is a universal non-negative constant. Many matrices satisfy the concentration property.

- **Subgaussian sketches.** Matrices with i.i.d. Subgaussian (such as Gaussian or Bernoulli) entries satisfy (26) with some universal constant \( c_0' \) and \( \beta = 0 \). More generally, if the rows of \( G \) are independent (scaled) copies of an isotropic \( \psi_2 \) vector, then \( G \) also satisfies (26) (Mendelson et al., 2008).

- **Randomized orthogonal system (ROS) sketches.** As noted in (Krahmer and Ward, 2011), matrix that satisfies restricted isometric property from compressed sensing with randomized column signs satisfies (26). Particularly, random partial Fourier matrix, or random partial Hadamard matrix with randomized column signs satisfies (26) with \( \beta = 4 \) for some universal constant \( c_0' \). Using OS sketches has an advantage in computation, as that for suitably chosen orthonormal matrices such as the DFT and Hadamard matrices, a matrix-vector product can be executed in \( O(n \log m) \) time, in contrast to \( O(nm) \) time required for the same operation with generic dense sketches.

The following corollary shows that sketched-regularized algorithms have optimal rates provided the sketch dimension \( m \) is not too small.

**Corollary 3** Under the assumptions and notations of Theorem 1, let \( S = \text{range}\{S^*_xG^*\} \), where \( G \in \mathbb{R}^{m \times n} \) is a randomized matrix satisfying (26). Let

\[
m \gtrless \log^3 n \log^3 \frac{3}{\delta} \begin{cases} \frac{n^7}{(1 \log n)^{2(\zeta-\gamma)}}, & \text{if } 2\zeta + \gamma \leq 1, \\ \frac{1}{n^{1-(1-a)(\zeta-\gamma)}}, & \text{if } \zeta \geq 1, \\ \frac{\zeta-a}{n^{2\zeta+\gamma}}, & \text{otherwise}. \end{cases} \tag{27}
\]

Then with confidence at least \( 1 - \delta \), the following holds

\[
\|L^{-a}(f_z - f_H)\|_\rho \lesssim \log^3 \frac{3}{\delta} \begin{cases} \left(\frac{1}{1 \log n}\right)^{\zeta-a}, & \text{if } 2\zeta + \gamma \leq 1, \\ \frac{\zeta-a}{n^{2\zeta+\gamma}}, & \text{if } 2\zeta + \gamma > 1. \end{cases} \tag{28}
\]
Furthermore, if $\zeta \geq 1/2$,

$$\|T^{1/2-a}(\omega_{\lambda}^{Z} - \omega_{H})\|_{H} \lesssim \log^{3} \frac{3}{\delta} n^{-\frac{\zeta}{2}+\gamma}.$$ 

The above results assert that sketched-regularized algorithms converge optimally, provided the sketch dimension is not too small, or in another words the error caused by projection is negligible when the sketch dimension is large enough. Ignoring the logarithmic factors, the minimal sketch dimension from the above is at most $Cn$, and it is smaller than $Cn$ when the regularity parameter $\zeta$ is large or the effective-dimensional parameter $\gamma$ is small. Furthermore, the minimal sketch dimension is proportional to the effective dimension $\lambda^{-\gamma}$ up to a logarithmic factor for the case $\zeta \leq 1$.

**Remark 3**
1) Considering only the case $\zeta = 1/2$ and $a = 0$, Yang et al. (2017) provide optimal error bounds for sketched ridge regression within the fixed design setting.
2) Wang et al. (2017) provide error estimates on the target function values (i.e., the regularized empirical risks) for sketched ridge regression over a finite-dimensional space in the fixed design setting, and they also show a similar bias-variance trade-off phenomenon when choosing the optimal regularization parameter for the algorithm.

Corollary 3 is proved by applying Corollary 1, combing with an estimate on the projection error developed in Subsection 5.5. As we mentioned before, the Nyström regularized algorithm can be viewed as a projected-regularized algorithm with the projection operator $P$ being the subspace range$\{S^{*}_{\lambda}G^{*}\}$, where $G \in \mathbb{R}^{m \times n}$ is a sketch matrix with rows drawn randomly from an identity matrix. However, for the latter case, in general, we need alternative arguments for estimating the projection error.

**3.4. Results for Nyström Regularized Algorithms**

As a byproduct of the paper, using Corollary 1 and an estimation on the projection error, we derive the following results for Nyström regularized algorithms.

**Corollary 4** Under the assumptions and notations of Theorem 1, let $S = \text{span}\{x_{1}, \cdots, x_{m}\}$, $2\zeta + \gamma > 1$, and $\lambda = n^{-\frac{1}{2\zeta+\gamma}}$. If

$$m \gtrsim (1 + \log n^{\gamma}) \begin{cases} \frac{n^{-\frac{\zeta-a}{1-\alpha}(2\zeta+\gamma)}}{\frac{1}{n^{2\zeta+\gamma}}} & \text{if } \zeta \geq 1, \\ \frac{n^{\frac{1}{2\zeta+\gamma}}}{} & \text{if } \zeta \leq 1, \end{cases}$$

then the conclusions in Corollary 3 are true.

**Remark 4**
1) Considering only the case $1/2 \leq \zeta \leq 1$ and $a = 0$, (Rudi et al., 2015) provides optimal generalization error bounds for Nyström ridge regression. This result was further extended in (Myleiko et al., 2017) to a general Nyström regularized algorithm with a general source assumption indexed with an operator monotone function (but only in the attainable cases). Note that as in classic ridge regression, Nyström ridge regression saturates over $\zeta \geq 1$, i.e., it does not have a better rate even for a bigger $\zeta \geq 1$.
2) For the case $\zeta \geq 1$ and $a = 0$, (Myleiko et al., 2017) provides certain generalization error
bounds for plain Nyström regularized algorithms, but the rates are capacity-independent, and the minimal projection dimension $O(n^{2\sqrt{\frac{2}{\gamma+1}}})$ is larger than ours (considering the case $\gamma = 1$ for the sake of fairness).

In the above lemma, we consider the plain Nyström subsampling. Using the ALS Nyström subsampling (Drineas et al., 2012; Gittens and Mahoney, 2016; Alaoui and Mahoney, 2015), we can improve the projection dimension condition to (27).

**ALS Nyström subsampling** Let $K = S_xS_x^\ast$. For $\lambda > 0$, the leveraging scores of $K(K + \lambda I)$ is the set \{\(l_i(\lambda)\)\}$_{i=1}^n$ with

$$l_i(\lambda) = (K(K + \lambda I)^{-1})_{ii}, \quad \forall i \in [n].$$

The $L$-approximated leveraging scores (ALS) of $K(K + \lambda I)$ is a set \{\(\hat{l}_i(\lambda)\)\}$_{i=1}^n$ satisfying

$$L^{-1}l_i(\lambda) \leq \hat{l}_i(\lambda) \leq Ll_i(\lambda),$$

for some $L \geq 1$. In an $(L, \lambda)$-ALS Nyström subsampling regime, $S = \text{range}\{\tilde{x}_1, \ldots, \tilde{x}_m\}$, where each $\tilde{x}_j$ is i.i.d. drawn according to

$$\mathbb{P}(\tilde{x} = x_i) \sim \hat{l}_i(\lambda).$$

The $i$-th leveraging score $l_i(\lambda)$ measures the “importance” of the $i$-th input $x_i$. In ALS Nyström scheme, the element corresponding with a higher score will be selected with a higher probability, which is different from the uniform selection in plain Nyström.

**Corollary 5** Under the assumptions of Theorem 1, let $\lambda = n^{-\frac{1}{(2\zeta+\gamma)\vert\gamma\vert}}$ and $S = \text{range}\{\tilde{x}_1, \ldots, \tilde{x}_m\}$ with $\tilde{x}_j$ drawn following an $(L, \lambda)$-ALS Nyström subsampling scheme. If

$$m \geq L^2 \log^3 \delta \begin{cases} n^\gamma [1 \lor \log n^\gamma]^{1-\gamma}, & \text{if } 2\zeta + \gamma \leq 1, \\ n^{\frac{\gamma(\zeta-\alpha)}{1-\zeta}} [1 \lor \log n^\gamma], & \text{if } \zeta \geq 1, \\ n^{\frac{\gamma}{2\zeta+\gamma}} [1 \lor \log n^\gamma], & \text{otherwise}, \end{cases}$$

then the conclusions in Corollary 3 are true.

### 4. Results for Projected Stochastic Gradient Method

In this section, we introduce stochastic gradient methods with projections (projected-SGM) and then state statistical results for the projected-SGM. As corollaries, we provide convergence results for the sketched/Nyström SGM methods.

SGM is one of the most popular and scalable algorithms for large-scale learning problems. We refer to (Lin and Rosasco, 2017b a) and references therein for further introductions on SGM. In this paper, we study the following projected-SGM, a variant of classic SGM considering an orthogonal projection operator.
Algorithm 2 The stochastic gradient method with projection is defined by \( \omega_1 = 0 \),

\[
\omega_{t+1} = \omega_t - \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} ((\omega_t, x_{j_i})_H - y_{j_i}) Px_{j_i}, \quad t = 1, \cdots, T,
\]

where \( \eta \) is a step-size, \( j_1, j_2, \cdots, j_{bt} \) are i.i.d. random variables from the uniform distribution on \( \{1, \cdots, n\} \), and \( b \in \mathbb{N}^+ \).

The step-size \( \eta \), the number of iterations \( T \), and the minibatch size \( b \), are free parameters in the above algorithm. They dictate the performance of the algorithm, as shown in our coming results.

The random variables \( j_1, \cdots, j_{bt} \) are conditionally independent given the sample \( z \). We write \( J = \{j_1, \cdots, j_{bt}\} \) and denote the conditional expectation with respect to \( J \) given \( z \) by \( \mathbb{E}_J \).

In order to state our results, we need to introduce the following assumption on the moment condition of \(|y|^2\).

Assumption 4 There exist constants \( M \in [0, \infty[ \) and \( Q \in [1, \infty[ \) such that

\[
\int_Y y^{2l} d\rho(y|x) \leq l! M^l Q, \quad \forall l \in \mathbb{N}, \tag{30}
\]

\( \rho_X \)-almost surely.

A simple calculation shows that the above assumption can imply Assumption 1. With this assumption, we have the following general results for projected-SGM.

Theorem 2 Under Assumptions 2, 3 and 4, let \( \delta \in (0, 1) \), and for some \( C'_1 \geq 1 \),

\[
\|(I - P)T^{\frac{1}{2}}\|^2 \leq C'_1 \lambda \frac{1 + \zeta - a}{\log \gamma} \log^2 \frac{2}{\delta}, \quad \lambda = n^{-\frac{1}{1 + (2\zeta + 1) \log n}} \tag{31}
\]

Consider Algorithm 2 with any of the following choices on \( \eta \), \( b \) and \( T \):

I) \( \eta \simeq \lambda^{2\zeta}, \quad b = 1 \) and \( T \simeq \lambda^{-(1 + 2\zeta)} \);

II) \( \eta \simeq (\log n)^{-1}, \quad b \simeq \lambda^{-2\zeta} \) and \( T \simeq \lambda^{-1} \log n \);

III) \( \eta \simeq n^{-1}, \quad b = 1 \) and \( T \simeq n \lambda^{-1} \);

IV) \( \eta \simeq n^{-1/2}, \quad b \simeq \sqrt{n} \) and \( T \simeq \sqrt{n} \lambda^{-1} \).

Then for any \( a \in [0, \frac{1}{2} \wedge \zeta] \), the following holds with probability at least \( 1 - \delta \).

1) If \( 2\zeta + \gamma \leq 1 \),

\[
\mathbb{E}_J \|L^{-a}(S_{\rho'} \omega_{T+1} - f_H)\|_\rho^2 \lesssim n^{-2(\zeta - a) \log n} (1 + \log n)^{1(2a \neq 1)} \log^{3} \frac{2}{\delta}. \tag{32}
\]

2) If \( 2\zeta + \gamma > 1 \),

\[
\mathbb{E}_J \|L^{-a}(S_{\rho'} \omega_{T+1} - f_H)\|_\rho^2 \lesssim n^{-2(\zeta - a) \log n} (1 + \log n)^{1(2a \neq 1)} \log^{3} \frac{2}{\delta}. \tag{33}
\]

Furthermore, if \( \zeta \geq 1/2 \), then the above conclusions still hold if we replace \( \|L^{-a}(S_{\rho'} \omega_{T+1} - f_H)\|_\rho \) by \( \|T^{\frac{1}{2} - a}(\omega_{T+1} - \omega_H)\|_H \).
The above results assert that with appropriate choices on the step-size and mini-batch size, if the projection error is small enough, the projected-SGM at some number of iterations performs optimally.

As direct corollaries, we have the following results for projected-SGM, considering specific projection operators as in Section 3.

**Corollary 6** Under the assumptions and notations of Theorem 2, if $P = I$, then the conclusions in Theorem 2 are true.

**Corollary 7** Under the assumptions and notations of Theorem 2, let $P$ and $m$ be as in Corollary 3/4/5, then the conclusions in Theorem 2 are true.

**Remark 5**
1) Similar results for classic (multi-pass) SGM were proved for $a = 0$ (Lin and Rosasco, 2017b; Lin and Cevher, 2018b) and $a = \frac{1}{2}$ (Lin and Rosasco, 2017b), where the derived rate $O(n^{-\frac{\alpha_1 - a}{2\zeta + \gamma}} \log^2 n)$ for $a = \frac{1}{2}$ from (Lin and Rosasco, 2017b) has an extra logarithmic factor in comparisons with our results.

2) Similar results with $a = 0$ for plain Nyström SGM were derived in (Lin and Rosasco, 2017a), but only for $\zeta \in [\frac{1}{2}, 1]$.

**Remark 6** Making an additional assumption on the so-called embedding property (Steinwart et al., 2009), optimal rates for the regime $2\zeta + \gamma \leq 1$ can be derived for ridge regression (Steinwart et al., 2009; Fischer and Steinwart, 2017) and multiple passes SGM with averaging (Pillaud-Vivien et al., 2018).

All the main results stated above will be proved in the remaining sections.

5. Proof for Section 3

In this section, we prove the results stated in Section 3. We first introduce some basic operator inequalities that are necessary for the proof in Subsection 5.1. We then give some deterministic estimates in Lemma 13, and with these basic operator inequalities and deterministic estimates, we prove a deterministically analytic result (i.e., Proposition 3) in Subsection 5.2. The analytic result involves three random quantities $\Delta_{1,2,3}$ and the projection error. The random quantities $\Delta_{1,2,3}$ will be estimated in Lemmas 14–16, see Subsection 5.3. Applying the probabilistic estimates on $\Delta_{1,2,3}$ from Lemmas 14–16 into Proposition 3, in Subsection 5.4 we prove the results (i.e., Theorem 1 and Corollary 1) for projected-regularized algorithms. We finally estimate the projection errors and use Corollary 1 to prove the results (i.e., Corollaries 3–5) for sketched-regularized and Nyström-regularized algorithms in Subsections 5.5–5.6.

5.1. Operator Inequalities

To proceed with the proof, we need to recall some basic operator inequalities, and we provide some of the proofs for completeness.

**Lemma 8** (Fujii et al., 1993) Let $A$ and $B$ be two positive bounded linear operators on a separable Hilbert space. Then

$$\|A^s B^s\| \leq \|AB\|^s, \quad \text{when } 0 \leq s \leq 1.$$
Lemma 9 Let $H_1, H_2$ be two separable Hilbert spaces and $S : H_1 \to H_2$ a compact operator. Then for any function $f : [0, ||S||] \to [0, \infty[$, 

$$f(SS^*)S = Sf(S^*S).$$

Proof The result can be proved using singular value decomposition of a compact operator.

Lemma 10 Let $A$ and $B$ be two non-negative bounded linear operators on a separable Hilbert space with $\max(||A||, ||B||) \leq \kappa^2$ for some non-negative $\kappa^2$. Then for any $\zeta > 0$,

$$\|A^\zeta - B^\zeta\| \leq C_{\zeta, \kappa} \|A - B\|^{\zeta \wedge 1},$$

where

$$C_{\zeta, \kappa} = \begin{cases} 1 & \text{when } \zeta \leq 1, \\ 2\zeta \kappa^{2\zeta - 2} & \text{when } \zeta > 1. \end{cases}$$

Proof The proof is based on the fact that $u^\zeta$ is operator monotone if $0 < \zeta \leq 1$. While for $\zeta \geq 1$, the proof can be found in, e.g., (Dicker et al., 2017).

Lemma 11 Let $X$ and $A$ be bounded linear operators on a separable Hilbert space $H$. Suppose that $A \succeq 0$ and $\|X\| \leq 1$. Then for any $\lambda \geq 0$, and any bounded linear operator $F$ on $H$,

$$\| (A + \lambda I)^{\frac{1}{2}}XF^*\| = \|FX^*(A + \lambda I)^{\frac{1}{2}}\| \leq \|F(X^*AX + \lambda I)^{\frac{1}{2}}\|.$$  

Proof Note that $X^*X \preceq I$ since $\|X\| \leq 1$. In fact, for any $\omega \in H$,

$$\langle X^*X\omega, \omega \rangle_H = \|X\omega\|^2_H \leq \|\omega\|^2_H = \langle \omega, \omega \rangle_H.$$

It thus follows that

$$X^*(A + \lambda I)X \preceq X^*AX + \lambda I.$$

Therefore,

$$\|FX^*(A + \lambda I)^{\frac{1}{2}}\|^2 = \|FX^*(A + \lambda I)XF^*\| \leq \|F(X^*AX + \lambda I)F^*\| = \|F(X^*AX + \lambda I)^{\frac{1}{2}}\|^2.$$

Lemma 12 Let $P$ be a projection operator in a Hilbert space $H$, and $A$, $B$ be two semidefinite positive operators on $H$. For any $0 \leq s, t \leq \frac{1}{2}$, we have

$$\|A^s(I - P)A^t\| \leq \|A - B\|^{s+t} + \|B^\frac{1}{2}(I - P)B^\frac{1}{2}\|^{s+t}.$$
**Proof** Since $P$ is a projection operator, $(I - P)^2 = I - P$. Then it holds that
\[
\|A^s(I - P)A^t\| = \|A^s(I - P)(I - P)A^t\| \leq \|A^s(I - P)\|\|I - P\|A^t\|.
\]
Moreover, by Lemma 8, we have
\[
\|A^s(I - P)\| = \|A^{\frac{1}{2}2s}(I - P)^{2s}\| \leq \|A^{\frac{1}{2}}(I - P)\|^{2s}.
\]
Similarly, $\|(I - P)A^t\| \leq \|(I - P)A^{\frac{1}{2}}\|^{2t}$. Thus, it follows that
\[
\|A^s(I - P)A^t\| \leq \|A^{\frac{1}{2}}(I - P)\|^{2s}\|I - P\|A^t\|\|I - P\|A^{\frac{1}{2}}\|^{2t} = \|(I - P)A^{\frac{1}{2}}\|^{2(t + s)}.
\]
Using $\|D\|^2 = \|D^*D\|$, we have
\[
\|A^s(I - P)A^t\| \leq \|(I - P)A(I - P)\|^{t + s}.
\]
Adding and subtracting with the same term, using the triangle inequality, and noting that $\|I - P\| \leq 1$ and $s + t \leq 1$,
\[
\|A^s(I - P)A^t\| \leq \|I - P\|A(I - P)\|^{t + s}
\]
\[
\leq \|(I - P)(A - B)(I - P)\| + \|(I - P)B(I - P)\|^{t + s}
\]
\[
\leq \|A - B\|^{s + t} + \|(I - P)B(I - P)\|^{s + t},
\]
which leads to the desired result using $\|D^*D\| = \|DD^*\|$.

## 5.2. Deterministic Estimates

In this subsection, we introduce some deterministic estimates. For notational simplicity, throughout this paper, we denote
\[
\mathcal{T}_\lambda = \mathcal{T} + \lambda I, \quad \mathcal{T}_x = \mathcal{T}_x + \lambda I.
\]
We also denote
\[
\mathcal{R}_\lambda(u) = 1 - \mathcal{G}_\lambda(u)u. \tag{37}
\]
For any $\lambda > 0$, we introduce a deterministic vector $\omega_\lambda^H$, defined by
\[
\omega_\lambda^H = \mathcal{G}_\lambda(\mathcal{T})S^*_\rho f_H, \tag{38}
\]
where $\mathcal{G}_\lambda(u)$ is given by (19). We have the following lemma for the properties of $\omega_\lambda^H$. We assume $\tau \geq \zeta - a$ throughout.

**Lemma 13** Under Assumption 2, the following holds.
1) For any $a \leq \zeta$, we have
\[
\|\mathcal{L}^{-a}(S_\rho \omega_\lambda^H - f_H)\|_{\rho} \leq R\lambda^{\zeta - a}. \tag{39}
\]
2) We have
\[
\|\mathcal{T}^{a - 1/2}\omega_\lambda^H\|_{H} \leq R \cdot \begin{cases} 
\lambda^{\zeta + a - 1}, & \text{if } -\zeta \leq a \leq 1 - \zeta, \\
\kappa^2(\zeta + a - 1), & \text{if } a \geq 1 - \zeta.
\end{cases} \tag{40}
\]
The above lemma could be proved using the spectral theorem, see (Lin and Cevher, 2018b) for details. The left hand-side of (39) is often called “true bias”.

Using the above lemma and some basic operator inequalities, we can prove the following analytic, deterministic result.

**Proposition 3** Under Assumption 2, let

\[
1 \vee \| T_{\lambda}^{1/2} T_{\lambda}^{-1/2} \|^{2} \vee \| T_{\lambda}^{-1/2} T_{\lambda}^{1/2} \|^{2} \leq \Delta_{1},
\]

\[
\| T_{\lambda}^{-1/2} [(T_{\lambda}^{\lambda} - S_{x}^{*} y) - (T_{\lambda}^{\lambda} - S_{\rho}^{*} f_{H})] \|_{H} \leq \Delta_{2},
\]

\[
\| T - T_{x} \| \leq \Delta_{3},
\]

\[
\| (I - P) T_{\lambda}^{1/2} \|^{2} = \Delta_{5}.
\]

Then, for any \(0 < a \leq [\zeta \wedge \frac{1}{2}]\), the following holds.

1) If \(\zeta \in [0, 1]\), then we have

\[
\| L^{-a}(S_{\rho}^{*} \omega_{\lambda}^{\rho} - f_{H})\|_{\rho} \leq \lambda^{-a} \Delta_{1}^{-a} \left( E \Delta_{2} + (2E + F + 1) R \lambda^{\zeta} + R \lambda^{\zeta-1} (E \Delta_{5} + \Delta_{5}^{1-a} \lambda^{a}) \right).
\]

(41)

2) If \(\zeta \geq 1\), then we have

\[
\| L^{-a}(S_{\rho}^{*} \omega_{\lambda}^{\rho} - f_{H})\|_{\rho} \leq \lambda^{-a} \Delta_{1}^{-a} \left( E \Delta_{2} + (E + F + 1) R \lambda^{\zeta} + \kappa^{2(\zeta-1)} R (E \Delta_{3} + E \Delta_{5} + \Delta_{5}^{1-a} \lambda^{a}) \right.
\]

\[
+ C_{\zeta}^{-\frac{1}{2}} \kappa \cdot FR(\lambda (E \Delta_{3} + \Delta_{5})^{1-a} + \lambda^{\frac{1}{2}} \Delta_{3}^{(\zeta-\frac{1}{2})} \lambda^{1-a}) \).
\]

(42)

The above proposition is key to our proof. The upper bounds from the proposition involve four quantities \(\Delta_{1,2,3,5}\). They will be estimated in the subsequent subsections. The proof of the above proposition \(\zeta \in [\frac{1}{2}, 1]\) borrows ideas from (Smale and Zhou, 2007; Caponnetto and De Vito, 2007; Rudi et al., 2015; Lin et al., 2018), whereas the key step is an error decomposition from (Lin and Cevher, 2018b). Our novelty lies in the proof for the cases \(\zeta \geq 1\) and \(\zeta \leq 1/2\), as well as some refined analysis and considering convergences under variants of norms.

**Proof** Adding and subtracting with the same term, and using the triangle inequality, we have

\[
\| L^{-a}(S_{\rho}^{*} \omega_{\lambda}^{\rho} - f_{H})\|_{\rho} \leq \| L^{-a} S_{\rho}(\omega_{\lambda}^{\rho} - \omega_{H}^{\lambda})\|_{\rho} + \| L^{-a}(S_{\rho}^{*} \omega_{\lambda}^{\rho} - f_{H})\|_{\rho}.
\]

Applying Part 1) of Lemma 13 to bound the last term, with \(0 \leq a \leq \zeta\),

\[
\| L^{-a}(S_{\rho}^{*} \omega_{\lambda}^{\rho} - f_{H})\|_{\rho} \leq \| L^{-a} S_{\rho}(\omega_{\lambda}^{\rho} - \omega_{H}^{\lambda})\|_{\rho} + R \lambda^{\zeta-a}
\]

\[
\leq \| L^{-a} S_{\rho} T^{a-\frac{1}{2}} \| \| T^{\frac{1}{2}-a}(\omega_{\lambda}^{\rho} - \omega_{H}^{\lambda})\|_{H} + R \lambda^{\zeta-a}.
\]

Using the spectral theorem for compact operators, \(L = S_{\rho} S_{\rho}^{*}\), and \(T = S_{\rho}^{*} S_{\rho}\), we have

\[
\| L^{-a} S_{\rho} T^{a-\frac{1}{2}} \| \leq 1,
\]
and thus
\[ \| L^{-a}(S_\rho \omega^2 - f_H) \|_\rho \leq \| T^{\frac{1}{2} - a}(\omega^2 - \omega_H^2) \|_H + R\lambda^{\zeta - a}. \] (43)

Adding and subtracting with the same term, and using the triangle inequality,
\[ \| L^{-a}(S_\rho \omega^2 - f_H) \|_\rho \leq \| T^{\frac{1}{2} - a}(\omega^2 - P\omega_H^2) \|_H + \| T^{\frac{1}{2} - a}(I - P)\omega_H^2 \|_H + R\lambda^{\zeta - a}. \]

Since \( P \) is an orthogonal projected operator and \( a \in [0, \frac{1}{2}] \), we have
\[
\| T^{\frac{1}{2} - a}(I - P)\omega_H^2 \|_H \\
= \| T^{\frac{1}{2}(1 - 2a)}(I - P)^{1 - 2a}(I - P)\omega_H^2 \|_H \\
\leq \| T^{\frac{1}{2}(1 - 2a)}(I - P)^{1 - 2a} \| \| (I - P)T^{\frac{1}{2}} \| \| T^{\frac{1}{2}}\omega_H^2 \|_H \\
\leq \| T^{\frac{1}{2}}(I - P)^{1 - 2a} \| \| (I - P)T^{\frac{1}{2}} \| R\lambda^{2(\zeta - 1) + \lambda^{\zeta - 1} - \Delta^{1 - a}} \\
= \Delta^{1 - a} R\lambda^{2(\zeta - 1) + \lambda^{\zeta - 1} - \Delta^{1 - a}},
\]

where for the last second inequality, we use Lemma 8 and Part 2) of Lemma 13, and we subsequently obtain
\[ \| L^{-a}(S_\rho \omega^2 - f_H) \|_\rho \leq \| T^{\frac{1}{2} - a}(\omega^2 - P\omega_H^2) \|_H + R\lambda^{2(\zeta - 1) + \lambda^{\zeta - 1} - \Delta^{1 - a}} + R\lambda^{\zeta - a}. \]

Since for all \( \omega \in H \), and \( a \in [0, \frac{1}{2}] \),
\[ \| T^{\frac{1}{2} - a} \|_H \leq \| T^{\frac{1}{2} - a} T^{\frac{1}{2} - a} \|_H \leq \| T^{\frac{1}{2}} \| \| T^{\frac{1}{2}} \| \| T^{\frac{1}{2}} \| \| \omega \|_H \\
\leq \lambda^{-a} \| T^{\frac{1}{2} - a} \|_H \| T^{\frac{1}{2}} \| \| T^{\frac{1}{2}} \| \| \omega \|_H \\
\leq \lambda^{-a} \| T^{\frac{1}{2} - a} \|_H \| T^{\frac{1}{2}} \| \| T^{\frac{1}{2}} \| \| \omega \|_H \\
\leq \lambda^{-a} \Delta^{1 - a} \| T^{\frac{1}{2}} \| \| \omega \|_H 
\] (44)

(where we use Lemma 8 for the last second inequality), we get
\[ \| L^{-a}(S_\rho \omega^2 - f_H) \|_\rho \leq \lambda^{-a} \Delta^{1 - a} \| T^{\frac{1}{2}}(\omega^2 - P\omega_H^2) \|_H + R\lambda^{2(\zeta - 1) + \lambda^{\zeta - 1} - \Delta^{1 - a}} + R\lambda^{\zeta - a}. \] (45)

In what follows, we estimate \( \| T^{\frac{1}{2}}(\omega^2 - P\omega_H^2) \|_H \).

Introducing with (15), with \( P^2 = P \),
\[ \| T^{\frac{1}{2}}(\omega^2 - P\omega_H^2) \|_H = \| T^{\frac{1}{2}} P(\mathcal{G}(PTxP)PS_y^*y - P\omega_H^2) \|_H. \]

Since for any \( \omega \in H \),
\[ \| T^{\frac{1}{2}} P \|_H^2 = \langle PTxP \omega, \omega \rangle_H \leq \langle (PTxP + \lambda I) \omega, \omega \rangle_H = \| (PTxP + \lambda I)^{\frac{1}{2}} \|_H^2, \]
and we thus get
\[ \| T^{\frac{1}{2}}(\omega^2 - P\omega_H^2) \|_H \leq \| U^{\frac{1}{2}}(\mathcal{G}(U)PS_y^*y - P\omega_H^2) \|_H, \]

18
where we denote
\[ U = PT \lambda P, \quad U_\lambda = U + \lambda I. \] (46)

Subtracting and adding with the same term, and applying the triangle inequality, with the notation \( R_\lambda \) given by (37) and \( P^2 = P \), we have
\[ \| T_{\lambda\lambda}^{1/2} (\omega^*_H - P\omega^*_H) \|_H \leq \| U_{\lambda 1}^{1/2} G_\lambda(U) P(S^*_X \tilde{y} - T_\lambda P\omega^*_H) \|_H + \| U_{\lambda 1}^{1/2} R_\lambda(U) P\omega^*_H \|_H. \] (47)

We will estimate the above two terms of the right-hand side.

**Estimating \( \| \text{Term.A} \|_H \):**
Using Lemma 11,
\[ \| U_{\lambda 1}^{1/2} G_\lambda(U) P T_{\lambda\lambda}^{1/2} \| \leq \| U_{\lambda 1}^{1/2} G_\lambda(U) U_{\lambda 1}^{1/2} \| = \| U_{\lambda 1} G_\lambda(U) \|. \]
Using the spectral theorem, with \( \| U \| \leq \| T_\lambda \| \leq \kappa^2 \) (implied by (6)), and then applying (16),
\[ \| U_{\lambda 1}^{1/2} G_\lambda(U) P T_{\lambda\lambda}^{1/2} \| \leq \sup_{w \in [0, \kappa^2]} |(u + \lambda)G_\lambda(u)| \leq E. \] (48)

Using the above inequality, and by a simple calculation,
\[ \| \text{Term.A} \|_H \leq \| U_{\lambda 1}^{1/2} G_\lambda(U) P T_{\lambda\lambda}^{1/2} (S^*_X \tilde{y} - T_\lambda P\omega^*_H) \| \leq E \| T_{\lambda\lambda}^{-1/2} (S^*_X \tilde{y} - T_\lambda P\omega^*_H) \|. \]

Adding and subtracting with the same terms, and using the triangle inequality,
\[ \| \text{Term.A} \|_H \leq E \| T_{\lambda\lambda}^{1/2} (S^*_X \tilde{y} - T_\lambda P\omega^*_H) \|_H \]
\[ \leq \| T_{\lambda\lambda}^{1/2} T_{\lambda\lambda}^{-1/2} (S^*_X \tilde{y} - T_\lambda P\omega^*_H) \|_H + E \| T_{\lambda\lambda}^{-1/2} T_\lambda P(I - P)\omega^*_H \|_H \]
\[ \leq E \Delta_{1/2} \| T_{\lambda\lambda}^{-1/2} (S^*_X \tilde{y} - T_\lambda P\omega^*_H) \|_H + E \| T_{\lambda\lambda}^{-1/2} T_\lambda P(I - P)\omega^*_H \|_H \]
\[ \leq E \Delta_{1/2} \| \Delta_2 + \| T_{\lambda\lambda}^{-1/2} (T_\lambda \omega^*_H - S^*_p f_H) \|_H \| + E \| T_{\lambda\lambda}^{-1/2} T_\lambda P(I - P)\omega^*_H \|_H \]
\[ \leq E \Delta_{1/2} \| \Delta_2 + \| T_{\lambda\lambda}^{-1/2} S^*_p \| \| S^*_p \omega^*_H - f_H \|_\rho \| + E \| T_{\lambda\lambda}^{1/2} T_\lambda P(I - P) \| \| (I - P) T_{\lambda\lambda}^{1/2} \| \| T_{\lambda\lambda}^{-1/2} \omega^*_H \|_H, \]

where we used \( T = S^*_X S^*_p \) and \( (I - P)^2 = I - P \) for the last inequality. Applying Lemma 13 and \( \| T_{\lambda\lambda}^{-1/2} S^*_p \| \leq 1 \),
\[ \| \text{Term.A} \|_H \leq E \Delta_{1/2} \| \Delta_2 + R\lambda^\xi \| + ER \Delta_{1/2} \| T_{\lambda\lambda}^{1/2} (I - P) \| \| \kappa^{2(\xi - 1)} + \lambda^{(\xi - 1)} \|. \] (49)

In what follows, we estimate \( \| T_{\lambda\lambda}^{1/2} (I - P) \| \), considering two different cases.

**Case \( \xi \leq 1 \).**
We have
\[ \| T_{\lambda\lambda}^{1/2} (I - P) \| \leq \Delta_{1/2} \| T_{\lambda\lambda}^{1/2} (I - P) \|. \]
Note that for any \( \omega \in H \) with \( \|\omega\|_H = 1 \),
\[
\|T^\frac{1}{2}(I - P)\omega\|_H^2 = \langle T^\frac{1}{2}(I - P)\omega, (I - P)\omega \rangle_H = \|T^\frac{1}{2}(I - P)\omega\|_H^2 + \lambda \|(I - P)\omega\|_H^2 \\
\leq \|T^\frac{1}{2}(I - P)\|_H^2 + \lambda \leq \Delta_5 + \lambda.
\]
It thus follows that
\[
\|T^\frac{1}{2}(I - P)\| \leq (\Delta_5 + \lambda)^\frac{1}{2},
\]
and thus
\[
\|T^\frac{1}{2}(I - P)\| \leq \Delta_1^2 (\Delta_5 + \lambda)^\frac{1}{2}.
\]
Introducing the above into (49), we know that Term.A can be estimated as \((\zeta \leq 1)\)
\[
\|\text{Term.A}\|_H \leq E \Delta_1^2 \left( \Delta_2 + 2R\lambda^\zeta + R\lambda^{\zeta-1}\Delta_5 \right). \tag{51}
\]

Case \( \zeta \geq 1 \).
Applying Lemma 12, we obtain
\[
\|T^\frac{1}{2}(I - P)\|^2 = \|T^\frac{1}{2}(I - P)T^\frac{1}{2}\| \leq \Delta_3 + \|T^\frac{1}{2}(I - P)\| = \Delta_3 + \Delta_5.
\]
Introducing the above into (49), we get for \( \zeta \geq 1 \),
\[
\|\text{Term.A}\|_H \leq E \Delta_1^2 \left( \Delta_2 + R\lambda^\zeta + (\Delta_3 + \Delta_5) \kappa^{2(\zeta-1)}R \right). \tag{52}
\]

Estimating \( \|\text{Term.B}\|_H \):
We estimate \( \|\text{Term.B}\|_H \), considering two different cases.

Case I: \( \zeta \leq 1 \).
Using a same argument as that for (48) and (17),
\[
\|U^\frac{1}{2}_X R_\lambda(U)PT^\frac{1}{2}_X\| \leq \sup_{u \in [0, \kappa^2]} |R_\lambda(u)(u + \lambda)| \leq F\lambda.
\]
Using the above inequality and by a direct calculation,
\[
\|\text{Term.B}\|_H \leq \|U^\frac{1}{2}_X R_\lambda(U)PT^\frac{1}{2}_X\| \|T^\frac{1}{2}_X T^\frac{1}{2}_X\| \|T^{-\frac{1}{2}}\omega_0^\lambda\|_H \leq F\lambda \Delta_1^2 \|T^{-\frac{1}{2}}\omega_0^\lambda\|_H.
\]
Applying Part 2) of Lemma 13, we get
\[
\|\text{Term.B}\|_H \leq FR\lambda^\zeta \Delta_1^2. \tag{53}
\]
Applying the above and (51) into (47), we know that for any \( \zeta \in [0, 1] \),
\[
\|T^\frac{1}{2}_X(\omega_0^\lambda - P\omega_0^\lambda)\|_H \leq \Delta_1^2 \left( E\Delta_2 + (2E + F)R\lambda^\zeta + ER\Delta_5\lambda^{\zeta-1} \right).
\]
Using the above into (45), we can prove the first desired result.

Case II: \( \zeta \geq 1 \)
We denote
\[
\mathcal{V} = T^\frac{1}{2}_X P T^\frac{1}{2}_X, \quad \mathcal{V}_\lambda = \mathcal{V} + \lambda I. \tag{54}
\]
Noting that $U = PT_x P = PT_x^2 (PT_x^2)^*$, thus following from Lemma 9 (with $f(u) = (u + \lambda)^\frac{1}{2} R_\lambda(u)$) and $P^2 = P$,

$$\|U^\frac{1}{2} R_\lambda(U) PT_x^2 \| = \|U^\frac{1}{2} R_\lambda(U)(PT_x^2) T_x^{-1}\| = \|(PT_x^2) V^\frac{1}{2} \lambda R_\lambda(V) T_x^{-1}\|.$$

Adding and subtracting with the same term, using the triangle inequality, we obtain

$$\|U^\frac{1}{2} R_\lambda(U) PT_x^2 \| \leq \|PT_x^2 V^\frac{1}{2} \lambda R_\lambda(V) V^\frac{1}{2} \lambda R_\lambda(V) - V^\frac{1}{2} \lambda R_\lambda(V) T_x^{-1}\| + \|PT_x^2 V^\frac{1}{2} \lambda R_\lambda(V) T_x^{-1}\| + \|PT_x^2 V^\frac{1}{2} \lambda R_\lambda(V) T_x^{-1}\|.$$

Now we are ready to estimate $\|T_x - V\|$. Noting that $U^\frac{1}{2} = \lambda T_x - V$, we have

$$|R_\lambda(u) u^{\frac{1}{2}} (u + \lambda)^\frac{1}{2}| \leq F\lambda^s,$$

and thus we get

$$\|U^\frac{1}{2} R_\lambda(U) PT_x^2 \| \leq F(\lambda^\frac{1}{2} + \lambda C_{\zeta - 1, \kappa} |T_x - V|)^{(\zeta - 1)^\frac{1}{2}}.$$

Using Lemma 12, $(I - P)^2 = I - P$ and $\|A^* A\| = \|A\|^2$, we have

$$\|T_x - V\| = \|T_x^\frac{1}{2} (I - P) T_x^\frac{1}{2}\| \leq \|T_x - T\| + \|T_x^\frac{1}{2} (I - P) T_x^\frac{1}{2}\| \leq \Delta_3 + \Delta_5,$$

and we thus get

$$\|U^\frac{1}{2} R_\lambda(U) PT_x^2 \| \leq F(\lambda^\frac{1}{2} + \lambda C_{\zeta - 1, \kappa} (\Delta_3 + \Delta_5)^{(\zeta - 1)^\frac{1}{2}}).$$

Now we are ready to estimate $\|\textbf{Term.B}\|_H$. By some direct calculations and Part 2) of Lemma 13,

$$\|\textbf{Term.B}\|_H \leq \|U^\frac{1}{2} R_\lambda(U) PT_x^2 \| \|T_x^\frac{1}{2} - \omega^\frac{1}{2}_H\| \leq \|U^\frac{1}{2} R_\lambda(U) PT_x^2 \| R.$$

Adding and subtracting with the same term, using the triangle inequality,

$$\|\textbf{Term.B}\|_H \leq R \left( \|U^\frac{1}{2} R_\lambda(U) PT_x^2 \| + \|U^\frac{1}{2} R_\lambda(U) \| \|T_x^\frac{1}{2} - \omega^\frac{1}{2}_H\| \right).$$

Using the spectral theorem, with $\|U\| \leq \|T_x\| \leq \kappa^2$ by (6) and (17),

$$\|U^\frac{1}{2} R_\lambda(U) \| = \sup_{u \in (0, \kappa^2]} |R_\lambda(u) u^{\frac{1}{2}} (u + \lambda)^\frac{1}{2}| \leq F\lambda^\frac{1}{2};$$
and we thus get
\[ \| \text{Term.B} \|_H \leq R \left( \| U_\lambda^2 R(\lambda) P T_\lambda^\frac{\lambda}{2} \| + F \lambda^\frac{1}{2} \| T_\lambda^\frac{\lambda}{2} \| \right). \]

Applying Lemma 10, with (3) and (6), it follows that
\[ \| \text{Term.B} \|_H \leq R \left( \| U_\lambda^2 R(\lambda) P T_\lambda^\frac{\lambda}{2} \| + F \lambda^\frac{1}{2} \| T_\lambda^\frac{\lambda}{2} \| \right). \]

Introducing with (55), we obtain
\[ \| \text{Term.B} \|_H \leq FR \left( \lambda \xi + C \xi \lambda (\Delta_3 + \Delta_5) (\xi^{-1})^\lambda + \lambda^\frac{1}{2} \Delta_3 (\xi^{-\frac{1}{2}})^\lambda \right). \]

Introducing the above inequality and (52) into (47), noting that \( \Delta_1 \geq 1 \) and \( \kappa_2 \geq 1 \), we know that for any \( \zeta \geq 1 \), the following holds
\[ \| T_\lambda^\frac{\lambda}{2} (\omega - P \omega) \|_H \leq \Delta_1 \left( E \Delta_2 + (F + E) R \lambda \xi + E \kappa^2 (\xi^{-1}) R (\Delta_3 + \Delta_5) \right. \]
\[ \left. + C \xi \lambda (\Delta_3 + \Delta_5) (\xi^{-1})^\lambda + \lambda^\frac{1}{2} \Delta_3 (\xi^{-\frac{1}{2}})^\lambda \right). \]

Using the above into (45), and by a simple calculation, we can prove the second desired result.

### 5.3. Probabilistic Estimates

To derive total error bounds from Proposition 3, it is necessary to develop probabilistic estimates for the random quantities \( \Delta_1 \), \( \Delta_2 \), and \( \Delta_3 \). We thus introduce the following three lemmas.

**Lemma 14** Under Assumption 3, let \( \delta \in (0, 1) \), and \( \lambda = n^{-\theta} \) with \( \theta \in [0, 1) \) or \( \lambda = [1 \lor \log n^\gamma]/n \). Then with probability at least \( 1 - \delta \),
\[ \| (T + \lambda I)^{1/2} (T_\lambda + \lambda I)^{-1/2} \|_F \leq \sqrt{2} \kappa \log \left( \frac{4 \kappa^2 (c_\gamma + 1)}{\delta \| T \|} \right) \]
where \( a(\delta) = 8 \kappa^2 \log \left( \frac{4 \kappa^2 (c_\gamma + 1)}{\delta \| T \|} \right) \) if \( \lambda = [1 \lor \log n^\gamma]/n \), or \( a(\delta) = 8 \kappa^2 \left( \log \left( \frac{4 \kappa^2 (c_\gamma + 1)}{\delta \| T \|} \right) + \frac{\theta \gamma}{e(1 - \theta)} \right) \)
otherwise.

The proof of the above result for \( \lambda = n^{-\theta} \) with \( \theta \in [0, 1) \) is given in (Lin and Cevher, 2018b). Here, we also provide a similar result for \( \lambda = [1 \lor \log n^\gamma]/n \) using the same argument. We report the proof in Appendix.

**Lemma 15** Let \( 0 < \delta < 1/2 \). The following holds with probability at least \( 1 - \delta \):
\[ \| T - T_\lambda \| \leq \| T - T_\lambda \|_{HS} \leq \frac{2 \kappa^2 \log (2/\delta)}{n} + \sqrt{\frac{2 \kappa^4 \log (2/\delta)}{n}}. \]
Here, \( \| \cdot \|_{HS} \) denotes the Hilbert-Schmidt norm.
Using (Smale and Zhou, 2007, Lemma 2) (which is a direct corollary of the concentration inequality for Hilbert-space valued random variables from (Pinelis and Sakhanenko, 1986)), one can prove the desired result.

**Lemma 16** Under Assumptions 1 and 2, with probability at least $1 - \delta$, the following holds:

$$\left\| T_{\lambda}^{\frac{1}{2}}(T_{\lambda}^{\frac{1}{2}} - S_{\lambda}^{\frac{1}{2}} Y_T - T_{\lambda} Y_H + S_{\lambda} f_H) \right\|_H \leq 2 \left( \frac{4\kappa(M + \kappa^{1/(2\zeta)} R\lambda^{(\zeta - \frac{1}{2})^-})}{n\sqrt{\lambda}} + \sqrt{\frac{8(2R^2\kappa^2\lambda^{2\zeta - 1} + (2B^2 + Q^2)N(\lambda))}{n}} \right) \log \frac{2}{\delta}. \quad (56)$$

The above lemma is essentially proved in (Lin and Cevher, 2018b; Lin et al., 2018). We include a proof in Appendix for completeness.

### 5.4. Proof for Projected-regularized Algorithms

With the above probabilistic estimates and the analytic result, Proposition 3, we are now ready to prove the following proposition and the results for the projected-regularized algorithms stated in Theorem 1.

**Proposition 4** Under Assumptions 1 and 2, let $\left\| (I - P) T_{\lambda}^{\frac{1}{2}} \right\|^2 = \Delta_5$, and $\lambda = n^{-\theta}$ for some $\theta \in [0, 1)$ or $\lambda = \frac{1}{\sqrt{n}}$. Then, for any $0 \leq a \leq \left\lfloor \zeta / 2 \right\rfloor$, with probability at least $1 - 3\delta$ ($\delta \in (0, 1/3)$), the following statements hold.

1) If $\zeta \in [0, 1]$, we have

$$\left\| L^{-a}(S_{\lambda}^{\frac{1}{2}} - f_H) \right\|_\rho \leq \bar{C}_1^{1-a} \log^{1-a} \frac{2}{\delta} \lambda^{\zeta-1-a}(E\Delta_5 + \Delta_5^{1-a}\lambda^a)R$$

$$+ \bar{C}_1^{1-a} \log^{2-a} \frac{2}{\delta} \lambda^{-a} \left( \bar{C}_2(\lambda^\zeta \sqrt{\frac{1}{n\sqrt{\lambda}}}R + 8E\sqrt{\frac{N(\lambda)}{n}}(B + Q) + 8\kappa E \frac{M}{n\sqrt{\lambda}}) \right).$$

2) If $\zeta \geq 1$ and $\lambda \geq n^{-1/2}$, we have

$$\left\| L^{-a}(S_{\lambda}^{\frac{1}{2}} - f_H) \right\|_\rho \leq \bar{C}_1^{1-a} C_{\zeta,1} \log^{1-a} \frac{2}{\delta} \lambda^{-a}(E\Delta_5 + \Delta_5^{1-a}\lambda^a + F\lambda \Delta_5^{\left(\zeta - 1\right)\lambda})R$$

$$+ \bar{C}_1^{1-a} \log^{2-a} \frac{2}{\delta} \lambda^{-a} \left( \bar{C}_3(\lambda^\zeta \sqrt{\frac{1}{n\sqrt{\lambda}}}R + 8E\sqrt{\frac{N(\lambda)}{n}}(B + Q) + 8\kappa E \frac{M}{n\sqrt{\lambda}}) \right).$$

Here, the constants $\bar{C}_{1,2,3}$ are defined by

$$\bar{C}_1 = \begin{cases} 
24\kappa^2 \left( \log \frac{2\kappa^{2\zeta}(\zeta + 1)}{\| \|_\|} + 1 \right), & \text{if } \lambda = \frac{1}{\sqrt{n}}, \\
24\kappa^2 \left( \log \frac{2\kappa^{2\zeta}(\zeta + 1)}{\| \|} + \frac{\theta_1}{(1-\theta)} \right), & \text{otherwise},
\end{cases}$$

$$\bar{C}_2 = 8E\kappa(\frac{1}{\zeta + 2}) + 2E + F + 1.$$  

$$\bar{C}_3 = 8E\kappa(\frac{1}{\zeta + 2} + 1) + E + F + 1 + C_{\zeta,1}(E + F)\kappa^2(2 + \sqrt{2}).$$

Furthermore, if $\zeta \geq 1/2$, then the above conclusions still hold if we replace $\left\| L^{-a}(f_H^{\frac{1}{2}} - f_H) \right\|_\rho$ by $\left\| T_{\lambda}^{\frac{1}{2}}(\omega_H^{\frac{1}{2}} - \omega_H) \right\|_H$. 

23
Proof We use Proposition 3 to prove the statement. We thus need to estimate $\Delta_1$, $\Delta_2$ and $\Delta_3$. Following from Lemmas 14, 15 and 16, with $n^{-1} \leq \lambda \leq 1$, we know that with probability at least $1 - 3\delta$,

$$
\Delta_1 \leq \bar{C}_1 \log \frac{2}{\delta},
$$

$$
\Delta_2 \leq \left( C_2 (\lambda^{\xi} \vee \frac{1}{n^{\sqrt{\lambda}}}) R + 8 \sqrt{\frac{N(\lambda)}{n}} (B + Q) + 8 \kappa \frac{M}{n^{\sqrt{\lambda}}} \right) \log \frac{2}{\delta}, \quad C_2 = 8 \kappa (\kappa^{1/2} + 1),
$$

$$
\Delta_3 \leq C_3 \frac{1}{\sqrt{n}} \log \frac{2}{\delta}, \quad C_3 = \kappa^2 (\sqrt{2} + 2).
$$

The convergence results in $L^2_{\rho_X}$-norm thus follow by introducing the above estimates into (41) or (42), combining with a direct calculation and the assumption of $1/n \leq \lambda \leq 1$.

The proof for the convergence results in $H$-norm in the attainable case parallelizes to that for results in $L^2_{\rho_X}$-norm, as we can replace (43) by

$$
\|T^{1/2-a}(\omega^X - \omega_H)\|_H \leq \|T^{1/2-a}(\omega^X - \omega^X_H)\|_H + R \lambda^{\xi-a}.
$$

Proof of Theorem 1 Theorem 1 is a direct consequence of Proposition 4 with Assumption 3 and using a simple calculation. Corollary 1 is a direct consequence of Theorem 1.

5.5. Proof for Sketched-regularized Algorithms

In order to use Corollary 1 for sketched-regularized algorithms, we need to estimate the projection error. The basic idea is to approximate the projection error in terms of its 'empirical' version, $\|(I - P)T^{1/2}_X\|^2$. The estimate for $\|(I - P)T^{1/2}_X\|^2$ is quite lengthy and it is divided into several steps.

Lemma 17 Let $0 < \delta < 1$ and $\theta \in [0, 1]$. Given a fixed input set $x \subseteq H^n$, assume that for $\lambda \in [0, 1]$,

$$
\text{tr}((T_x + \lambda I)^{-1} T_x) \leq b_\gamma \lambda^{-\gamma}
$$

holds for some $b_\gamma > 0$, $\gamma \in [0, 1]$. Then there exists a subset $U_x$ of $\mathbb{R}^{m \times n}$ with measure at least $1 - \delta$, such that for all $G \in U_x$, the following holds:

$$
\|(I - P)T^{1/2}_X\|^2 \leq 6 \lambda,
$$

provided that

$$
m \geq 100c_0^3 \log^3 n \lambda^{-\gamma} \log \frac{3}{\delta} (1 + 10b_\gamma).
$$

Under the condition (57), Lemma 17 provides an upper bound for $\|(I - P)T^{1/2}_X\|$. The left-hand side of (57) is called empirical effective dimension. It can be estimated as follows.
Lemma 18  Under Assumption 3, let $0 < \delta < 1$. For any fixed $\lambda = n^{-\theta}$ with $\theta \in [0, 1)$, or $\lambda = \frac{1 + \log n}{n}$, with probability at least $1 - \delta$, the following holds:

$$\text{tr}((T_x + \lambda I)^{-1}T_x) \leq b_\gamma \log^2 \frac{4}{\delta} \lambda^{-\gamma}. \quad (59)$$

Here, $b_\gamma$ is a positive constant given by

$$b_\gamma = 24\kappa^2(4\kappa^2 + 2\kappa \sqrt{c_\gamma + c_\gamma}) \left( \log \frac{2\kappa^2(c_\gamma + 1)}{\|T\|} + 1 + \tilde{c} \right), \quad \tilde{c} = \begin{cases} \frac{1}{\theta \gamma}, & \text{if } \lambda = \frac{1 + \log n}{n}, \\ \frac{\theta \gamma}{\theta (1 - \theta)}, & \text{otherwise}. \end{cases}$$

The above lemma improves (Rudi et al., 2015, Proposition 1). It does not require the extra assumption that the sample size is large enough, and our proof is simpler.

Now we are ready to estimate the projection error with randomized sketches as follows.

Lemma 19  Under Assumption 3, let

$$S = \text{range}\{S^*_xG^*\},$$

where $G \in \mathbb{R}^{m \times n}$ is a random matrix satisfying (26), and $P$ be the projection operator with its range $S$. Then with probability at least $1 - 3\delta$ ($\delta \in (0, 1/3)$), we have

$$\| (I - P)T \|^2 \leq \frac{1}{n^\theta} \left( 1 + \frac{\log n^\gamma}{n^{1 - \theta}} \right) 7a_\gamma \log^4 \frac{4}{\delta},$$

provided that

$$m \geq \tilde{C} n^{\theta \gamma} \log^\beta n (1 + \log n^\gamma)^c \log^3 \frac{4}{\delta}, \quad c = \begin{cases} 0, & \text{if } \theta < 1, \\ -\gamma, & \text{if } \theta = 1. \end{cases} \quad (60)$$

Here, $a_\gamma = 24\kappa^2 \log \frac{\kappa^2 (c_\gamma + 1)}{\|T\|}$, and $\tilde{C} = 100c_0 (1 + 10b_\gamma)$ with

$$b_\gamma = 24\kappa^2(4\kappa^2 + 2\kappa \sqrt{c_\gamma + c_\gamma}) \left( \log \frac{2\kappa^2(c_\gamma + 1)}{\|T\|} + 1 + \tilde{c} \right), \quad \tilde{c} = \begin{cases} \frac{\theta \gamma}{\theta (1 - \theta)}, & \text{if } \theta < 1, \\ 1, & \text{if } \theta = 1. \end{cases}$$

The proofs for Lemmas 17-19 are given in the appendix.

With Lemma 19, we can use Corollary 1 to prove Corollary 3 for the sketched-regularized algorithms as follows.

Proof of Corollary 3  Applying Lemma 19 with

$$\theta = \begin{cases} 1, & \text{if } 2\zeta + \gamma \leq 1, \\ \zeta - a \frac{(1 - a)(2\zeta + \gamma)}{2\zeta + \gamma}, & \text{if } \zeta \geq 1, \\ \frac{1}{2\zeta + \gamma}, & \text{otherwise} \end{cases}$$

we get that under the condition (27), with probability at least $1 - 3\delta$, it holds that

$$\Delta_5 \leq \frac{1}{n^\theta} \left( 1 + \frac{\log n^\gamma}{n^{1 - \theta}} \right) \log^4 \frac{4}{\delta} \lesssim \log^4 \delta \begin{cases} \frac{\lambda}{\lambda} \frac{\zeta - a}{\zeta - a}, & \text{if } \zeta \leq 1, \\ \frac{\lambda}{\lambda} \frac{\zeta - a}{\zeta - a}, & \text{if } \zeta \geq 1, \end{cases}$$

25
where we use the following fact
\[
\frac{\log n^\gamma}{n^{1-\theta}} = \frac{\gamma}{1-\theta} \frac{\log n^{1-\theta}}{n^{1-\theta}} \leq \frac{\gamma}{1-\theta}, \quad \text{if } \theta < 1,
\]
within the last inequality. Combining with Corollary 1, and by a direct calculation, with \(\lambda \leq 1\), one can prove the desired result.

**Remark 7** Roughly speaking, and ignoring the logarithmic factors, in the proof of Lemma 19 for the case \(\gamma \in (0, 1]\), we have the following high-probability upper bound for the projection error:
\[
\|(I - P)^T \frac{1}{2}\|^2 \lesssim m^{-\frac{1}{2}}.
\]
Introducing this estimate into Theorem 1, we observe that the following conclusions hold with high probability for \(\lambda \in (n^{-1}, 1]\) and \(\alpha \in [0, \frac{1}{2} \wedge \zeta]\):

For \(\zeta \in [0, 1]\),
\[
\|\mathcal{L}^{-\alpha}(f^2_\lambda - f_H)\|_\rho \lesssim \lambda^{\zeta-\alpha} + \frac{1}{n^{\frac{1}{2}} \lambda^{\alpha + \frac{1}{2}}} + \frac{1}{m^{\frac{1}{2}} \lambda^\alpha} + \frac{1}{m^{\frac{1}{2}} \gamma} + \frac{\lambda^{1-\alpha}}{m^{\frac{(\zeta-1)\lambda}{\gamma}}}.
\]
while for \(\zeta \geq 1\) and \(\lambda \geq n^{-1/2}\), we have
\[
\|\mathcal{L}^{-\alpha}(f^2_\lambda - f_H)\|_\rho \lesssim \lambda^{\zeta-\alpha} + \frac{1}{n^{\frac{1}{2}} \lambda^{\alpha + \frac{1}{2}}} + \frac{1}{m^{\frac{1}{2}} \lambda^\alpha} + \frac{1}{m^{\frac{1}{2}} \gamma} + \frac{\lambda^{1-\alpha}}{m^{\frac{(\zeta-1)\lambda}{\gamma}}}.
\]

Clearly, the regularization parameter, the sample size, and the sketching dimension have a direct impact on the upper bound. To minimize the upper bound, it is necessary to trade off these parameters.

5.6. Proof for Nyström-regularized Algorithms

In this subsection, we first estimate the projection errors for Nyström-regularized algorithms and then leverage Corollary 1 to prove Corollaries 4 and 5.

The following lemma estimates projection errors with the plain Nyström subsampling scheme.

**Lemma 20** Under Assumption 3, let \(P\) be the projection operator with range
\[
S = \text{span}\{x_1, \ldots, x_m\}.
\]
Then with probability at least \(1 - \delta\) (\(\delta \in (0, 1]\)), the following inequality holds:
\[
\|(I - P)^T \frac{1}{2}\|^2 \leq \|(I - P)^T \frac{1}{2}\|^2 \leq \frac{1}{m} \log m^\gamma \frac{4\kappa^2 \gamma^2 (c_\gamma + 1)}{\delta \|T\|}, \quad (61)
\]
where \(\mu = \frac{1}{m} \log m^\gamma\).

The next lemma provides upper bounds for projection errors with ALS Nyström subsampling scheme.
Lemma 21  Under Assumption 3, let $S = \text{range}\{	ilde{x}_1, \cdots, \tilde{x}_m\}$, with each $\tilde{x}_j$ drawn following an $(L, \lambda)$-ALS Nystr"om subsampling scheme, and $P$ be the projection operator with its range $S$. Let $\lambda = n^{-\theta}$ if $\theta \in [0, 1)$, or $\lambda = \frac{1 + \log n^\gamma}{n}$ if $\theta = 1$. Then with probability at least $1 - 3\delta$ ($\delta \in (0, 1/3)$), we have
\[
\|(I - P)T^{\frac{1}{2}}\|^2 \leq \frac{1}{n^\theta} \left(1 + \frac{\log n^\gamma}{n^{1-\theta}}\right) 4a_\gamma \log \frac{4}{\delta},
\]
provided that
\[
m \geq \tilde{C}_1 n^{\theta \gamma} (1 \lor \log n^\gamma)^c \log \frac{4}{\delta}, \quad c = \begin{cases} 1, & \text{if } \theta < 1, \\ 1 - \gamma, & \text{if } \theta = 1. \end{cases}
\]
Here, $\tilde{C}_1 = 8b_\gamma L^2 (4 + \log (2b_\gamma))$ where $a_\gamma$ and $b_\gamma$ are given by Lemma 19.

The proofs for the two above lemmas will be given in the appendix.

Proof of Corollary 4  Combining Corollary 1 with Lemma 20, one can prove the desired result. ■

Proof of Corollary 5  Combining Corollary 1 with Lemma 21, one can prove the desired result. ■

6. Proof for Section 4
In this section, we prove the results in Section 4. We first prove the following result.

Theorem 5  Under Assumptions 2, 3 and 4, let
\[
T = [(\eta \lambda)^{-1}], \quad \lambda = n^{-\frac{1}{\sqrt{1 + \zeta + \gamma}}} (1 \lor \log n^\gamma)^{1 \{2 \zeta \gamma \leq 1\}},
\]
and let
\[
0 < \eta \leq \frac{1}{8\kappa^2 (\log T + 1)}.
\]
Then for any $a \in [0, \frac{1}{2} \lor \zeta]$, the following holds with probability at least $1 - \delta$ ($0 < \delta < 1$).
1) If $\zeta \leq 1$, we have
\[
\mathbb{E}_j \|L^{-a}(S_\rho \omega_{T+1} - f_H)\|_{\rho}^2 \lesssim \lambda^{2(\zeta - a)} (1 + \lambda^{-\frac{1}{\Delta_5}}) 2^{\log \frac{4}{\delta}} + \eta b^{-1} \lambda^{-2a} (\log T)^{1 \{2a \not\in 1\}} \log \frac{4}{\delta}.
\]
2) If $\zeta \geq 1$, we have
\[
\mathbb{E}_j \|L^{-a}(S_\rho \omega_{T+1} - f_H)\|_{\rho}^2 \lesssim \lambda^{-2a} \left(\lambda^\zeta + \Delta_5 \left(1 + \left(\frac{\lambda}{\Delta_5}\right)^{\Delta_5^a} + \left(\frac{\Delta_5}{\lambda}\right)^{-a}\right)\right)^2 \log \frac{4}{\delta} + \eta b^{-1} \lambda^{-2a} (\log T)^{1 \{2a \not\in 1\}} \log \frac{4}{\delta}.
\]
Furthermore, if $\zeta \geq 1/2$, then the above conclusions still hold if we replace $\|L^{-a}(S_\rho \omega_{T+1} - f_H)\|_{\rho}$ by $\|T^{\frac{1}{2} - a}(\omega_{T+1} - \omega_H)\|_H$. Here, $\Delta_5$ is the projection error $\|(I - P)T^{\frac{1}{2}}\|^2$. 

27
Proof We only provide the proof sketches and omit the universal constants in the proof. We first introduce an auxiliary sequence \( \{\nu_t\}_{t=1}^T \), generated by projected gradient descent and given by \( \nu_1 = 0 \),

\[
\nu_{t+1} = \tilde{g}_t(PT_xP)S^*_x\tilde{y}, \quad \tilde{g}_t(\cdot) = \sum_{k=1}^t \eta_k \prod_{i=k+1}^t (I - \eta_i).
\]

Following (Lin and Rosasco, 2017a, (5.17)), which is originally motivated by (Lin and Rosasco, 2017b), we can prove the following decomposition:

\[
E_J \|\mathcal{L}^{-a}(S_\rho \omega_{T+1} - f_H)\|^2_\rho = \|\mathcal{L}^{-a}(S_\rho \nu_{T+1} - f_H)\|^2_\rho + E_J \|\mathcal{L}^{-a}S_\rho(\omega_{T+1} - \nu_{T+1})\|^2_\rho.
\]

In what follows, we estimate the last two terms separately.

We first estimate \( \|\mathcal{L}^{-a}(S_\rho \nu_{T+1} - f_H)\|^2_\rho \). As noted in Remark 1, \( \tilde{g}_t(\cdot) \) is a filter function with regularization parameter \((\eta t)^{-1}\). As \( \lambda \simeq (\eta T)^{-1} \) by our assumptions, with a simple modification of the proof for Corollary 1, we know that the error estimates in Corollary 1 hold with \( f^*_x = S_\rho \omega_{T+1} \).

What remains is to prove the following error bounds:

\[
E_J \|\mathcal{L}^{-a}S_\rho(\omega_{T+1} - \nu_{T+1})\|^2_\rho \lesssim \eta b^{-1}\lambda^{-2a}(\log n)^1(2\varphi^{-1}) \log^2 \frac{2}{\delta}. \tag{67}
\]

We first consider the case \( a < 1/2 \). From the proof for (43) and using (44), we have

\[
\|\mathcal{L}^{-a}S_\rho(\omega_{T+1} - \nu_{T+1})\|_\rho \leq \|T_x^\frac{1}{2-a}(\omega_{T+1} - \nu_{T+1})\|_H \leq \lambda^{-a} \Delta_1^{1-a} \|\mathcal{T}_x^\frac{1}{2}(\omega_{T+1} - \nu_{T+1})\|_H.
\]

Following the proof for (Lin and Rosasco, 2017a, Proposition 5.21), under Condition (64), we have

\[
E_J \|T_x^\frac{1}{2}(\omega_{T+1} - \nu_{T+1})\|^2_H \leq 48\kappa^2 \mathcal{E}_2(0) \eta b^{-1} \log(3T).
\]

Thus, we have

\[
E_J \|\mathcal{L}^{-a}S_\rho(\omega_{T+1} - \nu_{T+1})\|^2_\rho \leq \lambda^{-2a} \Delta_1^{1-2a} 48\kappa^2 \mathcal{E}_2(0) \eta b^{-1} \log(3T).
\]

Applying (Lin and Rosasco, 2017b, Lemma 25) and Lemma 14 to estimate \( \mathcal{E}_2(0) \) and \( \Delta_1 \) respectively, we can prove that (67) holds with probability at least \( 1 - \delta \). The proof for the case \( a = 1/2 \) is simpler. In fact, by (5), we have

\[
\|\mathcal{L}^{-1/2}S_\rho(\omega_{T+1} - \nu_{T+1})\|_\rho \leq \|\omega_{T+1} - \nu_{T+1}\|_H.
\]

Following the similar arguments as that for (Lin and Rosasco, 2017b, (77)) and (Lin and Rosasco, 2017a, Proposition 5.21), under Condition (64), we can prove

\[
E_J \|\omega_{T+1} - \nu_{T+1}\|^2_H \lesssim \eta b^{-1}\lambda^{-1} \mathcal{E}_2(0).
\]

Combining with (Lin and Rosasco, 2017b, Lemma 25), we can prove that (67) holds with probability at least \( 1 - \delta \).
From the above analysis, we conclude the proof.

Now we are ready to prove Theorem 2 and its corollaries.

**Proof of Theorem 2** Simply applying Theorem 5 with specific choices on $\eta, b$ and $T$, one can prove the desired results.

**Proof of Corollary 6** Simply applying Theorem 2 and noting that Condition (31) is satisfied trivially since $P = I$.

**Proof of Corollary 7** The proof can be done by combing Theorem 2 with Lemmas 19-21, and following exactly the same steps as that for Corollaries 3-5.

7. Conclusion

In this paper, we first prove optimal statistical results with respect to variants of norms for sketched or Nyström regularized algorithms. Our contributions are mainly on theoretical aspects. First, our results for sketched-regularized algorithms generalize previous results (Yang et al., 2017) from the fixed design setting to the random design setting. Moreover, our results involve the regularity/smoothness of the target function and thus can have a faster convergence rate. Second, our results cover the non-attainable cases. Third, our results provide optimal, capacity-dependent rates even when $\zeta \geq 1$. This may suggest that sketched/Nyström regularized algorithms have certain advantages in comparison with distributed learning algorithms (Zhang et al., 2015), as the latter suffer a saturation effect over $\zeta = 1$. We then extend our analysis to stochastic gradient methods with projections, allowing multi-pass over the data and minibatches, and we derive similar optimal statistical results. A future direction is to extend our analysis to learning with random features, see (Rahimi and Recht, 2008; Sriperumbudur and Sterge, 2017) and the references therein.

Acknowledgements

This work was sponsored by the Department of the Navy, Office of Naval Research (ONR) under a grant number N62909-17-1-2111. It has also received funding from Hasler Foundation Program: Cyber Human Systems (project number 16066), the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (grant agreement number 725594-time-data), the NSF of China under grant number 11971427, the Fundamental Research Funds for the Central Universities under grant number 2019QN81010.
Appendix A. Proofs for Section 5

In this appendix, we prove the lemmas stated in Section 5.

A.1. Proof of Lemma 14

We first introduce the following basic probabilistic estimate.

**Lemma 22** Let $X_1, \ldots, X_m$ be a sequence of independently and identically distributed self-adjoint Hilbert-Schmidt operators on a separable Hilbert space. Assume that $E[X_1] = 0$, and $\|X_1\| \leq B$ almost surely for some $B > 0$. Let $V$ be a positive trace-class operator such that $E[X_1^2] \preceq V$. Then with probability at least $1 - \delta$, ($\delta \in ]0, 1[), there holds

$$\left\| \frac{1}{m} \sum_{i=1}^{m} X_i \right\| \leq \frac{2B\beta}{3m} + \sqrt{\frac{2\|V\|\beta}{m}}, \quad \beta = \log \frac{4\text{tr} V}{\|V\|\delta}.$$ 

The above lemma was first proved in (Hsu et al., 2014; Tropp, 2012) for the matrix case, and it was later extended to the general operator case in (Minsker, 2011), see also (Rudi et al., 2015; Bach, 2017; Dicker et al., 2017). We refer to (Rudi et al., 2015; Dicker et al., 2017) for the proof.

Using Lemma 22, we can prove the following result. Refer to (Lin and Cevher, 2018b) for proof details.

**Lemma 23** Let $0 < \delta < 1$ and $\lambda > 0$. With probability at least $1 - \delta$, the following holds:

$$\left\| (T + \lambda I)^{-1/2}(T - T_x)(T + \lambda I)^{-1/2} \right\| \leq \frac{4\kappa^2\beta}{3|x|\lambda} + \sqrt{\frac{2\kappa^2\beta}{|x|\lambda}}, \quad \beta = \log \frac{4\kappa^2(N(\lambda) + 1)}{\delta\|T\|}.$$ 

We are now ready to proof Lemma 14.

**Proof of Lemma 14** By a simple calculation, we have if $0 \leq u \leq 1/2$, then $2u^2/3 + u \leq 2/3$. Letting $\sqrt{\frac{2\kappa^2\beta}{|x|\lambda}} = u$, and combining with Lemma 23, we know that if

$$\sqrt{\frac{2\kappa^2\beta}{|x|\lambda}} \leq \frac{1}{2},$$

which is equivalent to

$$|x| \geq \frac{8\kappa^2\beta}{\lambda'}, \quad \beta = \log \frac{4\kappa^2(1 + N'(\lambda'))}{\delta\|T\|}, \quad (68)$$

then with probability at least $1 - \delta$,

$$\left\| T_{\lambda'}^{-1/2}(T - T_x)T_{\lambda'}^{-1/2} \right\| \leq 2/3. \quad (69)$$

Note that (69) implies

$$\|T_{\lambda'}^{1/2}T_{x\lambda'}^{-1/2}\|^2 \vee \|T_{\lambda'}^{1/2}T_{\lambda'}^{-1/2}\|^2 \leq 3. \quad (70)$$
Indeed,
\[ \|T_{\lambda'}^{1/2}T_{\lambda}^{-1/2}\|^2 = \|T_{\lambda'}^{1/2}T_{\lambda}^{-1}T_{\lambda'}^{1/2}\| = \|(I - T_{\lambda'}^{1/2}(T - T_{\lambda})T_{\lambda'}^{1/2})^{-1}\| \leq 3, \]
and
\[ \|T_{\lambda'}^{-1/2}T_{\lambda}^{1/2}\|^2 = \|T_{\lambda'}^{-1/2}T_{\lambda}T_{\lambda'}^{-1/2}\| = \|T_{\lambda'}^{-1/2}(T - T_{\lambda})T_{\lambda'}^{-1/2} + I\| \leq 3. \]
From the above analysis, we know that for any fixed \( \lambda' > 0 \) such that (68), then with probability at least \( 1 - \delta, \) (70) holds.

Let \( \lambda' = a\lambda, \) where for notational simplicity, we denote \( a(\delta) \) by \( a. \) We will prove that the choice on \( \lambda' \) ensures the condition (68) is satisfied, and thus with probability at least \( 1 - \delta, \) (70) holds. Obviously, one can easily prove that \( a \geq 1. \) Therefore, \( \lambda' \geq \lambda, \) and
\[ \|T_{\lambda'}^{1/2}T_{\lambda}^{-1/2}\| \leq \|T_{\lambda'}^{1/2}T_{\lambda}^{-1/2}\|\|T_{\lambda'}^{1/2}T_{\lambda}^{1/2}\|\|T_{\lambda}^{1/2}T_{\lambda}^{-1/2}\| \leq \|T_{\lambda'}^{-1/2}T_{\lambda}^{-1/2}\|^2 \|T_{\lambda}^{1/2}\| \sqrt{\lambda'/\lambda}, \]
where for the last inequality, we used \( \|T_{\lambda'}^{1/2}T_{\lambda}^{-1/2}\|^2 \leq \sup_{u \geq 0} \frac{u + \lambda}{u + \lambda} \leq \lambda'/\lambda. \) Similarly,
\[ \|T_{\lambda'}^{-1/2}T_{\lambda}^{1/2}\| \leq \|T_{\lambda'}^{-1/2}T_{\lambda}^{1/2}\| \sqrt{\lambda'/\lambda}. \]
Combining with (70), and by a simple calculation, one can prove the desired bounds. What remains is to prove that the condition (68) is satisfied. By Assumption 3 and \( a \geq 1, \) for \( \lambda = |x|^{-\theta} \) with \( \theta \in [0, 1], \)
\[ \beta \leq \log \frac{4\kappa^2(1 + c_\gamma a^{-\gamma}|x|^{\theta\gamma})}{\log \|T\|} \leq \log \frac{4\kappa^2(1 + c_\gamma)|x|^{\theta\gamma}}{\log \|T\|} = \log \frac{4\kappa^2(1 + c_\gamma)}{\log \|T\|} + \log |x|^{\theta\gamma}, \]
while for \( \lambda = (1 \vee \log |x|^{\gamma})/|x|, \)
\[ \beta \leq \log \frac{4\kappa^2(1 + c_\gamma a^{-\gamma}|x|^{-\theta\gamma})}{\log \|T\|} \leq \log \frac{4\kappa^2(1 + c_\gamma)|x|^{\gamma}}{\log \|T\|} = \log \frac{4\kappa^2(1 + c_\gamma)}{\log \|T\|} + \log |x|^\gamma, \]
If \( \lambda = |x|^{-\theta} \) with \( \theta \in [0, 1] \) and \( \theta\gamma = 0, \) or \( \lambda = (1 \vee \log |x|^{\gamma})/|x|, \) then the condition (68) follows trivially. Now consider the case \( \lambda = |x|^{-\theta} \) with \( \theta \in (0, 1) \) and \( \theta\gamma \neq 0. \) The maximum of the function \( g(u) = e^{-cu}u^\alpha \) (with \( c > 0 \)) over \( \mathbb{R}_+ \) is achieved at \( u_{\max} = \alpha/c, \) and thus
\[ \sup_{u \geq 0} e^{-cu}u^\alpha = \left(\frac{\alpha}{ec}\right)^\alpha. \]
We apply the above with \( u = |x|^{\theta\gamma\zeta'}, \alpha = 1/\zeta', \) we know that for any \( \zeta', \zeta' > 0 \)
\[ \beta \leq \log \frac{4\kappa^2(1 + c_\gamma)}{\log \|T\|} + c'|x|^{\theta\gamma\zeta'} + 1/\zeta' \log \frac{1}{\zeta'ec}. \]
Selecting \( \zeta' = \frac{1 - \theta}{\theta\gamma} \) and \( \zeta' = \frac{\theta\gamma}{e(1 - \theta)}, \) we know that a sufficient condition for (68) is
\[ \frac{|x|^{1 - \theta}a}{8\kappa^2} \geq \log \frac{4\kappa^2(1 + c_\gamma)}{\log \|T\|} + \frac{\theta\gamma}{e(1 - \theta)}|x|^{1 - \theta}. \]
From the definition of \( a, \) and by a direct calculation, one can prove that the condition (68) is satisfied.
A.2. Proof of Lemma 16

To prove the result, we need the following concentration inequality.

**Lemma 24** Let $w_1, \cdots, w_m$ be i.i.d random variables in a separable Hilbert space with norm $\| \cdot \|$. Suppose that there are two positive constants $B$ and $\sigma^2$ such that

$$E[\|w_1 - E[w_1]\|^l] \leq \frac{1}{2} l! B^{l-2} \sigma^2, \quad \forall l \geq 2. \tag{72}$$

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

$$\left\| \frac{1}{m} \sum_{k=1}^{m} w_m - E[w_1] \right\| \leq 2 \left( \frac{B}{m} + \frac{\sigma}{\sqrt{m}} \right) \log \frac{2}{\delta}. \tag{73}$$

In particular, (72) holds if

$$\|w_1\| \leq B/2 \ a.s., \quad \text{and} \quad E[\|w_1\|^2] \leq \sigma^2. \tag{74}$$

The above lemma is a reformulation of the concentration inequality for sums of Hilbert-space-valued random variables from (Pinelis and Sakhanenko, 1986). We refer to (Smale and Zhou, 2007; Caponnetto and De Vito, 2007) for the detailed proof.

**Proof of Lemma 16** We use Lemma 24 to prove the result. We let $\xi_i = T_{\lambda}^{-\frac{1}{2}}(\langle \omega_H^\lambda, x_i \rangle_H - y_i)x_i$ for all $i \in [n]$. It is easy to see that $\xi_i$ is a random variable depending on $(x_i, y_i)$. From the definition of the regression function $f_\rho$ in (8) and (9), a simple calculation shows that

$$E[\xi] = E[T_{\lambda}^{-\frac{1}{2}}(\langle \omega_H^\lambda, x \rangle_H - f_\rho(x))x] = T_{\lambda}^{-\frac{1}{2}}(T \omega_H^\lambda - S_\rho f_\rho) = T_{\lambda}^{-\frac{1}{2}}(T \omega_H^\lambda - S_\rho^* f_H). \tag{75}$$

Combining with the definition of $T_\lambda$ and $S_\rho^*$, we have

$$\|T_{\lambda}^{-\frac{1}{2}}(T_\lambda \omega_H^\lambda - S_\rho^* y - \omega_H^\lambda + S_\rho^* f_H)\|_H = \left\| \frac{1}{n} \sum_{i=1}^{n} (\xi_i - E[\xi]) \right\|_H \tag{76}$$

In order to apply Lemma 24, we need to estimate $E[\|\xi - E[\xi]\|^l_H]$ for any $l \in N$ with $l \geq 2$. In fact, using Hölder’s inequality twice,

$$E[\|\xi - E[\xi]\|^l_H] \leq E(\|\xi\|_H + E[\|\xi\|_H])^l \leq 2^{l-1}(E[\|\xi\|_H^l + (E[\|\xi\|_H])^l]) \leq 2^l E[\|\xi\|_H^l]. \tag{77}$$

We now estimate $E[\|\xi\|_H^l]$. By Hölder’s inequality,

$$E[\|\xi\|_H^l] = E[\|T_{\lambda}^{-\frac{1}{2}} x\|_H^l y - \langle \omega_H^\lambda, x \rangle_H^l] \leq 2^{l-1} E[\|T_{\lambda}^{-\frac{1}{2}} x\|_H^l (|y - f_\rho(x)|^l + |f_\rho(x) - \langle \omega_H^\lambda, x \rangle_H^l|)]. \tag{78}$$

According to (2), one has

$$\|T_{\lambda}^{-\frac{1}{2}} x\|_H \leq \|T_{\lambda}^{-\frac{1}{2}} \|_H \|x\|_H \leq \frac{1}{\sqrt{\lambda}}. \tag{79}$$
Moreover, by Cauchy-Schwarz inequality and (2),\( |\langle \omega_H^\lambda, x \rangle_H| \leq \|\omega_H^\lambda\|_H \|x\|_H \leq \kappa \|\omega_H^\lambda\|_H \).

Thus, with \( |f_\rho(x)| \leq M \) by Assumption 1, we get

\[
E[\|\xi\|_H^4] \leq 2^{l-1} \left( \frac{\kappa}{\sqrt{\lambda}} \right)^{l-2} E[\|T^{-\frac{1}{2}}_\lambda x\|_H^2|y - f_\rho(x)|] + (M + \kappa \|\omega_H^\lambda\|_H)^{-2} \|\omega_H^\lambda, x\rangle_H - f_\rho(x)^2).
\]

(77)

Note that by (11),

\[
E[\|T^{-\frac{1}{2}}_\lambda x\|_H^2|y - f_\rho(x)|] = \int_H \|T^{-\frac{1}{2}}_\lambda x\|_H^2 \int_{\mathbb{R}} |y - f_\rho(x)|^2 d\rho(y|x) d\rho_X(x)
\]

\[
\leq \frac{1}{2} M^{l-2} Q^2 \int_H \|T^{-\frac{1}{2}}_\lambda x\|_H^2 d\rho_X(x).
\]

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

\[
\int_H \|T^{-\frac{1}{2}}_\lambda x\|_H^2 d\rho_X(x) = \int_H \text{tr}(T^{-\frac{1}{2}}_\lambda x \otimes x T^{-\frac{1}{2}}_\lambda) d\rho_X(x) = \text{tr}(T^{-\frac{1}{2}}_\lambda T T^{-\frac{1}{2}}_\lambda) = \mathcal{N}(\lambda),
\]

we get

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

(79)

Besides, by Cauchy-Schwarz inequality,

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

By (76) and (39),

\[
E[\|T^{-\frac{1}{2}}_\lambda x\|_H^2(|\langle \omega_H^\lambda, x \rangle_H - f_H|)^2] \leq \frac{\kappa^2}{\lambda} E[|\langle \omega_H^\lambda, x \rangle_H - f_H|^2] = \frac{\kappa^2}{\lambda} \|S_\rho \omega_H^\lambda - f_H\|_2^2 \leq R^2 \kappa^2 \lambda^{2\kappa - 1}.
\]

Therefore,

\[
E[\|T^{-\frac{1}{2}}_\lambda x\|_H^2|\langle \omega_H^\lambda, x \rangle_H - f_\rho(x)^2|] \leq 2 \left( R^2 \kappa^2 \lambda^{2\kappa - 1} + E[\|T^{-\frac{1}{2}}_\lambda x\|_H^2|f_H(x) - f_\rho(x)|^2] \right).
\]

Using \( \|w\|_H^2 = \text{tr}(w \otimes w) \) and (12), we have

\[
E[\|T^{-\frac{1}{2}}_\lambda x\|_H^2|f_H(x) - f_\rho(x)|^2] = E[|f_H(x) - f_\rho(x)|^2 \text{tr}(T^{-\frac{1}{2}}_\lambda x \otimes x T^{-\frac{1}{2}}_\lambda)]
\]

\[
= \text{tr}(T^{-1}_\lambda E[|f_H(x) - f_\rho(x)|^2 x \otimes x])
\]

\[
\leq B^2 \text{tr}(T^{-1}_\lambda T) = B^2 \mathcal{N}(\lambda),
\]

and therefore,

\[
E[\|T^{-\frac{1}{2}}_\lambda x\|_H^2|\langle \omega_H^\lambda, x \rangle_H - f_\rho(x)^2|] \leq 2 \left( \kappa^2 R^2 \lambda^{2\kappa - 1} + B^2 \mathcal{N}(\lambda) \right).
\]
Introducing the above estimate and (79) into (77), we derive
\[
\mathbb{E}\|\xi\|_H^2 \leq 2^{l-1} \left( \frac{\kappa \sqrt{\lambda}}{\sqrt{\lambda}} \right)^{l-2} \left( \frac{1}{2} \|M\|_2 Q^2 \mathcal{N}(\lambda) + 2(M + \kappa \|\omega_H^\lambda\|_H) \|L^2(\kappa^2 \lambda^{2\zeta-1} + B^2 \mathcal{N}(\lambda)) \right)
\]
\[
\lesssim 2^{l-1} \left( \frac{\kappa M + \kappa^2 \|\omega_H^\lambda\|_H}{\sqrt{\lambda}} \right)^{l-2} \frac{1}{2} ! \left( 2R^2 \kappa^2 \lambda^{2\zeta-1} + (2B^2 + Q^2) \mathcal{N}(\lambda) \right).
\]

Introducing the above estimate into (75), and then substituting with (40), we get
\[
\mathbb{E}[\|\xi - E[\xi]\|_H^2] \leq \frac{1}{2} ! \left( \frac{4\kappa (M + \kappa^{1/2} \lambda R \mathcal{N}(\frac{\kappa}{\lambda} - \frac{1}{2}))}{\sqrt{\lambda}} \right)^{l-2} 8 \left( 2R^2 \kappa^2 \lambda^{2\zeta-1} + (2B^2 + Q^2) \mathcal{N}(\lambda) \right).
\]

Applying Lemma 24, we get the desired result. \hfill \qed

A.3. Proof of Lemma 17

Let \( S_X = U \Sigma V^* \) be the singular value decomposition of \( S_X \), where \( V : \mathbb{R}^r \to H, U \in \mathbb{R}^{n \times r} \) and \( \Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r) \) with \( V^* V = I_r, U^* U = I_r \) and \( \sigma_1 \geq \sigma_2, \ldots, \sigma_r > 0 \). In fact, we can write \( V = [v_1, \ldots, v_r] \) with
\[
V a = \sum_{i=1}^r a(i) v_i, \quad \forall a \in \mathbb{R}^r,
\]
with \( v_i \in H \) such that \( \langle v_i, v_j \rangle_H = 0 \) if \( i \neq j \) and \( \langle v_i, v_i \rangle_H = 1 \). Similarly, we write
\( U = [u_1, \ldots, u_r] \), and
\[
S_X = \sum_{i=1}^r \sigma_i \langle v_i, \cdot \rangle_H u_i = \sum_{i=1}^r \sigma_i u_i \otimes v_i.
\]

For any \( \mu \geq 0 \), we decompose \( S_X \) as \( S_{1,\mu} + S_{2,\mu} \) with
\[
S_{1,\mu} = \sum_{\sigma_i > \mu} \sigma_i u_i \otimes v_i, \quad S_{2,\mu} = \sum_{\sigma_i \leq \mu} \sigma_i u_i \otimes v_i,
\]
and we will drop \( \mu \) to write \( S_{j,\mu} \) as \( S_j \) when it is clear in the text. Denote \( d \) the cardinality of \( \{\sigma_i : \sigma_i > \mu\} \). Correspondingly,
\[
S_1 = U_1 \Sigma_1 V_1^*, \quad S_2 = U_2 \Sigma_2 V_2^*, \quad (80)
\]
where \( V_1 = [v_1, \ldots, v_d], V_2 = [v_{d+1}, \ldots, v_r], U_1 = [u_1, \ldots, u_d], U_2 = [u_{d+1}, \ldots, u_r], \Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_d) \), and \( \Sigma_2 = \text{diag}(\sigma_{d+1}, \ldots, \sigma_r) \). As the range of \( P \) is range(\( S_2^* G^* \)), we can let
\[
P = P_1 + P_2,
\]
where \( P_1 \) and \( P_2 \) are projection operators on range(\( S_1^* G^* \)) and range(\( S_2^* G^* \)), respectively. As
\[
T_X = S_X^* S_X = (U \Sigma V^*)^* U \Sigma V^* = V \Sigma^2 V^*,
\]

34
we have
\[\|(I - P)T_2^\frac{1}{2}\| = \|(I - P)V\Sigma V^*\| = \|(I - P_1 - P_2)\sum_{i=1}^{2} V_i\Sigma_i V_i^*\|.

As \(P_1\) is a projection operator on \(\text{range}(S_i^*G^*)\) (\(\subseteq\) \(\text{range}(V_1)\)) and \(\text{range}(S_i^*G^*)\) (\(\subseteq\) \(\text{range}(V_2)\)), and \(V_i^*V_2 = 0\), we know that \(P_iV_j = 0\) when \(i \neq j\). Thus, it follows that
\[\|(I - P)T_2^\frac{1}{2}\| = \sum_{i=1}^{2} \|(I - P_i)(V_i\Sigma_i V_i^*)\|
\leq \sum_{i=1}^{2} \|(I - P_i)(V_i\Sigma_i V_i^*)\|
\leq \|(I - P_1)(V_1\Sigma_1 V_1^*)\| + \|(I - P_2\|\|\Sigma_2\|\|V_2\|\|.

As \(\Sigma_2 = \text{diag}(\sigma_{d+1}, \ldots, \sigma_r)\) with \(\sigma_r \leq \cdots \sigma_{d+1} \leq \mu\), we get
\[\|(I - P)T_2^\frac{1}{2}\| \leq \|(I - P_1)(V_1\Sigma_1 V_1^*)\| + \mu. \tag{81}\]

As \(P_1\) is the projection operator on \(\text{range}(S_i^*G^*)\), letting \(W = GS_1\) and for any \(\lambda > 0\),
\[P_1 = W^*(WW^*)^{-1}W \geq W^*(WW^* + \lambda I)^{-1}W = W^*W(W^*W + \lambda I)^{-1},\]
and thus
\[I - P_1 \leq I - W^*W(W^*W + \lambda I)^{-1} = \lambda(W^*W + \lambda I)^{-1}.

It thus follows that
\[T_1^\frac{1}{2}(I - P_1)T_1^\frac{1}{2} \leq \lambda T_1^\frac{1}{2}(W^*W + \lambda I)^{-1}T_1^\frac{1}{2},\]
where for notational simplicity, we write
\[T_1 = (V_1\Sigma_1 V_1^*)^2. \tag{82}\]

Combing with
\[\|(I - P)T_2^\frac{1}{2}\|^2 = \|T_1^\frac{1}{2}(I - P)^2T_1^\frac{1}{2}\| = \|T_1^\frac{1}{2}(I - P)T_1^\frac{1}{2}\|,
\]
we know that
\[\|(I - P)T_1^\frac{1}{2}\|^2 \leq \lambda\|T_1^\frac{1}{2}(W^*W + \lambda I)^{-1}T_1^\frac{1}{2}\| \leq \lambda\|T_{1\lambda}^\frac{1}{2}(W^*W + \lambda I)^{-1}T_{1\lambda}^\frac{1}{2}\|.

As
\[T_{1\lambda}^\frac{1}{2}(W^*W + \lambda I)^{-1}T_{1\lambda}^\frac{1}{2} = \left(T_{1\lambda}^{-\frac{1}{2}}(W^*W + \lambda I)T_{1\lambda}^{-\frac{1}{2}}\right)^{-1} = \left(I - T_{1\lambda}^{-\frac{1}{2}}(T_1 - W^*W)T_{1\lambda}^{-\frac{1}{2}}\right)^{-1},\]
and if
\[\|T_{1\lambda}^{-\frac{1}{2}}(T_1 - W^*W)T_{1\lambda}^{-\frac{1}{2}}\| \leq c < 1, \tag{83}\]
then according to Neumann series,
\[
\| (I - P) T_{\lambda}^{-\frac{1}{2}} \|^2 \leq \lambda \| T_{\lambda}^{-\frac{1}{2}} (W^* W + \lambda I)^{-1} T_{\lambda}^{-\frac{1}{2}} \| \leq (1-c)^{-1} \lambda.
\]
(84)
If we choose \( \mu = \sqrt{\lambda} \), and introduce the above with \( c = \frac{1}{2} \) into (81), one can get
\[
\| (I - P) T_{\lambda}^{-\frac{1}{2}} \|^2 \leq (\sqrt{2} + 1)^2 \lambda \leq 6 \lambda,
\]
(85)
which leads to the desired bound.

In what follows, we show that (83) with \( c = \frac{1}{2} \) holds in high probability under the constraint (58). Recall (82) and that \( W = G S_1 \) with \( S_1 \) given by (80). Thus, \( T_1 = V_1 \Sigma_1 V^*_1 \Sigma_1 V_1^* = V_1 \Sigma_1^2 V_1^* \), and
\[
W^* W = S_1^* G^* G S_1 = V_1 \Sigma_1 U^*_1 G^* G U_1 \Sigma_1 V_1^*.
\]
Therefore, with \( V_1^* V_1 = I \),
\[
T_{\lambda}^{-\frac{1}{2}} (T_1 - W^* W) T_{\lambda}^{-\frac{1}{2}} = V_1 (\Sigma_1^2 + \lambda I)^{-1/2} V^*_1 \Sigma_1 (I - U_1 G^* G U_1) \Sigma_1 V^*_1 V_1 (\Sigma_1^2 + \lambda I)^{-1/2} V_1^* = V_1 (\Sigma_1^2 + \lambda I)^{-1/2} \Sigma_1 (I - U_1 G^* G U_1) \Sigma_1 (\Sigma_1^2 + \lambda I)^{-1/2} V_1^*.
\]
(86)
It follows that
\[
\| T_{\lambda}^{-\frac{1}{2}} (T_1 - W^* W) T_{\lambda}^{-\frac{1}{2}} \| \leq \| V_1 \| \| (\Sigma_1^2 + \lambda I)^{-1/2} \Sigma_1 \| \| I - U_1 G^* G U_1 \| \| V_1^* \| \leq \| I - U_1 G^* G U_1 \|.
\]
Using \( U_1^* U_1 = I \),
\[
\| I - U_1 G^* G U_1 \| = \| U_1^* (I - G^* G) U_1 \| = \max_{a \in \mathbb{R}^d, \| a \|_2 = 1} | \langle U_1^* (I - G^* G) U_1, a \rangle | = \max_{a \in \mathbb{R}^d, \| a \|_2 = 1} \| U_1 a \|^2 - \| G U_1 a \|^2.
\]
Based on a standard argument as that for (Baraniuk et al., 2008, Lemma 5.1), we know that
\[
\max_{a \in \mathbb{R}^d, \| a \|_2 = 1} \| U_1 a \|^2 - \| G U_1 a \|^2 \leq \frac{1}{2}
\]
with probability at least
\[
1 - 2(60)^d \exp \left( -\frac{m}{100c_0' \log^6 n} \right) \geq 1 - \delta,
\]
provided that
\[
m \geq 100c_0' \log^6 n \left( \log \frac{2}{\delta} + 5d \right).
\]
(87)
Note that by (57),
\[
b \lambda^{-\gamma} \geq \text{tr} (T \gamma T_\lambda^{-1}) = \sum_i \frac{\sigma_i}{\sigma_i^2 + \lambda} \geq \sum \frac{\sigma_i^2}{\sigma_i^2 + \lambda} \geq \frac{d}{2}.
\]
Thus, a stronger condition for (87) is (58). The proof is complete.
A.4. Proof of Lemma 18
We first use Lemma 24 to estimate $\text{tr}(T_\lambda^{-\frac{1}{2}}(T_x - T)T_\lambda^{-\frac{1}{2}})$. Note that
\[ \text{tr}(T_\lambda^{-\frac{1}{2}}T_x T_\lambda^{-\frac{1}{2}}) = \frac{1}{n} \sum_{j=1}^{n} \|T_\lambda^{-\frac{1}{2}}x_j\|_H^2 = \frac{1}{n} \sum_{j=1}^{n} \xi_j, \]
where we let $\xi_j = \|T_\lambda^{-\frac{1}{2}}x_j\|_H^2$ for all $j \in [n]$. Besides, it is easy to see that
\[ \text{tr}(T_\lambda^{-\frac{1}{2}}(T_x - T)T_\lambda^{-\frac{1}{2}}) = \frac{1}{n} \sum_{j=1}^{n} (\xi_j - \mathbb{E}[\xi_j]). \]
Using Assumption (2),
\[ \xi_1 \leq \frac{1}{\lambda} \|x_1\|_H^2 \leq \frac{\kappa^2}{\lambda}, \]
and
\[ \mathbb{E}[\|\xi_1\|^2] \leq \frac{\kappa^2}{\lambda} \mathbb{E}[\|T_\lambda^{-\frac{1}{2}}x_1\|_H^2] = \frac{\kappa^2}{\lambda} \mathbb{E} \text{tr}(T_\lambda^{-\frac{1}{2}}x_1 \otimes x_1 T_\lambda^{-\frac{1}{2}}) = \frac{\kappa^2 \mathcal{N}(\lambda)}{\lambda}. \]
Applying Lemma 24, we get that there exists a subset $\Omega_1$ of $H^n$ with measure at least $1 - \delta$, such that for all $x \in \Omega_1$,
\[ \text{tr}(T_\lambda^{-\frac{1}{2}}(T_x - T)T_\lambda^{-\frac{1}{2}}) \leq 2 \left( \frac{2\kappa^2}{n\lambda} + \sqrt{\frac{\kappa^2 \mathcal{N}(\lambda)}{n\lambda}} \right) \log \frac{2}{\delta}. \]
Combining with Lemma 14, taking the union bounds, rescaling $\delta$, and noting that
\[ \text{tr}(T_\lambda^{-1}T_x) = \text{tr}(T_\lambda^{-\frac{1}{2}}T_\lambda^\frac{1}{2}T_\lambda^{-\frac{1}{2}}T_x T_\lambda^{-\frac{1}{2}}T_\lambda^{-\frac{1}{2}}T_\lambda^\frac{1}{2}T_\lambda^{-\frac{1}{2}}T_x T_\lambda^{-\frac{1}{2}}) \]
\[ \leq \|T_\lambda^{-\frac{1}{2}}T_\xi^{-\frac{1}{2}}\|_2^2 \text{tr}(T_\lambda^{-\frac{1}{2}}T_x T_\lambda^{-\frac{1}{2}}) \]
\[ = \|T_\lambda^{-\frac{1}{2}}T_\xi^{-\frac{1}{2}}\|_2^2 \left( \text{tr}(T_\lambda^{-\frac{1}{2}}(T_x - T)T_\lambda^{-\frac{1}{2}}) + \mathcal{N}(\lambda) \right). \]
we get that there exists a subset $\Omega$ of $H^n$ with measure at least $1 - \delta$, such that for all $x \in \Omega$,
\[ \text{tr}(T_\lambda^{-1}T_x) \leq 3a(\delta/2) \left( 2 \left( \frac{2\kappa^2}{n\lambda} + \sqrt{\frac{\kappa^2 \mathcal{N}(\lambda)}{n\lambda}} \right) \log \frac{4}{\delta} + \mathcal{N}(\lambda) \right), \]
which leads to the desired result using $\lambda \leq 1$, $n\lambda \geq 1$ and Assumption 3.

A.5. Estimating Projection Errors with Random Sketches

**Proof of Lemma 19** Let $\mu = \frac{1}{\sqrt{n}}\log n^\gamma$, and $\lambda = n^{-\theta}$ with $\theta \in [0, 1)$ or $\lambda = \frac{1}{\sqrt{n}}\log n^\gamma$. By a simple calculation,
\[ \|(I - P)T^\frac{1}{2}\|_2^2 \leq \|(I - P)T^\frac{1}{2}\mu\|^2_2 \|T^\frac{1}{2}\mu\|^2_2. \]
Using
\[(I - P)^{1/2}T_{\mu}^{\frac{1}{2}} = \| (I - P)^{1/2}T_{\mu}^L(I - P) \| \leq \| (I - P)^{1/2}T_L^L(I - P) \| + \mu \| (I - P)^{1/2} \| \leq \| (I - P)^{1/2}L^\frac{1}{2} \|^2 + \mu, \]
we get
\[\| (I - P)^{1/2} \|^2 \leq \left( \| (I - P)^{1/2}L^\frac{1}{2} \|^2 + \mu \right) \| T_{\mu}^L \|^2 \| T_{\mu}^L \|^2. \] (88)

Following from Lemma 18 and Lemma 14, we know that there exists a subset $\Omega_1$ of $H^n$ with measure at least $1 - 2\delta$ such that for every $x \in \Omega_1$,
\[
\text{tr}(T_{\mu\lambda}^{-1}T_L) \leq b_{\gamma,\delta} \lambda^{-\gamma},
\]
and
\[\| T_{\mu\lambda}^{-1}T_L \|^2 \leq a_{\gamma} \log \frac{4}{\delta}, \] (89)
where $b_{\gamma,\delta} = b_{\gamma} \log^2 \frac{4}{\delta}$. For every $x \in \Omega_1$, according to Lemma 17, we know that there exists a subset $U_x$ of $R^{m \times n}$ with measure at least $1 - \delta$, such that for all $G \in U_x$,
\[\| (I - P)^{1/2}L^\frac{1}{2} \| \leq 6\lambda, \] (90)
provided that,
\[m \geq 100c_0 \log^3 n \lambda^{-\gamma} \log^3 \frac{4}{\delta} (1 + 10b_{\gamma}), \]
which is satisfied under the constraint (60). From the above analysis, we can conclude that if (60) holds, then with probability at least $1 - 3\delta$, (90) and (89) hold. Introducing (90) and (89) into (88), one gets that with probability at least $1 - 3\delta$,
\[\| (I - P)^{1/2} \|^2 \leq (6\lambda + \mu) a_{\gamma} \log \frac{4}{\delta}, \]
which leads to the desired result. \hfill \blacksquare


**Proof of Lemma 20**

As $P$ is the projection operator onto $\text{range}\{S_{\bar{x}}^\perp\}$ with $\bar{x} = \{x_1, \ldots, x_m\}$,
\[P = S_{\bar{x}}^\perp(S_{\bar{x}}S_{\bar{x}}^\perp)^{\dagger}S_{\bar{x}} \geq S_{\bar{x}}^\perp(S_{\bar{x}}S_{\bar{x}}^\perp + \mu I)^{-1}S_{\bar{x}} = S_{\bar{x}}^\perp(S_{\bar{x}}S_{\bar{x}}^\perp + \mu I)^{-1} = T_{\bar{x}}(T_{\bar{x}} + \mu I)^{-1}, \]
where for the last second equality, we used Lemma 9. Thus,
\[I - P \leq I - T_{\bar{x}}(T_{\bar{x}} + \mu I)^{-1} = \mu(T_{\bar{x}} + \mu I)^{-1}. \]

It thus follows that
\[T_{\mu}^{1/2}(I - P)^{1/2}T_{\mu}^{1/2} \leq \mu T_{\mu}^{1/2}(T_{\bar{x}} + \mu I)^{-1}T_{\mu}^{1/2}. \]

Using $\| A^* A \|^2 = \| A \|^2$ and the above,
\[\| (I - P)^{1/2} \|^2 = \| T_{\mu}^{1/2}(I - P)^{1/2}T_{\mu}^{1/2} \| \leq \mu \| T_{\mu}^{1/2}(T_{\bar{x}} + \mu I)^{-1/2}T_{\mu}^{1/2} \|^2 = \mu \| (T_{\bar{x}} + \mu I)^{-1/2} \|^2. \] (91)

Thus,
\[\| (I - P)^{1/2} \|^2 \leq \| (I - P)^{1/2} \|^2 \leq \mu \| (T_{\bar{x}} + \mu I)^{-1/2} \|^2. \]

Using Lemma 14 with $\mu = \frac{\sqrt{\log m\delta}}{m}$, one can prove the desired result. \hfill \blacksquare
A.7. Estimating Projection Errors with ALS Nyström Subsampling

We first note that in an L-ALS Nyström subsampling regime, \( S \) can be rewritten as \( S = \text{range}\{S_x^T G^T\} \), where each row \( \frac{1}{\sqrt{m}} a_j^T \) of \( G \) is i.i.d. drawn according to

\[
\mathbb{P}\left( a = \frac{1}{\sqrt{q_i}} e_i \right) = q_i, \quad i \in \{1, \ldots, n\}
\]

Here \( \{e_i : i \in [n]\} \) is the standard basis of \( \mathbb{R}^n \) and

\[
q_i := q_i(\lambda) = \frac{i_i(\lambda)}{\sum_j i_j(\lambda)}.
\]

Using Lemma 22 and with a similar argument as that for Lemma 17, we can estimate the empirical version of the projection error as follows.

**Lemma 25** Let \( 0 < \delta < 1 \) and \( \theta \in [0, 1] \). Given a fix input subset \( x \subseteq H^n \), assume that for \( \lambda \in [0, 1] \), (57) holds for some \( b_\gamma > 0, \gamma \in [0, 1] \). Then there exists a subset \( U_x \) of \( \mathbb{R}^{m \times n} \) with measure at least \( 1 - \delta \), such that for all \( G \in U_x \),

\[
\|(I - P)T^\frac{1}{2}\|^2 \leq 3\lambda,
\]

provided that

\[
m \geq 8b_\gamma \lambda^{-\gamma} L^2 \log \frac{8b_\gamma \lambda^{-\gamma}}{\delta}.
\]

**Proof** If we choose \( u = 0 \) in the proof of Lemma 17, then \( S_x = S_1 \) and \( S_2 = 0 \). Similarly, \( T_x = T_1 \). In this case, (86) reads as

\[
T_{x\lambda}^{-\frac{1}{2}}(T_x - W^*W)T_{x\lambda}^{-\frac{1}{2}} = V(\Sigma^2 + \lambda I)^{-1/2} \Sigma(I - U^*G^*GU)\Sigma(\Sigma^2 + \lambda I)^{-1/2} V^*.
\]

Thus, using \( V^*V = I, U^*U = I \) and \( U \) is of full column rank,

\[
\|T_{x\lambda}^{-\frac{1}{2}}(T_x - W^*W)T_{x\lambda}^{-\frac{1}{2}}\| \leq \|V\|\|U^*U(\Sigma^2 + \lambda I)^{-1/2} \Sigma U^*(I - G^*G)U\Sigma(\Sigma^2 + \lambda I)^{-1/2} U^*V\|
\]

\[
\leq \|U(\Sigma^2 + \lambda I)^{-1/2} \Sigma U^*(I - G^*G)U\Sigma(\Sigma^2 + \lambda I)^{-1/2} U^*\|.
\]

Using \( K := K_{xx} = S_x S_x^T = U \Sigma^2 U^* \), we get

\[
\|T_{x\lambda}^{-\frac{1}{2}}(T_x - W^*W)T_{x\lambda}^{-\frac{1}{2}}\| \leq \|(K(K + \lambda I)^{-1})^{1/2} (I - G^*G) (K(K + \lambda I)^{-1})^{1/2}\|.
\]

Letting \( \lambda_i = (K(K + \lambda I)^{-1})^{1/2} a_i a_i^* (K(K + \lambda I)^{-1})^{1/2}, \) it is easy to prove that \( \mathbb{E}[a_i a_i^*] = I \), according to the definition of ALS Nyström subsampling. Then the above inequality can be written as

\[
\|T_{x\lambda}^{-\frac{1}{2}}(T_x - W^*W)T_{x\lambda}^{-\frac{1}{2}}\| \leq \|\frac{1}{m} \sum_{i=1}^{m} (\mathbb{E}[-\lambda_i^2 - \lambda_i])\|.
\]
A simple calculation shows that
\[ \|X_i\| = a_i^* (K(K + \lambda I)^{-1}) a_i \leq \max_{j \in [n]} \frac{(K(K + \lambda I)^{-1})_{jj}}{q_j} \]
\[ = \max_{j \in [n]} \frac{l_j(\lambda)}{q_j} = \max_{j \in [n]} \frac{l_j(\lambda) \sum_k \hat{l}_k(\lambda)}{l_j(\lambda)} \leq L^2 \sum_j l_j(\lambda) = L^2 \text{tr}(KK^{-1}), \]
and
\[ E[X_i^2] = E[a_i^* (K(K + \lambda I)^{-1}) a_i X_i] \leq L^2 \text{tr}(KK^{-1}) E[X_i] = L^2 \text{tr}(KK^{-1}) KK^{-1}. \]
Thus,
\[ \|E[X_i] - X_i\| \leq E\|X_i\| + \|X_i\| \leq 2L^2 \text{tr}(KK^{-1}), \]
and
\[ E[(X_i - E[X_i])^2] \leq E[X_i^2] \leq 2L^2 \text{tr}(KK^{-1}) KK^{-1}. \]
Letting \( V = L^2 \text{tr}(KK^{-1}) KK^{-1}, \) we have
\[ \|V\| \leq L^2 \text{tr}(KK^{-1}), \]
and
\[ \frac{\text{tr}(V)}{\|V\|} = \frac{\text{tr}(KK^{-1})}{\|KK^{-1}\|} = \text{tr}(KK^{-1}) \left(1 + \frac{\lambda}{\|K\|}\right). \]
Applying Lemma 22, noting that \( \text{tr}(KK^{-1}) = \text{tr}(T_x T_{x\lambda}^{-1}) \) and \( \|K\| = \|T_x\| \) as \( T_x = S_x S_x, \)
we get that there exists a subset \( U_x \subseteq \mathbb{R}^{m \times n} \) with measure at least \( 1 - \delta \) such that for all \( G \in U_x, \)
\[ \|T_{x\lambda}^{-\frac{1}{2}} (T_x - W^* W) T_{x\lambda}^{-\frac{1}{2}} \| \leq \frac{4L^2 \text{tr}(T_x T_{x\lambda}^{-1})\beta}{3m} + \sqrt{\frac{2L^2 \text{tr}(T_x T_{x\lambda}^{-1})\beta}{m}}, \quad \beta = \log \frac{4 \text{tr}(T_x T_{x\lambda}^{-1})(1 + \lambda/\|T_x\|)}{\delta}. \]
If \( \lambda \leq \|T_x\|, \) using Condition (57), we have
\[ \beta \leq \log \frac{4b_x \lambda^{-\gamma} (1 + \lambda/\|T_x\|)}{\delta} \leq \log \frac{8b_x \lambda^{-\gamma}}{\delta}, \]
and, combining with (93),
\[ \frac{4L^2 \text{tr}(T_x T_{x\lambda}^{-1})\beta}{3m} + \sqrt{\frac{2L^2 \text{tr}(T_x T_{x\lambda}^{-1})\beta}{m}} \leq \frac{2}{3}. \]
Thus,
\[ \|T_{x\lambda}^{-\frac{1}{2}} (T - W^* W) T_{x\lambda}^{-\frac{1}{2}} \| \leq \frac{2}{3}, \quad \forall G \in U_x. \]
Following from (83) and (84), one can prove (92) for the case \( \lambda \leq \|T_x\|. \) The proof for the case \( \lambda \geq \|T_x\| \) is trivial:
\[ \|(I - P)T_x^\frac{1}{2}\|^2 \leq \|I - P\|^2 \|T_x^\frac{1}{2}\|^2 \leq \|T_x\| \leq \lambda. \]
The proof is complete.

With the above lemma, and using a similar argument as that for Lemma 19, we can prove Lemma 21. We thus skip it.

Appendix B. Learning with Kernel Methods

In this appendix, we review how the regression setting considered in this paper covers non-parametric regression with kernel methods.

Let the input space \( \Xi \) be a closed subset of Euclidean space \( \mathbb{R}^d \), the output space \( Y \subseteq \mathbb{R} \). Let \( \mu \) be an unknown but fixed Borel probability measure on \( \Xi \times Y \). Assume that \( \{(\xi_i, y_i)\}_{i=1}^m \) are i.i.d. from the distribution \( \mu \). 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 \( \Xi \). 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: \( f(\xi) = \langle K_\xi, f \rangle_K \).

Example B.1 (Sobolev Spaces) Let \( X = [0, 1] \) and the kernel

\[
K(x, x') = \begin{cases} 
(1 - y)x, & x \leq y; \\
(1 - x)y, & x \geq y.
\end{cases}
\]

Then the kernel induces a Sobolev Space \( H = \{f : X \to \mathbb{R}| f \text{ is absolutely continuous}, f(0) = f(1) = 0, f \in L^2(X)\} \).

In learning with kernel methods, one considers the following minimization problem

\[
\inf_{f \in \mathcal{H}_K} \int_{\Xi \times Y} (f(\xi) - y)^2 d\mu(\xi, y).
\]

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

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

Letting \( X = \{K_\xi : \xi \in \Xi\} \) and defining another probability measure \( \rho(K_\xi, y) = \mu(\xi, y) \), the above reduces to the learning setting in Section 1.

References


