Journal of Machine Learning Research 21 (2020) 1-44 Submitted 1/19; Revised 12/19; Published 1/20

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

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

where the measure ρ is known only through a sample $\mathbf{z} = \{z_i = (x_i, y_i)\}_{i=1}^n$ of size $n \in \mathbb{N}$, independently and identically distributed (i.i.d.) according to ρ.

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 $\overline{\mathrm{span}\{\mathbf{x}\}}$ of H, where $\mathbf{x} = \{x_1, \cdots, x_n\}$. 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 $m < n$, $S = \overline{\mathrm{span}\{\tilde{x}_j : 1 \leq j \leq m\}}$ where $\tilde{x}_j$ is chosen randomly from the input set $\mathbf{x}$, and more generally, $S = \overline{\mathrm{span}\{\sum_{j=1}^n G_{ij} x_j : 1 \leq i \leq m\}}$ where $\mathbf{G} = [G_{ij}]_{1\leq i\leq m, 1\leq j\leq n}$ is a general randomized matrix whose rows are drawn according to a distribution. We call[1] 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 $\mathbf{y} = \{y_1, \cdots, y_n\}$ 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 $\mathbf{G}$ 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 $\mathbf{z}$ 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 $\mathbf{z}$ 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. Section 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 ρ, and let $\rho(\cdot|x)$ be the conditional probability measure on $\mathbb{R}$ with respect to $x \in H$ and ρ. For simplicity, we assume that the support of ρ_X is compact and that there exists a constant $\kappa \in [1, \infty[$, such that

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

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 ρ_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 $\{1, \cdots, r\}$ is denoted by $[r]$. For any real number a, $a_+ = \max(a, 0)$, $a_- = \min(0, a)$.

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

$$\|\mathcal{L}\| = \|\mathcal{T}\| \le \operatorname{tr}(\mathcal{T}) = \int_H \operatorname{tr}(x \otimes x) d\rho_X(x) = \int_H \|x\|^2_H d\rho_X(x) \le \kappa^2. \tag{3}$$

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

$$\|\mathcal{S}_\rho \omega\|_\rho = \|\mathcal{T}^{\frac{1}{2}} \omega\|_H. \tag{4}$$

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

$$\|\mathcal{L}^{-\frac{1}{2}} \mathcal{S}_\rho \omega\|_\rho \le \|\omega\|_H. \tag{5}$$

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

$$\|\mathcal{T}_{\mathbf{x}}\| \leq \operatorname{tr}(\mathcal{T}_{\mathbf{x}}) \leq \kappa^2. \tag{6}$$

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

$$\inf_{f \in H_\rho} \mathcal{E}(f), \quad \mathcal{E}(f) = \int_{H \times \mathbb{R}} (f(x) - y)^2 d\rho(x,y), \tag{7}$$

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,

$$f_\rho(x) = \int_{\mathbb{R}} y d\rho(y|x), \qquad x \in H, \rho_X\text{-almost surely.} \tag{8}$$

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_ρ 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_ρ onto the closure of H_ρ in $L^2_{\rho_X}$, and moreover, for all $f \in H_\rho$ (Lin and Rosasco, 2017b),

$$\mathcal{S}_\rho^* f_\rho = \mathcal{S}_\rho^* f_H, \tag{9}$$

and

$$\mathcal{E}(f) - \mathcal{E}(f_H) = \|f - f_H\|_\rho^2. \tag{10}$$

Note that f_H does not necessarily lie in H_ρ.

Throughput 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 $l \in \mathbb{N}$,*

$$\int_{\mathbb{R}} |y - f_\rho(x)|^l d\rho(y|x) \leq \frac{1}{2} l! M^{l-2} Q^2, \tag{11}$$

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

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

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

Assumption 2 *f_H satisfies*

$$\int_H (f_H(x) - f_\rho(x))^2 x \otimes x d\rho_X(x) \preceq B^2 \mathcal{T}, \tag{12}$$

and the following Hölder source condition

$$f_H = \mathcal{L}^\zeta g_0, \quad \textit{with} \quad \|g_0\|_\rho \leq R. \tag{13}$$

Here, B, R, ζ 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 ζ 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

$$\omega_H = \mathcal{T}^{\zeta-1} \mathcal{S}_\rho^* g_0.$$

Using the singular value decomposition of $\mathcal{S}_\rho$, one can prove that $\mathcal{S}_\rho \omega_H = f_H$.

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

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

$$\mathcal{N}(\lambda) := \operatorname{tr}(\mathcal{T}(\mathcal{T} + \lambda I)^{-1}) \leq c_\gamma \lambda^{-\gamma}, \quad \textit{for all } \lambda > 0. \tag{14}$$

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 $\widetilde{\mathcal{E}}(\omega)$ in (1) can not be computed exactly. It can be only approximated through the empirical risk $\widetilde{\mathcal{E}}_{\mathbf{z}}(\omega)$,

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

2. This can be verified by using (10).

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, $\inf_{\omega \in S} \widetilde{\mathcal{E}}_{\mathbf{z}}(\omega)$. 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 $P\mathcal{T}_{\mathbf{x}} P\hat{\alpha} = P\mathcal{S}_{\mathbf{x}}^* \bar{\mathbf{y}}$. The inversion of the linear operator $P\mathcal{T}_{\mathbf{x}}P$ may have a bad condition number or be unbounded . Motivated by the classic (iterated) ridge regression, we replace the inversion of $P\mathcal{T}_{\mathbf{x}}P$ 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 τ over the sample $\mathbf{z}$ and subspace S is given by $f_\lambda^{\mathbf{z}} = \mathcal{S}_\rho \omega_\lambda^{\mathbf{z}}$, where*[3]

$$\omega_\lambda^{\mathbf{z}} = P\mathcal{G}_\lambda(P\mathcal{T}_{\mathbf{x}}P)P\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}}, \quad \mathcal{G}_\lambda(u) = \sum_{i=1}^{\tau} \lambda^{i-1}(\lambda+u)^{-i}. \tag{15}$$

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

$$\sup_{\lambda \in \Lambda} \sup_{u \in]0,\kappa^2]} |\mathcal{G}_\lambda(u)(u+\lambda)| \leq E. \tag{16}$$

and

$$\sup_{\alpha\in[0,\tau]} \sup_{\lambda \in \Lambda} \sup_{u \in]0,\kappa^2]} |1 - \mathcal{G}_\lambda(u)u|(u+\lambda)^\alpha \lambda^{-\alpha} \leq F. \tag{17}$$

The filter function $\mathcal{G}_\lambda(u)$ 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

$$\mathcal{G}_\lambda(u) = \frac{1-q^\tau}{u} = \frac{\sum_{i=0}^{\tau-1} q^i}{u+\lambda}, \quad q = \frac{\lambda}{\lambda+u}. \tag{18}$$

Thus, $\mathcal{G}_\lambda(u)$ is a filter function with qualification τ, $E = \tau$ 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

$$\mathcal{G}_\lambda(u) = \begin{cases} u^{-1}, & \text{if } u \geq \lambda, \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

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 τ could be any positive

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

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

In the above, λ 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 $f_\lambda^{\mathbf{z}}$ can be measured in terms of *excess risk* (*generalization error*), $\mathcal{E}(f_\lambda^{\mathbf{z}}) - \inf_{H_\rho} \mathcal{E} = \widetilde{\mathcal{E}}(\omega_\lambda^{\mathbf{z}}) - \inf_H \widetilde{\mathcal{E}}$, which is exactly $\|f_\lambda^{\mathbf{z}} - f_H\|_\rho^2$ according to (10). Assuming that $f_H \in H_\rho$, i.e., $f_H = \mathcal{S}_\rho \omega_*$ for some $\omega_* \in H$, it can be measured in terms of H-norm, $\|\omega_\lambda^{\mathbf{z}} - \omega_*\|_H$, which is closely related to $\|\mathcal{L}^{-\frac{1}{2}} \mathcal{S}_\rho(\omega_\lambda^{\mathbf{z}} - \omega_*)\|_H = \|\mathcal{L}^{-\frac{1}{2}}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho$, according to (5). In what follows, we will measure the performance of an estimator $f_\lambda^{\mathbf{z}}$ in terms of a broader class of norms, $\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho$, where $a \in [0, \frac{1}{2}]$ is such that $\mathcal{L}^{-a} f_H$ is well defined. In the attainable cases, i.e., $f_H \in H_\rho$, according to (5), $\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho$ is close to $\|\mathcal{T}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \omega_H)\|_H$. 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 $\kappa^2, c_\gamma, \gamma, \zeta$ B, M, Q, R, τ and $\|\mathcal{T}\|$, 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* $\lambda = \frac{1 \vee \log n^\gamma}{n}$. *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 $\zeta \in [0,1]$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \lambda^{-a} \log^2 \frac{3}{\delta} \left(\lambda^\zeta + \frac{1}{\sqrt{n\lambda^\gamma}} + \lambda^{\zeta-1}\big(\Delta_5 + \Delta_5^{1-a}\lambda^a\big) \right). \tag{20}$$

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

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \lambda^{-a} \log^2 \frac{3}{\delta} \left(\lambda^\zeta + \frac{1}{\sqrt{n\lambda^\gamma}} + (\Delta_5 + \lambda \Delta_5^{(\zeta-1)\wedge 1} + \Delta_5^{1-a}\lambda^a) \right). \tag{21}$$

Furthermore, if $\zeta \geq 1/2$, *then the above conclusions still hold if we replace* $\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho$ *by* $\|\mathcal{T}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \omega_H)\|_H$. *Here,* Δ_5 *is the projection error* $\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2$.

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 ζ, 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 λ and the following result.

Corollary 1 *Under the assumptions and notations of Theorem 1, let*

$$\lambda = n^{-\frac{1}{1\vee(2\zeta+\gamma)}} (1 \vee \log n^{\gamma})^{\mathbf{1}_{\{2\zeta+\gamma\leq 1\}}}. \tag{22}$$

Then the following statements hold with probability at least $1-\delta$.
1) If $\zeta \leq 1$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \lambda^{\zeta-a} \left(1 + \lambda^{-1}\Delta_5\right) \log^2 \frac{3}{\delta}. \tag{23}$$

2) If $\zeta \geq 1$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \lambda^{-a} \log^2 \frac{3}{\delta} \left(\lambda^{\zeta} + \Delta_5 \left(1 + \left(\frac{\lambda}{\Delta_5}\right) \Delta_5^{(\zeta-1)\wedge 1} + \left(\frac{\lambda}{\Delta_5}\right)^{a} \right)\right). \tag{24}$$

Furthermore, if $\zeta \geq 1/2$, *then the above conclusions still hold if we replace* $\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho$ *by* $\|\mathcal{T}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \omega_H)\|_H$.

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* λ *be given by* (22) *and* $P = I$. *Then with probability at least* $1-\delta$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \log^2 \frac{3}{\delta} \begin{cases} \left(\frac{1\vee\log n^{\gamma}}{n}\right)^{\zeta-a}, & \text{if } 2\zeta+\gamma \leq 1, \\ n^{-\frac{\zeta-a}{2\zeta+\gamma}}, & \text{if } 2\zeta+\gamma > 1. \end{cases} \tag{25}$$

Furthermore, if $\zeta \geq 1/2$,

$$\|\mathcal{T}^{1/2-a}(\omega_\lambda^{\mathbf{z}} - \omega_H)\|_H \lesssim \log^2 \frac{3}{\delta} n^{-\frac{\zeta-a}{2\zeta+\gamma}},$$

The rate from the above with $2\zeta+\gamma \leq 1$ improves the rate $O(n^{a-\zeta}[1 \vee \log n^{\gamma}]^{1-a})$ 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.*

Articles	Assumptions	Minimax Rate
(Caponnetto and De Vito, 2007, Theorem 2)	$a = 0, \zeta \in [\frac{1}{2}, 1]$	$N^{-\frac{2\zeta}{2\zeta+\gamma}}$
(Steinwart et al., 2009, Theorem 9)	$a = 0, \zeta \in (0, \frac{1}{2}]$	$N^{-\frac{2\zeta}{2\zeta+\gamma}}$
(Blanchard and Mücke, 2018, Theorem 3.5)	$a \in [0, \frac{1}{2}], \zeta \geq \frac{1}{2}$	$N^{-\frac{2(\zeta-a)}{2\zeta+\gamma}}$

Table 1: Minimax Rates on $\|\mathcal{L}^{-a}(f_{\mathbf{z}} - f_H)\|_\rho^2$

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 $\overline{\mathrm{range}\{\mathcal{S}_{\mathbf{x}}^*\mathbf{G}^*\}}$, where $\mathbf{G} \in \mathbb{R}^{m\times n}$ is a sketch matrix with $m < n$ satisfying the following concentration inequality: For any finite subset E in $\mathbb{R}^n$ and for any $t \in (0, 1)$,

$$\mathbb{P}(|\|\mathbf{Ga}\|_2^2 - \|\mathbf{a}\|_2^2 \geq t\|\mathbf{a}\|_2^2 : \exists \mathbf{a} \in E) \leq 2|E|\mathrm{e}^{\frac{-t^2 m}{c_0' \log^\beta n}}. \tag{26}$$

Here, c_0' is a universal positive constant and β 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 $\mathbf{G}$ are independent (scaled) copies of an isotropic ψ_2 vector, then $\mathbf{G}$ 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 $S = \overline{range\{\mathcal{S}_{\mathbf{x}}^*\mathbf{G}^*\}}$, where $\mathbf{G} \in \mathbb{R}^{m\times n}$ is a randomized matrix satisfying* (26)*. Let*

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

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

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \log^3 \frac{3}{\delta} \begin{cases} \left(\frac{1\vee \log n^\gamma}{n}\right)^{\zeta-a}, & \text{if } 2\zeta+\gamma \leq 1, \\ n^{-\frac{\zeta-a}{2\zeta+\gamma}}, & \text{if } 2\zeta+\gamma > 1. \end{cases} \tag{28}$$

Furthermore, if $\zeta \geq 1/2$,

$$\|\mathcal{T}^{1/2-a}(\omega_\lambda^{\mathbf{z}} - \omega_H)\|_H \lesssim \log^3 \frac{3}{\delta} n^{-\frac{\zeta-a}{2\zeta+\gamma}}.$$

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 ζ is large or the effective-dimensional parameter γ 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 $\overline{\mathrm{range}\{\mathcal{S}_{\mathbf{x}}^* \mathbf{G}^*\}}$, where $\mathbf{G} \in \mathbb{R}^{m \times n}$ 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* $S = \overline{span\{x_1, \cdots, x_m\}}$, $2\zeta + \gamma > 1$, *and* $\lambda = n^{-\frac{1}{2\zeta+\gamma}}$. *If*

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

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 $O(n^{\frac{2\zeta-1}{2\zeta+1}})$ 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 $\mathbf{K} = \mathcal{S}_{\mathbf{x}}\mathcal{S}_{\mathbf{x}}^*$. For $\lambda > 0$, the leveraging scores of $\mathbf{K}(\mathbf{K} + \lambda I)$ is the set $\{l_i(\lambda)\}_{i=1}^n$ with

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

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

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

for some $L \ge 1$. In an (L, λ)-ALS Nyström subsampling regime, $S = \overline{\mathrm{range}\{\tilde{x}_1, \cdots, \tilde{x}_m\}}$, where each $\tilde{x}_j$ is i.i.d. drawn according to

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

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 $\lambda = n^{-\frac{1}{(2\zeta+\gamma)\vee 1}}$ and $S = \overline{range\{\tilde{x}_1, \cdots, \tilde{x}_m\}}$ with $\tilde{x}_j$ drawn following an (L, λ)-ALS Nyström subsampling scheme. If*

$$m \gtrsim L^2 \log^3 \frac{3}{\delta} \begin{cases} n^{\gamma}[1 \vee \log n^{\gamma}]^{1-\gamma}, & \text{if } 2\zeta + \gamma \le 1, \\ n^{\frac{\gamma(\zeta-a)}{(1-a)(2\zeta+\gamma)}}[1 \vee \log n^{\gamma}], & \text{if } \zeta \ge 1, \\ n^{\frac{\gamma}{2\zeta+\gamma}}[1 \vee \log n^{\gamma}], & \text{otherwise}, \end{cases} \tag{29}$$

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* $\omega_1 = 0$,

$$\omega_{t+1} = \omega_t - \eta \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} (\langle \omega_t, x_{j_i} \rangle_H - y_{j_i}) P x_{j_i}, \quad t = 1, \cdots, T,$$

where η *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 η_t, 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 $\mathbf{z}$. We write $\mathbf{J} = \{j_1, \cdots, j_{bT}\}$ and denote the conditional expectation with respect to $\mathbf{J}$ given $\mathbf{z}$ by $\mathbb{E}_{\mathbf{J}}$.

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*

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

ρ_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* $C_1' \geq 1$,

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \leq C_1' \lambda^{\frac{1 \vee \zeta - a}{1-a}} \log \frac{2}{\delta}, \quad \lambda = n^{-\frac{1}{1 \vee (2\zeta+\gamma)}} (1 \vee \log n^{\gamma})^{\mathbf{1}_{\{2\zeta+\gamma \leq 1\}}}. \tag{31}$$

Consider Algorithm 2 with any of the following choices on η, b *and* T*:*
I) $\eta \simeq \lambda^{2\zeta}$, $b = 1$ *and* $T \simeq \lambda^{-(1+2\zeta)}$*;*
II) $\eta \simeq (\log n)^{-1}$, $b \simeq \lambda^{-2\zeta}$ *and* $T \simeq \lambda^{-1} \log n$*;*
III) $\eta \simeq n^{-1}$, $b = 1$ *and* $T \simeq n\lambda^{-1}$*;*
IV) $\eta \simeq n^{-1/2}$, $b \simeq \sqrt{n}$ *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 $2\zeta + \gamma \leq 1$,

$$\mathbb{E}_{\mathbf{J}} \|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_{T+1} - f_H)\|_\rho^2 \lesssim n^{-2(\zeta-a)} (\log n)^{\mathbf{1}_{\{2a \neq 1\}}} (1 \vee \log n^{\gamma})^{2(\zeta-a)} \log^3 \frac{2}{\delta}. \tag{32}$$

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

$$\mathbb{E}_{\mathbf{J}} \|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_{T+1} - f_H)\|_\rho^2 \lesssim n^{-\frac{2(\zeta-a)}{2\zeta+\gamma}} (\log n)^{\mathbf{1}_{\{2a \neq 1\}}} \log^3 \frac{2}{\delta}. \tag{33}$$

Furthermore, if $\zeta \geq 1/2$*, then the above conclusions still hold if we replace* $\|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_{T+1} - f_H)\|_\rho$ *by* $\|\mathcal{T}^{\frac{1}{2}-a}(\omega_{T+1} - \omega_H)\|_H$.

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 $O(n^{\frac{-(2\zeta-a)}{2\zeta+\gamma}} \log^2 n)$ 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 $2\zeta + \gamma \le 1$ 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*

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

Lemma 9 *Let H_1, H_2 be two separable Hilbert spaces and $\mathcal{S} : H_1 \to H_2$ a compact operator. Then for any function $f : [0, \|\mathcal{S}\|] \to [0, \infty[$,*

$$f(\mathcal{S}\mathcal{S}^*)\mathcal{S} = \mathcal{S} f(\mathcal{S}^*\mathcal{S}).$$

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 κ^2. Then for any $\zeta > 0$,*

$$\|A^\zeta - B^\zeta\| \leq C_{\zeta,\kappa} \|A - B\|^{\zeta \wedge 1}, \tag{34}$$

where

$$C_{\zeta,\kappa} = \begin{cases} 1 & \textit{when } \zeta \leq 1, \\ 2\zeta\kappa^{2\zeta-2} & \textit{when } \zeta > 1. \end{cases} \tag{35}$$

Proof The proof is based on the fact that u^ζ 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,*

$$\|(A + \lambda I)^{\frac{1}{2}} X F^*\| = \|F X^* (A + \lambda I)^{\frac{1}{2}}\| \leq \|F(X^* A X + \lambda I)^{\frac{1}{2}}\|. \tag{36}$$

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

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

It thus follows that

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

Therefore,

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

■

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

$$\|A^s(I-P)A^t\| \leq \|A^{\frac{1}{2}}(I-P)\|^{2s}\|(I-P)A^{\frac{1}{2}}\|^{2t} = \|(I-P)A^{\frac{1}{2}}\|^{2(t+s)}.$$

Using $\|D\|^2 = \|D^*D\|$,

$$\|A^s(I-P)A^t\| \leq \|(I-P)A(I-P)\|^{t+s}.$$

Adding and subtracting with the same term, using the triangle inequality, and noting that $\|I-P\| \leq 1$ and $s+t \leq 1$,

$$\begin{aligned}\|A^s(I-P)A^t\| \leq & \|(I-P)A(I-P)\|^{t+s} \\ \leq & (\|(I-P)(A-B)(I-P)\| + \|(I-P)B(I-P)\|)^{t+s} \\ \leq & \|A-B\|^{s+t} + \|(I-P)B(I-P)\|^{s+t},\end{aligned}$$

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

$$\mathcal{T}_\lambda = \mathcal{T} + \lambda I, \quad \mathcal{T}_{\mathbf{x}\lambda} = \mathcal{T}_{\mathbf{x}} + \lambda I.$$

We also denote

$$\mathcal{R}_\lambda(u) = 1 - \mathcal{G}_\lambda(u)u. \tag{37}$$

For any $\lambda > 0$, we introduce a deterministic vector ω_H^λ, defined by

$$\omega_H^\lambda = \bar{\mathcal{G}}_\lambda(\mathcal{T})\mathcal{S}_\rho^* f_H, \tag{38}$$

where $\bar{\mathcal{G}}_\lambda(u)$ is given by (19). We have the following lemma for the properties of ω_H^λ. We assume $\tau \geq \zeta - a$ throughout.

Lemma 13 *Under Assumption 2, the following holds.*
1) For any $a \leq \zeta$, we have

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_H^\lambda - f_H)\|_\rho \leq R\lambda^{\zeta-a}. \tag{39}$$

2) We have

$$\|\mathcal{T}^{a-1/2}\omega_H^\lambda\|_H \leq R \cdot \begin{cases}\lambda^{\zeta+a-1}, & \textit{if } -\zeta \leq a \leq 1-\zeta, \\ \kappa^{2(\zeta+a-1)}, & \textit{if } a \geq 1-\zeta.\end{cases} \tag{40}$$

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*

$$1 \vee \|\mathcal{T}_\lambda^{\frac{1}{2}} \mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\|^2 \vee \|\mathcal{T}_\lambda^{-\frac{1}{2}} \mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\|^2 \le \Delta_1,$$

$$\|\mathcal{T}_\lambda^{-1/2}[(\mathcal{T}_{\mathbf{x}}\omega_H^\lambda - \mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}}) - (\mathcal{T}\omega_H^\lambda - \mathcal{S}_\rho^* f_H)]\|_H \le \Delta_2,$$

$$\|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\| \le \Delta_3,$$

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 = \Delta_5.$$

Then, for any $0 \le a \le [\zeta \wedge \frac{1}{2}]$, the following holds.
1) If $\zeta \in [0,1]$, then we have

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \le \lambda^{-a}\Delta_1^{1-a}\Big(E\Delta_2 + (2E+F+1)R\lambda^\zeta + R\lambda^{\zeta-1}(E\Delta_5 + \Delta_5^{1-a}\lambda^a)\Big). \tag{41}$$

2) If $\zeta \ge 1$, then we have

$$\begin{aligned}\|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \le \lambda^{-a}\Delta_1^{1-a}\Big(&E\Delta_2 + (E+F+1)R\lambda^\zeta + \kappa^{2(\zeta-1)}R\big(E\Delta_3 + E\Delta_5 + \Delta_5^{1-a}\lambda^a\big) \\ &+ C_{\zeta-\frac{1}{2},\kappa}FR\big(\lambda(\Delta_3+\Delta_5)^{(\zeta-1)\wedge 1} + \lambda^{\frac{1}{2}}\Delta_3^{(\zeta-\frac{1}{2})\wedge 1}\big)\Big).\end{aligned} \tag{42}$$

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 $\zeta \ge 1$ and $\zeta \le 1/2$, 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

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

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

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

Using the spectral theorem for compact operators, $\mathcal{L} = \mathcal{S}_\rho\mathcal{S}_\rho^*$, and $\mathcal{T} = \mathcal{S}_\rho^*\mathcal{S}_\rho$, we have

$$\|\mathcal{L}^{-a}\mathcal{S}_\rho\mathcal{T}^{a-\frac{1}{2}}\| \le 1,$$

and thus

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \le \|\mathcal{T}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \omega_H^\lambda)\|_H + R\lambda^{\zeta-a}. \tag{43}$$

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

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \le \|\mathcal{T}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H + \|\mathcal{T}^{\frac{1}{2}-a}(I-P)\omega_H^\lambda\|_H + R\lambda^{\zeta-a}.$$

Since P is an orthogonal projected operator and $a \in [0, \frac{1}{2}]$, we have

$$\begin{aligned}
&\|\mathcal{T}^{\frac{1}{2}-a}(I-P)\omega_H^\lambda\|_H \\
=&\|\mathcal{T}^{\frac{1}{2}(1-2a)}(I-P)^{1-2a}(I-P)\omega_H^\lambda\|_H \\
\le&\|\mathcal{T}^{\frac{1}{2}(1-2a)}(I-P)^{1-2a}\|\|(I-P)\mathcal{T}^{\frac{1}{2}}\|\|\mathcal{T}^{-\frac{1}{2}}\omega_H^\lambda\|_H \\
\le&\|\mathcal{T}^{\frac{1}{2}}(I-P)\|^{1-2a}\|(I-P)\mathcal{T}^{\frac{1}{2}}\|R\kappa^{2(\zeta-1)_+}\lambda^{(\zeta-1)_-} \\
=&\Delta_5^{1-a}R\kappa^{2(\zeta-1)_+}\lambda^{(\zeta-1)_-},
\end{aligned}$$

where for the last second inequality, we use Lemma 8 and Part 2) of Lemma 13, and we subsequently obtain

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \le \|\mathcal{T}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H + R\kappa^{2(\zeta-1)_+}\lambda^{(\zeta-1)_-}\Delta_5^{1-a} + R\lambda^{\zeta-a}.$$

Since for all $\omega \in H$, and $a \in [0, \frac{1}{2}]$,

$$\begin{aligned}
\|\mathcal{T}^{\frac{1}{2}-a}\omega\|_H \le&\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{T}_{\mathbf{x}\lambda}^{a-\frac{1}{2}}\|\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}-a}\omega\|_H \\
\le&\lambda^{-a}\|\mathcal{T}_\lambda^{\frac{1}{2}-a}\mathcal{T}_{\mathbf{x}\lambda}^{a-\frac{1}{2}}\|\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\omega\|_H \\
\le&\lambda^{-a}\|\mathcal{T}_\lambda^{\frac{1}{2}}\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\|^{1-2a}\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\omega\|_H \\
\le&\lambda^{-a}\Delta_1^{\frac{1}{2}-a}\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\omega\|_H
\end{aligned} \tag{44}$$

(where we use Lemma 8 for the last second inequality), we get

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho \omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \le \lambda^{-a}\Delta_1^{\frac{1}{2}-a}\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H + R\kappa^{2(\zeta-1)_+}\lambda^{(\zeta-1)_-}\Delta_5^{1-a} + R\lambda^{\zeta-a}. \tag{45}$$

In what follows, we estimate $\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H$.

Introducing with (15), with $P^2 = P$,

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H = \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}P(\mathcal{G}_\lambda(P\mathcal{T}_{\mathbf{x}}P)P\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}} - P\omega_H^\lambda)\|_H.$$

Since for any $\omega \in H$,

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}P\omega\|_H^2 = \langle P\mathcal{T}_{\mathbf{x}\lambda}P\omega, \omega\rangle_H \le \langle (P\mathcal{T}_{\mathbf{x}}P + \lambda I)\omega, \omega\rangle_H = \|(P\mathcal{T}_{\mathbf{x}}P + \lambda I)^{\frac{1}{2}}\omega\|_H^2,$$

and we thus get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H \le \|\mathcal{U}_\lambda^{\frac{1}{2}}(\mathcal{G}_\lambda(\mathcal{U})P\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}} - P\omega_H^\lambda)\|_H,$$

where we denote

$$\mathcal{U} = P\mathcal{T}_{\mathbf{x}}P, \quad \mathcal{U}_\lambda = \mathcal{U} + \lambda I. \tag{46}$$

Subtracting and adding with the same term, and applying the triangle inequality, with the notation $\mathcal{R}_\lambda$ given by (37) and $P^2 = P$, we have

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H \le \|\underbrace{\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{G}_\lambda(\mathcal{U})P(\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}} - \mathcal{T}_{\mathbf{x}}P\omega_H^\lambda)}_{\mathbf{Term.A}}\|_H + \|\underbrace{\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\omega_H^\lambda}_{\mathbf{Term.B}}\|_H. \tag{47}$$

We will estimate the above two terms of the right-hand side.
Estimating $\|$Term.A$\|_H$:
Using Lemma 11,

$$\|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{G}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\| \le \|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{G}_\lambda(\mathcal{U})\mathcal{U}_\lambda^{\frac{1}{2}}\| = \|\mathcal{U}_\lambda\mathcal{G}_\lambda(\mathcal{U})\|.$$

Using the spectral theorem, with $\|\mathcal{U}\| \le \|\mathcal{T}_{\mathbf{x}}\| \le \kappa^2$ (implied by (6)), and then applying (16),

$$\|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{G}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\| \le \sup_{u\in[0,\kappa^2]}|(u+\lambda)\mathcal{G}_\lambda(u)| \le E. \tag{48}$$

Using the above inequality, and by a simple calculation,

$$\|\mathbf{Term.A}\|_H \le \|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{G}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\|\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}(\mathcal{S}_{\mathbf{x}}^*\mathbf{y} - \mathcal{T}_{\mathbf{x}}P\omega_H^\lambda)\| \le E\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}(\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}} - \mathcal{T}_{\mathbf{x}}P\omega_H^\lambda)\|.$$

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

$$\begin{aligned}
\|\mathbf{Term.A}\|_H \le& E\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}(\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}} - \mathcal{T}_{\mathbf{x}}\omega_H^\lambda)\|_H + E\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\mathcal{T}_{\mathbf{x}}(I-P)\omega_H^\lambda)\|_H \\
\le& E\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\mathcal{T}_\lambda^{\frac{1}{2}}\|\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}} - \mathcal{T}_{\mathbf{x}}\omega_H^\lambda)\|_H + E\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\mathcal{T}_{\mathbf{x}}(I-P)\omega_H^\lambda)\|_H \\
\le& E\Delta_1^{\frac{1}{2}}\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}} - \mathcal{T}_{\mathbf{x}}\omega_H^\lambda)\|_H + E\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\mathcal{T}_{\mathbf{x}}(I-P)\omega_H^\lambda)\|_H \\
\le& E\Delta_1^{\frac{1}{2}}(\Delta_2 + \|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}\omega_H^\lambda - \mathcal{S}_\rho^* f_H)\|_H) + E\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\mathcal{T}_{\mathbf{x}}(I-P)\omega_H^\lambda\|_H \\
\le& E\Delta_1^{\frac{1}{2}}(\Delta_2 + \|\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{S}_\rho^*\|\|\mathcal{S}_\rho\omega_H^\lambda - f_H\|_\rho) + E\|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(I-P)\|\|(I-P)\mathcal{T}^{\frac{1}{2}}\|\|\mathcal{T}^{-\frac{1}{2}}\omega_H^\lambda\|_H,
\end{aligned}$$

where we used $\mathcal{T} = \mathcal{S}_\rho^*\mathcal{S}_\rho$ and $(I-P)^2 = I - P$ for the last inequality. Applying Lemma 13 and $\|\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{S}_\rho^*\| \le 1$,

$$\|\mathbf{Term.A}\|_H \le E\Delta_1^{\frac{1}{2}}(\Delta_2 + R\lambda^\zeta) + ER\Delta_5^{\frac{1}{2}}\|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(I-P)\|\kappa^{2(\zeta-1)_+}\lambda^{(\zeta-1)_-}. \tag{49}$$

In what follows, we estimate $\|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(I-P)\|$, considering two different cases.
Case $\zeta \le 1$.
We have

$$\|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(I-P)\| \le \Delta_1^{\frac{1}{2}}\|\mathcal{T}_\lambda^{\frac{1}{2}}(I-P)\|.$$

Note that for any $\omega \in H$ with $\|\omega\|_H = 1$,

$$\|\mathcal{T}_\lambda^{\frac{1}{2}}(I-P)\omega\|_H^2 = \langle \mathcal{T}_\lambda (I-P)\omega, (I-P)\omega\rangle_H = \|\mathcal{T}^{\frac{1}{2}}(I-P)\omega\|_H^2 + \lambda\|(I-P)\omega\|_H^2$$
$$\leq \|\mathcal{T}^{\frac{1}{2}}(I-P)\|^2 + \lambda \leq \Delta_5 + \lambda.$$

It thus follows that

$$\|\mathcal{T}_\lambda^{\frac{1}{2}}(I-P)\| \leq (\Delta_5+\lambda)^{\frac{1}{2}}, \tag{50}$$

and thus

$$\|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(I-P)\| \leq \Delta_1^{\frac{1}{2}}(\Delta_5+\lambda)^{\frac{1}{2}}.$$

Introducing the above into (49), we know that **Term.A** can be estimated as ($\zeta \leq 1$)

$$\|\mathbf{Term.A}\|_H \leq E\Delta_1^{\frac{1}{2}}\left(\Delta_2 + 2R\lambda^\zeta + R\lambda^{\zeta-1}\Delta_5\right). \tag{51}$$

Case $\zeta \geq 1$.
Applying Lemma 12, we obtain

$$\|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(I-P)\|^2 = \|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\| \leq \Delta_3 + \|\mathcal{T}^{\frac{1}{2}}(I-P)\mathcal{T}^{\frac{1}{2}}\| = \Delta_3 + \Delta_5.$$

Introducing the above into (49), we get for $\zeta \geq 1$,

$$\|\mathbf{Term.A}\|_H \leq E\Delta_1^{\frac{1}{2}}\left(\Delta_2 + R\lambda^\zeta + (\Delta_3+\Delta_5)\,\kappa^{2(\zeta-1)}R\right). \tag{52}$$

Estimating $\|\mathbf{Term.B}\|_H$:
We estimate $\|\mathbf{Term.B}\|_H$, considering two different cases.
Case I: $\zeta \leq 1$.
Using a same argument as that for (48) and (17),

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

Using the above inequality and by a direct calculation,

$$\|\mathbf{Term.B}\|_H \leq \|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}\|\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{-1}{2}}\mathcal{T}_\lambda^{\frac{1}{2}}\|\|\mathcal{T}^{-\frac{1}{2}}\omega_H^\lambda\|_H \leq F\lambda\Delta_1^{\frac{1}{2}}\|\mathcal{T}^{-\frac{1}{2}}\omega_H^\lambda\|_H.$$

Applying Part 2) of Lemma 13, we get

$$\|\mathbf{Term.B}\|_H \leq FR\lambda^\zeta\Delta_1^{\frac{1}{2}}. \tag{53}$$

Applying the above and (51) into (47), we know that for any $\zeta \in [0,1]$,

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H \leq \Delta_1^{\frac{1}{2}}\left(E\Delta_2 + (2E+F)R\lambda^\zeta + ER\Delta_5\lambda^{\zeta-1}\right).$$

Using the above into (45), we can prove the first desired result.
Case II: $\zeta \geq 1$
We denote

$$\mathcal{V} = \mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}, \quad \mathcal{V}_\lambda = \mathcal{V} + \lambda I. \tag{54}$$

Noting that $\mathcal{U} = P\mathcal{T}_{\mathbf{x}}P = P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}})^*$, thus following from Lemma 9 (with $f(u) = (u + \lambda)^{\frac{1}{2}}\mathcal{R}_\lambda(u)$) and $P^2 = P$,

$$\|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| = \|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})(P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}})\mathcal{T}_{\mathbf{x}}^{\zeta-1}\| = \|(P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}})\mathcal{V}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{V})\mathcal{T}_{\mathbf{x}}^{\zeta-1}\|.$$

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

$$\begin{aligned}\|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| \leq & \|P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\mathcal{V}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{V})\mathcal{V}^{\zeta-1}\| + \|P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\mathcal{V}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{V})(\mathcal{T}_{\mathbf{x}}^{\zeta-1} - \mathcal{V}^{\zeta-1})\| \\ \leq & \|P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\mathcal{V}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{V})\mathcal{V}^{\zeta-1}\| + \|P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\mathcal{V}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{V})\|\|\mathcal{T}_{\mathbf{x}}^{\zeta-1} - \mathcal{V}^{\zeta-1}\|.\end{aligned}$$

Using Lemma 10, with (6) and $\|\mathcal{V}\| \leq \|\mathcal{T}_{\mathbf{x}}\| \leq \kappa^2$, we then have

$$\|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| \leq \|P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\mathcal{V}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{V})\mathcal{V}^{\zeta-1}\| + \|P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\mathcal{V}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{V})\|C_{\zeta-1,\kappa}\|\mathcal{T}_{\mathbf{x}} - \mathcal{V}\|^{(\zeta-1)\wedge 1}.$$

Using $\|A\| = \|A^*A\|^{\frac{1}{2}}, P^2 = P$, the spectral theorem, and (17), for any $s \in [1, \tau]$, it holds that

$$\begin{aligned}\|P\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\mathcal{V}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{V})\mathcal{V}^{s-1}\| = & \|\mathcal{V}^{s-1}\mathcal{R}_\lambda(\mathcal{V})\mathcal{V}_\lambda\mathcal{V}\mathcal{R}_\lambda(\mathcal{V})\mathcal{V}^{s-1}\|^{\frac{1}{2}} \\ \leq & \sup_{u\in[0,\kappa^2]} |\mathcal{R}_\lambda(u)u^{s-\frac{1}{2}}(u+\lambda)^{\frac{1}{2}}| \leq F\lambda^s,\end{aligned}$$

and thus we get

$$\|\mathcal{U}_\lambda^{\frac{1}{2}-a}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| \leq F(\lambda^\zeta + \lambda C_{\zeta-1,\kappa}\|\mathcal{T}_{\mathbf{x}} - \mathcal{V}\|^{(\zeta-1)\wedge 1}).$$

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

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

and we thus get

$$\|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| \leq F(\lambda^\zeta + \lambda C_{\zeta-1,\kappa}(\Delta_3 + \Delta_5)^{(\zeta-1)\wedge 1}). \tag{55}$$

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

$$\|\mathbf{Term}.\mathbf{B}\|_H \leq \|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}^{\zeta-\frac{1}{2}}\|\|\mathcal{T}^{\frac{1}{2}-\zeta}\omega_H^\lambda\|_H \leq \|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}^{\zeta-\frac{1}{2}}\|R.$$

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

$$\|\mathbf{Term}.\mathbf{B}\|_H \leq R\left(\|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})P\mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| + \|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})\|\|\mathcal{T}^{\zeta-\frac{1}{2}} - \mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\|\right).$$

Using the spectral theorem, with $\|\mathcal{U}\| \leq \|\mathcal{T}_{\mathbf{x}}\| \leq \kappa^2$ by (6) and (17),

$$\|\mathcal{U}_\lambda^{\frac{1}{2}}\mathcal{R}_\lambda(\mathcal{U})\| = \sup_{u\in]0,\kappa^2]} |\mathcal{R}_\lambda(u)(u+\lambda)^{\frac{1}{2}}| \leq F\lambda^{\frac{1}{2}},$$

and we thus get

$$\|\mathbf{Term.B}\|_H \leq R\left(\|\mathcal{U}_\lambda^{\frac{1}{2}} \mathcal{R}_\lambda(\mathcal{U}) P \mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| + F\lambda^{\frac{1}{2}} \|\mathcal{T}^{\zeta-\frac{1}{2}} - \mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| \right).$$

Applying Lemma 10, with (3) and (6), it follows that

$$\|\mathbf{Term.B}\|_H \leq R\left(\|\mathcal{U}_\lambda^{\frac{1}{2}} \mathcal{R}_\lambda(\mathcal{U}) P \mathcal{T}_{\mathbf{x}}^{\zeta-\frac{1}{2}}\| + F\lambda^{\frac{1}{2}} C_{\zeta-\frac{1}{2},\kappa} \Delta_3^{(\zeta-\frac{1}{2})\wedge 1} \right).$$

Introducing with (55), we obtain

$$\|\mathbf{Term.B}\|_H \leq FR\left(\lambda^\zeta + C_{\zeta-1,\kappa}\lambda(\Delta_3+\Delta_5)^{(\zeta-1)\wedge 1} + C_{\zeta-\frac{1}{2},\kappa}\lambda^{\frac{1}{2}}\Delta_3^{(\zeta-\frac{1}{2})\wedge 1} \right).$$

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

$$\begin{aligned} \|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_\lambda^{\mathbf{z}} - P\omega_H^\lambda)\|_H \leq & \Delta_1^{\frac{1}{2}}\Big(E\Delta_2 + (F+E)R\lambda^\zeta + E\kappa^{2(\zeta-1)}R(\Delta_3+\Delta_5) \\ & + C_{\zeta-\frac{1}{2},\kappa}FR\big(\lambda(\Delta_3+\Delta_5)^{(\zeta-1)\wedge 1} + \lambda^{\frac{1}{2}}\Delta_3^{(\zeta-\frac{1}{2})\wedge 1}\big)\Big). \end{aligned}$$

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 Δ_1, Δ_2, and Δ_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 $\lambda = [1 \vee \log n^\gamma]/n$. Then with probability at least $1-\delta$,*

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

where $a(\delta) = 8\kappa^2 \log \frac{4\kappa^2 \mathrm{e}(c_\gamma+1)}{\delta\|\mathcal{T}\|}$ if $\lambda = [1 \vee \log n^\gamma]/n$, or $a(\delta) = 8\kappa^2 \left(\log \frac{4\kappa^2(c_\gamma+1)}{\delta\|\mathcal{T}\|} + \frac{\theta\gamma}{\mathrm{e}(1-\theta)} \right)$ 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 $\lambda = [1 \vee \log n^\gamma]/n$ using the same argument. We report the proof in Appendix.

Lemma 15 *Let $0 < \delta < 1/2$. The following holds with probability at least $1-\delta$:*

$$\|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\| \leq \|\mathcal{T} - \mathcal{T}_{\mathbf{x}}\|_{HS} \leq \frac{2\kappa^2 \log(2/\delta)}{n} + \sqrt{\frac{2\kappa^4\log(2/\delta)}{n}}.$$

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:*

$$\begin{aligned}&\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}}\omega_H^\lambda - \mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}} - \mathcal{T}\omega_H^\lambda + \mathcal{S}_\rho^* f_H)\|_H \\ &\leq 2\left(\frac{4\kappa(M+\kappa^{1\vee(2\zeta)}R\lambda^{(\zeta-\frac{1}{2})_-})}{n\sqrt{\lambda}} + \sqrt{\frac{8(2R^2\kappa^2\lambda^{2\zeta-1}+(2B^2+Q^2)\mathcal{N}(\lambda))}{n}}\right)\log\frac{2}{\delta}. \qquad (56)\end{aligned}$$

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 $\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 = \Delta_5$, and $\lambda = n^{-\theta}$ for some $\theta \in [0,1)$ or $\lambda = \frac{1\vee\log n^\gamma}{n}$. Then, for any $0 \leq a \leq [\zeta\wedge\frac{1}{2}]$, with probability at least $1-3\delta$ ($\delta\in(0,1/3)$), the following statements hold.*
1) If $\zeta\in[0,1]$, we have

$$\begin{aligned}\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda^{\mathbf{z}} - f_H)\|_\rho &\leq \bar{C}_1^{1-a}\log^{1-a}\frac{2}{\delta}\lambda^{\zeta-1-a}(E\Delta_5+\Delta_5^{1-a}\lambda^a)R \\ &+ \bar{C}_1^{1-a}\log^{2-a}\frac{2}{\delta}\lambda^{-a}\Big(\bar{C}_2(\lambda^\zeta\vee\frac{1}{n\sqrt{\lambda}})R + 8E\sqrt{\frac{\mathcal{N}(\lambda)}{n}}(B+Q) + 8\kappa E\frac{M}{n\sqrt{\lambda}}\Big).\end{aligned}$$

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

$$\begin{aligned}\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda^{\mathbf{z}} - f_H)\|_\rho &\leq \bar{C}_1^{1-a}C_{\zeta,\kappa}\log^{1-a}\frac{2}{\delta}\lambda^{-a}(E\Delta_5+\Delta_5^{1-a}\lambda^a + F\lambda\Delta_5^{(\zeta-1)\wedge 1})R \\ &+ \bar{C}_1^{1-a}\log^{2-a}\frac{2}{\delta}\lambda^{-a}\Big(\bar{C}_3(\lambda^\zeta\vee\sqrt{\frac{1}{n}})R + 8E\sqrt{\frac{\mathcal{N}(\lambda)}{n}}(B+Q) + 8\kappa E\frac{M}{n\sqrt{\lambda}}\Big).\end{aligned}$$

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

$$\bar{C}_1 = \begin{cases} 24\kappa^2\left(\log\frac{2\kappa^2\mathrm{e}(c_\gamma+1)}{\|\mathcal{T}\|}+1\right), & \text{if } \lambda = \frac{1\vee\log n^\gamma}{n}, \\ 24\kappa^2\left(\log\frac{2\kappa^2\mathrm{e}(c_\gamma+1)}{\|\mathcal{T}\|}+\frac{\theta\gamma}{\mathrm{e}(1-\theta)}\right), & \text{otherwise}, \end{cases}$$

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

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

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

Proof We use Proposition 3 to prove the statement. We thus need to estimate Δ_1, Δ_2 and Δ_3. Following from Lemmas 14, 15 and 16, with $n^{-1} \le \lambda \le 1$, we know that with probability at least $1 - 3\delta$,

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

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

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

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 $1/n \le \lambda \le 1$.

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

$$\|\mathcal{T}^{1/2-a}(\omega_\lambda^{\mathbf{z}} - \omega_H)\|_H \le \|\mathcal{T}^{\frac{1}{2}-a}(\omega_\lambda^{\mathbf{z}} - \omega_H^\lambda)\|_H + R\lambda^{\zeta - a}.$$

■

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, $\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2$. The estimate for $\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2$ 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 $\mathbf{x} \subseteq H^n$, assume that for $\lambda \in [0,1]$,*

$$\operatorname{tr}((\mathcal{T}_{\mathbf{x}} + \lambda I)^{-1}\mathcal{T}_{\mathbf{x}}) \le b_\gamma \lambda^{-\gamma} \tag{57}$$

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

$$\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 \le 6\lambda,$$

provided that

$$m \ge 100 c_0' \log^\beta n \lambda^{-\gamma} \log \frac{3}{\delta} \left(1 + 10 b_\gamma\right). \tag{58}$$

Under the condition (57), Lemma 17 provides an upper bound for $\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|$. 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 $\lambda = \frac{1 \vee \log n^{\gamma}}{n}$, with probability at least $1-\delta$, the following holds:*

$$\operatorname{tr}((\mathcal{T}_{\mathbf{x}} + \lambda I)^{-1} \mathcal{T}_{\mathbf{x}}) \leq b_{\gamma} \log^2 \frac{4}{\delta} \lambda^{-\gamma}. \tag{59}$$

Here, b_γ is a positive constant given by

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

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 $S = \overline{range\{\mathcal{S}_{\mathbf{x}}^* \mathbf{G}^*\}}$, where $\mathbf{G} \in \mathbb{R}^{m \times n}$ 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*

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

provided that

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

Here, $a_{\gamma} = 24\kappa^2 \log \frac{\kappa^2 \mathrm{e}^2 (c_{\gamma}+1)}{\|\mathcal{T}\|}$, and $\bar{C} = 100 c_0' (1 + 10 b_{\gamma})$ with

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

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

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

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

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

where we use the following fact

$$\frac{\log n^{\gamma}}{n^{1-\theta}} = \frac{\gamma}{1-\theta}\frac{\log n^{1-\theta}}{n^{1-\theta}} \le \frac{\gamma}{1-\theta}, \quad \text{if } \theta < 1,$$

within the last inequality. Combining with Corollary 1, and by a direct calculation, with $\lambda \le 1$, 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:*

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \lesssim m^{-\frac{1}{\gamma}}.$$

Introducing this estimate into Theorem 1, we observe that the following conclusions hold with high probability for $\lambda \in (n^{-1},1]$ and $a \in [0, \frac{1}{2} \wedge \zeta]$:
For $\zeta \in [0,1]$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \lambda^{\zeta-a} + \frac{1}{n^{\frac{1}{2}}\lambda^{a+\frac{\gamma}{2}}} + \frac{\lambda^{\zeta-a}}{m^{\frac{1}{\gamma}}\lambda},$$

while for $\zeta \ge 1$ and $\lambda \ge n^{-1/2}$, we have

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \lambda^{\zeta-a} + \frac{1}{n^{\frac{1}{2}}\lambda^{a+\frac{\gamma}{2}}} + \frac{1}{m^{\frac{1}{\gamma}}\lambda^{a}} + \frac{1}{m^{\frac{1-a}{\gamma}}} + \frac{\lambda^{1-a}}{m^{\frac{(\zeta-1)\wedge 1}{\gamma}}}.$$

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*

$$S = \overline{span\{x_1, \cdots, x_m\}}.$$

Then with probability at least $1-\delta$ ($\delta \in (0,1)$), the following inequality holds:

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \le \|(I-P)\mathcal{T}_\mu^{\frac{1}{2}}\|^2 \le \frac{1 \vee \log m^{\gamma}}{m} 24\kappa^2 \log \frac{4\kappa^2 \mathrm{e}(c_\gamma+1)}{\delta\|\mathcal{T}\|}, \tag{61}$$

where $\mu = \frac{1 \vee \log m^{\gamma}}{m}$.

The next lemma provides upper bounds for projection errors with ALS Nyström subsampling scheme.

Lemma 21 *Under Assumption 3, let $S = \overline{range\{\tilde{x}_1, \cdots, \tilde{x}_m\}}$, with each $\tilde{x}_j$ drawn following an (L, λ)-ALS Nyström subsampling scheme, and P be the projection operator with its range S. Let $\lambda = n^{-\theta}$ if $\theta \in [0, 1)$, or $\lambda = \frac{1 \vee \log n^\gamma}{n}$ if $\theta = 1$. Then with probability at least $1 - 3\delta$ ($\delta \in (0, 1/3)$), we have*

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

provided that

$$m \geq \bar{C}_1 n^{\theta\gamma}(1 \vee \log n^\gamma)^c \log^3\frac{4}{\delta}, \quad c = \begin{cases} 1, & \text{if } \theta < 1, \\ 1 - \gamma, & \text{if } \theta = 1. \end{cases} \tag{62}$$

Here, $\bar{C}_1 = 8b_\gamma L^2(4 + \log(2b_\gamma))$ where a_γ and b_γ 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*

$$T = \lceil(\eta\lambda)^{-1}\rceil, \quad \lambda = n^{-\frac{1}{1\vee(2\zeta+\gamma)}}(1 \vee \log n^\gamma)^{\mathbf{1}_{\{2\zeta+\gamma\leq 1\}}}, \tag{63}$$

and let

$$0 < \eta \leq \frac{1}{8\kappa^2(\log T + 1)}. \tag{64}$$

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

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_{T+1} - f_H)\|_\rho^2 \lesssim \lambda^{2(\zeta-a)}\left(1 + \lambda^{-1}\Delta_5\right)^2 \log^4\frac{2}{\delta} + \eta b^{-1}\lambda^{-2a}(\log T)^{\mathbf{1}_{\{2a\neq 1\}}}\log^2\frac{2}{\delta}. \tag{65}$$

2) If $\zeta \geq 1$, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{J}}\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_{T+1} - f_H)\|_\rho^2 \\ \lesssim&\lambda^{-2a}\left(\lambda^\zeta + \Delta_5\left(1 + \left(\frac{\lambda}{\Delta_5}\right)\Delta_5^{(\zeta-1)\wedge 1} + \left(\frac{\lambda}{\Delta_5}\right)^a\right)\right)^2 \log^4\frac{2}{\delta} + \eta b^{-1}\lambda^{-2a}(\log T)^{\mathbf{1}_{\{2a\neq 1\}}}\log^2\frac{2}{\delta}.\end{aligned} \tag{66}$$

Furthermore, if $\zeta \geq 1/2$, then the above conclusions still hold if we replace $\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_{T+1} - f_H)\|_\rho$ by $\|\mathcal{T}^{\frac{1}{2}-a}(\omega_{T+1} - \omega_H)\|_H$. Here, Δ_5 is the projection error $\|(I - P)\mathcal{T}^{\frac{1}{2}}\|^2$.

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 $\nu_1 = 0$,

$$\nu_{t+1} = \widetilde{\mathcal{G}}_t(P\mathcal{T}_{\mathbf{x}}P)\mathcal{S}_{\mathbf{x}}^*\bar{\mathbf{y}}, \quad \widetilde{\mathcal{G}}_t(\cdot) = \sum_{k=1}^t \eta_k \prod_{i=k+1}^t (I - \eta_i \cdot).$$

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

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_{T+1} - f_H)\|_\rho^2 = \|\mathcal{L}^{-a}(\mathcal{S}_\rho\nu_{T+1} - f_H)\|_\rho^2 + \mathbb{E}_{\mathbf{J}}\|\mathcal{L}^{-a}\mathcal{S}_\rho(\omega_{T+1} - \nu_{T+1})\|_\rho^2.$$

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

We first estimate $\|\mathcal{L}^{-a}(\mathcal{S}_\rho\nu_{T+1} - f_H)\|_\rho^2$. As noted in Remark 1, $\widetilde{\mathcal{G}}_t(\cdot)$ 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 $f_\lambda^{\mathbf{z}} = \mathcal{S}_\rho\omega_{T+1}$.

What remains is to prove the following error bounds:

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{L}^{-a}\mathcal{S}_\rho(\omega_{T+1} - \nu_{T+1})\|_\rho^2 \lesssim \eta b^{-1}\lambda^{-2a}(\log n)^{\mathbf{1}_{\{2a\neq 1\}}}\log^2\frac{2}{\delta}. \tag{67}$$

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

$$\|\mathcal{L}^{-a}\mathcal{S}_\rho(\omega_{T+1} - \nu_{T+1})\|_\rho \le \|\mathcal{T}^{\frac{1}{2}-a}(\omega_{T+1} - \nu_{T+1})\|_H \le \lambda^{-a}\Delta_1^{\frac{1}{2}-a}\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_{T+1} - \nu_{T+1})\|_H.$$

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

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{T}_{\mathbf{x}\lambda}^{\frac{1}{2}}(\omega_{T+1} - \nu_{T+1})\|_H^2 \le 48\kappa^2\mathcal{E}_{\mathbf{z}}(0)\eta b^{-1}\log(3T).$$

Thus, we have

$$\mathbb{E}_{\mathbf{J}}\|\mathcal{L}^{-a}\mathcal{S}_\rho(\omega_{T+1} - \nu_{T+1})\|_\rho^2 \le \lambda^{-2a}\Delta_1^{1-2a}48\kappa^2\mathcal{E}_{\mathbf{z}}(0)\eta b^{-1}\log(3T).$$

Applying (Lin and Rosasco, 2017b, Lemma 25) and Lemma 14 to estimate $\mathcal{E}_{\mathbf{z}}(0)$ and Δ_1 respectively, we can prove that (67) holds with probability at least $1-\delta$. The proof for the case $a = \frac{1}{2}$ is simpler. In fact, by (5), we have

$$\|\mathcal{L}^{-1/2}\mathcal{S}_\rho(\omega_{T+1} - \nu_{T+1})\|_\rho \le \|\omega_{T+1} - \nu_{T+1}\|_H.$$

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

$$\mathbb{E}_{\mathbf{J}}\|\omega_{T+1} - \nu_{T+1}\|_H^2 \lesssim \eta b^{-1}\lambda^{-1}\mathcal{E}_{\mathbf{z}}(0).$$

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 η, 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 $\mathcal{X}_1, \cdots, \mathcal{X}_m$ be a sequence of independently and identically distributed self-adjoint Hilbert-Schmidt operators on a separable Hilbert space. Assume that $\mathbb{E}[\mathcal{X}_1] = 0$, and $\|\mathcal{X}_1\| \leq B$ almost surely for some $B > 0$. Let $\mathcal{V}$ be a positive trace-class operator such that $\mathbb{E}[\mathcal{X}_1^2] \preccurlyeq \mathcal{V}$. Then with probability at least $1-\delta$, ($\delta \in]0,1[$), there holds*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \mathcal{X}_i \right\| \leq \frac{2B\beta}{3m} + \sqrt{\frac{2\|\mathcal{V}\|\beta}{m}}, \qquad \beta = \log \frac{4 \operatorname{tr} \mathcal{V}}{\|\mathcal{V}\|\delta}.$$

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:*

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

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 $\sqrt{\frac{2\kappa^2\beta}{|\mathbf{x}|\lambda'}} = u$, and combining with Lemma 23, we know that if

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

which is equivalent to

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

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

$$\left\|\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}-\mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2}\right\| \leq 2/3. \tag{69}$$

Note that (69) implies

$$\|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|^2 \vee \|\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\mathcal{T}_{\lambda'}^{-1/2}\|^2 \leq 3. \tag{70}$$

Indeed,

$$\|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|^2 = \|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1}\mathcal{T}_{\lambda'}^{1/2}\| = \|(I - \mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T} - \mathcal{T}_{\mathbf{x}})\mathcal{T}_{\lambda'}^{-1/2})^{-1}\| \leq 3,$$

and

$$\|\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\mathcal{T}_{\lambda'}^{-1/2}\|^2 = \|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}\mathcal{T}_{\lambda'}^{-1/2}\| = \|\mathcal{T}_{\lambda'}^{-1/2}(\mathcal{T}_{\mathbf{x}} - \mathcal{T})\mathcal{T}_{\lambda'}^{-1/2} + I\| \leq 3.$$

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 λ' 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

$$\|\mathcal{T}_{\lambda}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\| \leq \|\mathcal{T}_{\lambda}^{1/2}\mathcal{T}_{\lambda'}^{-1/2}\|\|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|\|\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\| \leq \|\mathcal{T}_{\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{-1/2}\|\sqrt{\lambda'/\lambda},$$

where for the last inequality, we used $\|\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\|^2 \leq \sup_{u\geq 0}\frac{u+\lambda'}{u+\lambda} \leq \lambda'/\lambda$. Similarly,

$$\|\mathcal{T}_{\lambda}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda}^{1/2}\| \leq \|\mathcal{T}_{\lambda'}^{-1/2}\mathcal{T}_{\mathbf{x}\lambda'}^{1/2}\|\sqrt{\lambda'/\lambda}.$$

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 $\lambda = |\mathbf{x}|^{-\theta}$ with $\theta \in [0,1)$,

$$\beta \leq \log\frac{4\kappa^2(1 + c_\gamma a^{-\gamma}|\mathbf{x}|^{\theta\gamma})}{\delta\|\mathcal{T}\|} \leq \log\frac{4\kappa^2(1+c_\gamma)|\mathbf{x}|^{\theta\gamma}}{\delta\|\mathcal{T}\|} = \log\frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \log|\mathbf{x}|^{\theta\gamma},$$

while for $\lambda = (1 \vee \log|\mathbf{x}|^{\gamma})/|\mathbf{x}|$,

$$\beta \leq \log\frac{4\kappa^2(1 + c_\gamma a^{-\gamma}\lambda^{-\gamma})}{\delta\|\mathcal{T}\|} \leq \log\frac{4\kappa^2(1+c_\gamma)|\mathbf{x}|^{\gamma}}{\delta\|\mathcal{T}\|} = \log\frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \log|\mathbf{x}|^{\gamma},$$

If $\lambda = |\mathbf{x}|^{-\theta}$ with $\theta \in [0,1)$ and $\theta\gamma = 0$, or $\lambda = (1 \vee \log|\mathbf{x}|^{\gamma})/|\mathbf{x}|$, then the condition (68) follows trivially. Now consider the case $\lambda = |\mathbf{x}|^{-\theta}$ with $\theta \in (0,1)$ and $\theta\gamma \neq 0$. The maximum of the function $g(u) = \mathrm{e}^{-cu}u^{\alpha}$ (with $c > 0$) over $\mathbb{R}_+$ is achieved at $u_{\max} = \alpha/c$, and thus

$$\sup_{u\geq 0}\mathrm{e}^{-cu}u^{\alpha} = \left(\frac{\alpha}{\mathrm{e}c}\right)^{\alpha}. \tag{71}$$

We apply the above with $u = |\mathbf{x}|^{\theta\gamma\zeta'}$, $\alpha = 1/\zeta'$, we know that for any $c', \zeta' > 0$

$$\beta \leq \log\frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + c'|\mathbf{x}|^{\theta\gamma\zeta'} + \frac{1}{\zeta'}\log\frac{1}{\zeta'\mathrm{e}c'}.$$

Selecting $\zeta' = \frac{1-\theta}{\theta\gamma}$ and $c' = \frac{\theta\gamma}{\mathrm{e}(1-\theta)}$, we know that a sufficient condition for (68) is

$$\frac{|\mathbf{x}|^{1-\theta}a}{8\kappa^2} \geq \log\frac{4\kappa^2(1+c_\gamma)}{\delta\|\mathcal{T}\|} + \frac{\theta\gamma}{\mathrm{e}(1-\theta)}|\mathbf{x}|^{1-\theta}.$$

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 σ^2 such that*

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

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

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

In particular, (72) holds if

$$\|w_1\| \leq B/2 \quad a.s., \quad \text{and} \quad \mathbb{E}[\|w_1\|^2] \leq \sigma^2. \tag{73}$$

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 $\xi_i = \mathcal{T}_\lambda^{-\frac{1}{2}}(\langle \omega_H^\lambda, x_i \rangle_H - y_i) x_i$ for all $i \in [n]$. It is easy to see that ξ_i is a random variable depending on (x_i, y_i). From the definition of the regression function f_ρ in (8) and (9), a simple calculation shows that

$$\mathbb{E}[\xi] = \mathbb{E}[\mathcal{T}_\lambda^{-\frac{1}{2}}(\langle \omega_H^\lambda, x \rangle_H - f_\rho(x)) x] = \mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}\omega_H^\lambda - \mathcal{S}_\rho^* f_\rho) = \mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}\omega_H^\lambda - \mathcal{S}_\rho^* f_H). \tag{74}$$

Combining with the definition of $\mathcal{T}_{\mathbf{x}}$ and $\mathcal{S}_{\mathbf{x}}^*$, we have

$$\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}}\omega_H^\lambda - \mathcal{S}_{\mathbf{x}}^* \bar{\mathbf{y}} - \mathcal{T}\omega_H^\lambda + \mathcal{S}_\rho^* f_H)\|_H = \left\| \frac{1}{n} \sum_{i=1}^{n} (\xi_i - \mathbb{E}[\xi]) \right\|_H$$

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

$$\mathbb{E}\|\xi - \mathbb{E}[\xi]\|_H^l \leq \mathbb{E}\left(\|\xi\|_H + \mathbb{E}\|\xi\|_H\right)^l \leq 2^{l-1}(\mathbb{E}\|\xi\|_H^l + (\mathbb{E}\|\xi\|_H)^l) \leq 2^l \mathbb{E}\|\xi\|_H^l. \tag{75}$$

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

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

According to (2), one has

$$\|\mathcal{T}_\lambda^{-\frac{1}{2}} x\|_H \leq \|\mathcal{T}_\lambda^{-\frac{1}{2}}\| \|x\|_H \leq \frac{1}{\sqrt{\lambda}} \kappa. \tag{76}$$

Moreover, by Cauchy-Schwarz inequality and (2), $|\langle \omega_H^\lambda, x\rangle_H| \le \|\omega_H^\lambda\|_H \|x\|_H \le \kappa\|\omega_H^\lambda\|_H$. Thus, with $|f_\rho(x)| \le M$ by Assumption 1, we get

$$\mathbb{E}\|\xi\|_H^l \le 2^{l-1}\left(\frac{\kappa}{\sqrt{\lambda}}\right)^{l-2} \mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2(|y-f_\rho(x)|^l + (M+\kappa\|\omega_H^\lambda\|_H)^{l-2}|\langle\omega_H^\lambda, x\rangle_H - f_\rho(x)|^2). \tag{77}$$

Note that by (11),

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

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

$$\int_H \|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2 d\rho_X(x) = \int_H \operatorname{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}}x\otimes x\mathcal{T}_\lambda^{-\frac{1}{2}})d\rho_X(x) = \operatorname{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{T}\mathcal{T}_\lambda^{-\frac{1}{2}}) = \mathcal{N}(\lambda), \tag{78}$$

we get

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2|y-f_\rho(x)|^l] \le \frac{1}{2} l! M^{l-2}Q^2\mathcal{N}(\lambda). \tag{79}$$

Besides, by Cauchy-Schwarz inequality,

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

By (76) and (39),

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2(|\langle\omega_H^\lambda, x\rangle_H - f_H(x)|^2] \le \frac{\kappa^2}{\lambda}\mathbb{E}[|\langle\omega_H^\lambda, x\rangle_H - f_H(x)|^2] = \frac{\kappa^2}{\lambda}\|\mathcal{S}_\rho\omega_H^\lambda - f_H\|_\rho^2 \le R^2\kappa^2\lambda^{2\zeta-1}.$$

Therefore,

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

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

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

and therefore,

$$\mathbb{E}[\|\mathcal{T}_\lambda^{-\frac{1}{2}}x\|_H^2|\langle\omega_H^\lambda, x\rangle_H - f_\rho(x)|^2] \le 2\left(\kappa^2R^2\lambda^{2\zeta-1} + B^2\mathcal{N}(\lambda)\right).$$

Introducing the above estimate and (79) into (77), we derive

$$\begin{aligned}\mathbb{E}\|\xi\|_H^l \leq & 2^{l-1}\left(\frac{\kappa}{\sqrt{\lambda}}\right)^{l-2}\left(\frac{1}{2}l!M^{l-2}Q^2\mathcal{N}(\lambda)+2(M+\kappa\|\omega_H^\lambda\|_H)^{l-2}(R^2\kappa^2\lambda^{2\zeta-1}+B^2\mathcal{N}(\lambda))\right)\\ \leq & 2^{l-1}\left(\frac{\kappa M+\kappa^2\|\omega_H^\lambda\|_H}{\sqrt{\lambda}}\right)^{l-2}\frac{1}{2}l!\left(2R^2\kappa^2\lambda^{2\zeta-1}+(2B^2+Q^2)\mathcal{N}(\lambda)\right).\end{aligned}$$

Introducing the above estimate into (75), and then substituting with (40), we get

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

Applying Lemma 24, we get the desired result. ■

A.3. Proof of Lemma 17

Let $\mathcal{S}_{\mathbf{x}} = U\Sigma V^*$ be the singular value decomposition of $\mathcal{S}_{\mathbf{x}}$, where $V:\mathbb{R}^r \to H$, $U \in \mathbb{R}^{n\times r}$ and $\Sigma = \mathrm{diag}(\sigma_1,\sigma_2,\cdots,\sigma_r)$ with $V^*V = I_r$, $U^*U = I_r$ and $\sigma_1 \geq \sigma_2,\cdots,\sigma_r > 0$. In fact, we can write $V = [v_1,\cdots,v_r]$ with

$$V\mathbf{a} = \sum_{i=1}^r \mathbf{a}(i)v_i, \quad \forall \mathbf{a}\in\mathbb{R}^r,$$

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 $U = [u_1,\cdots,u_r]$, and

$$\mathcal{S}_{\mathbf{x}} = \sum_{i=1}^r \sigma_i\langle v_i,\cdot\rangle_H u_i = \sum_{i=1}^r \sigma_i u_i\otimes v_i.$$

For any $\mu \geq 0$, we decompose $\mathcal{S}_{\mathbf{x}}$ as $\mathcal{S}_{1,\mu}+\mathcal{S}_{2,\mu}$ with

$$\mathcal{S}_{1,\mu} = \sum_{\sigma_i>\mu}\sigma_i u_i\otimes v_i, \quad \mathcal{S}_{2,\mu} = \sum_{\sigma_i\leq\mu}\sigma_i u_i\otimes v_i,$$

and we will drop μ to write $\mathcal{S}_{j,\mu}$ as $\mathcal{S}_j$ when it is clear in the text. Denote d the cardinality of $\{\sigma_i:\sigma_i>\mu\}$. Correspondingly,

$$\mathcal{S}_1 = U_1\Sigma_1V_1^*, \quad \mathcal{S}_2 = U_2\Sigma_2V_2^*, \tag{80}$$

where $V_1 = [v_1,\cdots,v_d]$, $V_2 = [v_{d+1},\cdots,v_r]$, $U_1 = [u_1,\cdots,u_d]$, $U_2 = [u_{d+1},\cdots,u_r]$, $\Sigma_1 = \mathrm{diag}(\sigma_1,\cdots,\sigma_d)$, and $\Sigma_2 = \mathrm{diag}(\sigma_{d+1},\cdots,\sigma_r)$. As the range of P is range($\mathcal{S}_{\mathbf{x}}^*\mathbf{G}^*$), we can let

$$P = P_1+P_2,$$

where P_1 and P_2 are projection operators on range($\mathcal{S}_1^*\mathbf{G}^*$) and range($\mathcal{S}_2^*\mathbf{G}^*$), respectively.

As

$$\mathcal{T}_{\mathbf{x}} = \mathcal{S}_{\mathbf{x}}^*\mathcal{S}_{\mathbf{x}} = (U\Sigma V^*)^*U\Sigma V^* = V\Sigma^2V^*,$$

we have

$$\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\| = \|(I-P)V\Sigma V^*\| = \|(I-P_1-P_2)\sum_{i=1}^{2} V_i\Sigma_i V_i^*\|.$$

As P_1 is a projection operator on range$(\mathcal{S}_1^*\mathbf{G}^*)(\subseteq \text{range}(V_1))$ and range$(\mathcal{S}_1^*\mathbf{G}^*)(\subseteq \text{range}(V_2))$, and $V_1^*V_2 = \mathbf{0}$, we know that $P_iV_j = 0$ when $i \neq j$. Thus, it follows that

$$\begin{aligned}\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\| =& \|\sum_{i=1}^{2}(I-P_i)(V_i\Sigma_i V_i^*)\| \\ \leq& \sum_{i=1}^{2}\|(I-P_i)(V_i\Sigma_i V_i^*)\| \\ \leq& \|(I-P_1)(V_1\Sigma_1 V_1^*)\| + \|I-P_2\|\|V_2\|\|\Sigma_2\|\|V_2^*\|.\end{aligned}$$

As $\Sigma_2 = \text{diag}(\sigma_{d+1},\cdots,\sigma_r)$ with $\sigma_r \leq,\cdots,\sigma_{d+1} \leq \mu$, we get

$$\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\| \leq \|(I-P_1)(V_1\Sigma_1 V_1^*)\| + \mu. \tag{81}$$

As P_1 is the projection operator on range$(\mathcal{S}_1^*\mathbf{G}^*)$, letting $W = \mathbf{G}\mathcal{S}_1$ and for any $\lambda > 0$,

$$P_1 = W^*(WW^*)^\dagger W \succeq W^*(WW^*+\lambda I)^{-1}W = W^*W(W^*W+\lambda I)^{-1},$$

and thus

$$I - P_1 \preceq I - W^*W(W^*W+\lambda I)^{-1} = \lambda(W^*W+\lambda I)^{-1}.$$

It thus follows that

$$T_1^{\frac{1}{2}}(I-P_1)T_1^{\frac{1}{2}} \preceq \lambda T_1^{\frac{1}{2}}(W^*W+\lambda I)^{-1}T_1^{\frac{1}{2}},$$

where for notational simplicity, we write

$$T_1 = (V_1\Sigma_1 V_1^*)^2. \tag{82}$$

Combing with

$$\|(I-P)T_1^{\frac{1}{2}}\|^2 = \|T_1^{\frac{1}{2}}(I-P)^2T_1^{\frac{1}{2}}\| = \|T_1^{\frac{1}{2}}(I-P)T_1^{\frac{1}{2}}\|,$$

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,

$$\|(I-P)T_1^{\frac{1}{2}}\|^2 \le \lambda \|T_{1\lambda}^{-\frac{1}{2}}(W^*W+\lambda I)^{-1}T_{1\lambda}^{-\frac{1}{2}}\| \le (1-c)^{-1}\lambda. \tag{84}$$

If we choose $\mu = \sqrt{\lambda}$, and introduce the above with $c = \frac{1}{2}$ into (81), one can get

$$\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 \le (\sqrt{2}+1)^2\lambda \le 6\lambda, \tag{85}$$

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 $W = \mathbf{G}\mathcal{S}_1$ with $\mathcal{S}_1$ given by (80). Thus, $T_1 = V_1\Sigma_1V_1^*V_1\Sigma_1V_1^* = V_1\Sigma_1^2V_1^*$, and

$$W^*W = \mathcal{S}_1^*\mathbf{G}^*\mathbf{G}\mathcal{S}_1 = V_1\Sigma_1U_1^*\mathbf{G}^*\mathbf{G}U_1\Sigma_1V_1^*.$$

Therefore, with $V_1^*V_1 = I$,

$$\begin{aligned} T_{1\lambda}^{-\frac{1}{2}}(T_1 - W^*W)T_{1\lambda}^{-\frac{1}{2}} =& V_1(\Sigma_1^2+\lambda I)^{-1/2}V_1^*V_1\Sigma_1(I-U_1^*\mathbf{G}^*\mathbf{G}U_1)\Sigma_1V_1^*V_1(\Sigma_1^2+\lambda I)^{-1/2}V_1^* \\ =& V_1(\Sigma_1^2+\lambda I)^{-1/2}\Sigma_1(I-U_1^*\mathbf{G}^*\mathbf{G}U_1)\Sigma_1(\Sigma_1^2+\lambda I)^{-1/2}V_1^*. \end{aligned} \tag{86}$$

It follows that

$$\|T_{1\lambda}^{-\frac{1}{2}}(T_1-W^*W)T_{1\lambda}^{-\frac{1}{2}}\| \le \|V_1\|\|(\Sigma_1^2+\lambda I)^{-1/2}\Sigma_1\|^2\|I-U_1^*\mathbf{G}^*\mathbf{G}U_1\|\|V_1^*\| \le \|I-U_1^*\mathbf{G}^*\mathbf{G}U_1\|.$$

Using $U_1^*U_1 = I$,

$$\begin{aligned} \|I-U_1^*\mathbf{G}^*\mathbf{G}U_1\| =& \|U_1^*(I-\mathbf{G}^*\mathbf{G})U_1\| \\ =& \max_{\mathbf{a}\in\mathbb{R}^d, \|\mathbf{a}\|_2=1} |\langle U_1^*(I-\mathbf{G}^*\mathbf{G})U_1\mathbf{a}, \mathbf{a}\rangle_2| \\ =& \max_{\mathbf{a}\in\mathbb{R}^d, \|\mathbf{a}\|_2=1} |\|U_1\mathbf{a}\|_2^2 - \|\mathbf{G}U_1\mathbf{a}\|_2^2|. \end{aligned}$$

Based on a standard argument as that for (Baraniuk et al., 2008, Lemma 5.1), we know that

$$\max_{\mathbf{a}\in\mathbb{R}^d, \|\mathbf{a}\|_2=1} |\|U_1\mathbf{a}\|_2^2 - \|\mathbf{G}U_1\mathbf{a}\|_2^2| \le \frac{1}{2}$$

with probability at least

$$1 - 2(60)^d \exp\left(-\frac{m}{100c_0' \log^\beta n}\right) \ge 1-\delta,$$

provided that

$$m \ge 100c_0' \log^\beta n \left(\log\frac{2}{\delta} + 5d\right). \tag{87}$$

Note that by (57),

$$b_\gamma \lambda^{-\gamma} \ge \operatorname{tr}(\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\lambda}^{-1}) = \sum_i \frac{\sigma_i^2}{\sigma_i^2+\lambda} \ge \sum_{\sigma_i^2>\lambda} \frac{\sigma_i^2}{\sigma_i^2+\lambda} \ge \frac{d}{2}.$$

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 $\mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - \mathcal{T})\mathcal{T}_\lambda^{-\frac{1}{2}})$. Note that

$$\mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{T}_{\mathbf{x}}\mathcal{T}_\lambda^{-\frac{1}{2}}) = \frac{1}{n}\sum_{j=1}^n \|\mathcal{T}_\lambda^{-\frac{1}{2}} x_j\|_H^2 = \frac{1}{n}\sum_{j=1}^n \xi_j,$$

where we let $\xi_j = \|\mathcal{T}_\lambda^{-\frac{1}{2}} x_j\|_H^2$ for all $j \in [n]$. Besides, it is easy to see that

$$\mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - \mathcal{T})\mathcal{T}_\lambda^{-\frac{1}{2}}) = \frac{1}{n}\sum_{j=1}^n (\xi_j - \mathbb{E}[\xi_j]).$$

Using Assumption (2),

$$\xi_1 \le \frac{1}{\lambda}\|x_1\|_H^2 \le \frac{\kappa^2}{\lambda},$$

and

$$\mathbb{E}[\|\xi_1\|^2] \le \frac{\kappa^2}{\lambda}\mathbb{E}\|\mathcal{T}_\lambda^{-\frac{1}{2}} x_1\|_H^2 = \frac{\kappa^2}{\lambda}\mathbb{E}\,\mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}} x_1 \otimes x_1 \mathcal{T}_\lambda^{-\frac{1}{2}}) = \frac{\kappa^2\mathcal{N}(\lambda)}{\lambda}.$$

Applying Lemma 24, we get that there exists a subset Ω_1 of H^n with measure at least $1-\delta$, such that for all $\mathbf{x} \in \Omega_1$,

$$\mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - \mathcal{T})\mathcal{T}_\lambda^{-\frac{1}{2}}) \le 2\left(\frac{2\kappa^2}{n\lambda} + \sqrt{\frac{\kappa^2\mathcal{N}(\lambda)}{n\lambda}}\right)\log\frac{2}{\delta}.$$

Combining with Lemma 14, taking the union bounds, rescaling δ, and noting that

$$\begin{aligned}
\mathrm{tr}(\mathcal{T}_{\mathbf{x}\lambda}^{-1}\mathcal{T}_{\mathbf{x}}) =& \mathrm{tr}(\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\mathcal{T}_\lambda^{\frac{1}{2}}\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{T}_{\mathbf{x}}\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{T}_\lambda^{\frac{1}{2}}\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}) \\
\le& \|\mathcal{T}_\lambda^{\frac{1}{2}}\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\|^2\,\mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}}\mathcal{T}_{\mathbf{x}}\mathcal{T}_\lambda^{-\frac{1}{2}}) \\
=& \|\mathcal{T}_\lambda^{\frac{1}{2}}\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\|^2\left(\mathrm{tr}(\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - \mathcal{T})\mathcal{T}_\lambda^{-\frac{1}{2}}) + \mathcal{N}(\lambda)\right).
\end{aligned}$$

we get that there exists a subset Ω of H^n with measure at least $1-\delta$, such that for all $\mathbf{x} \in \Omega$,

$$\mathrm{tr}(\mathcal{T}_{\mathbf{x}\lambda}^{-1}\mathcal{T}_{\mathbf{x}}) \le 3a(\delta/2)\left(2\left(\frac{2\kappa^2}{n\lambda} + \sqrt{\frac{\kappa^2\mathcal{N}(\lambda)}{n\lambda}}\right)\log\frac{4}{\delta} + \mathcal{N}(\lambda)\right),$$

which leads to the desired result using $\lambda \le 1$, $n\lambda \ge 1$ and Assumption 3.

A.5. Estimating Projection Errors with Random Sketches

Proof of Lemma 19 Let $\mu = \frac{1 \vee \log n^\gamma}{n}$, and $\lambda = n^{-\theta}$ with $\theta \in [0,1)$ or $\lambda = \frac{1\vee\log n^\gamma}{n}$. By a simple calculation,

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \le \|(I-P)\mathcal{T}_{\mathbf{x}\mu}^{\frac{1}{2}}\|^2\|\mathcal{T}_{\mathbf{x}\mu}^{-\frac{1}{2}}\mathcal{T}_\mu^{\frac{1}{2}}\|^2.$$

Using

$$\|(I-P)\mathcal{T}_{\mathbf{x}\mu}^{\frac{1}{2}}\|^2 = \|(I-P)\mathcal{T}_{\mathbf{x}\mu}(I-P)\| \le \|(I-P)\mathcal{T}_{\mathbf{x}}(I-P)\|+\mu\|(I-P)^2\| \le \|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2+\mu,$$

we get

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \le \left(\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 + \mu\right)\|\mathcal{T}_{\mathbf{x}\mu}^{-\frac{1}{2}}\mathcal{T}_{\mu}^{\frac{1}{2}}\|^2. \tag{88}$$

Following from Lemma 18 and Lemma 14, we know that there exists a subset Ω_1 of H^n with measure at least $1-2\delta$ such that for every $\mathbf{x} \in \Omega_1$,

$$\operatorname{tr}(\mathcal{T}_{\mathbf{x}\lambda}^{-1}\mathcal{T}_{\mathbf{x}}) \le b_{\gamma,\delta}\lambda^{-\gamma},$$

and

$$\|\mathcal{T}_{\mathbf{x}\mu}^{-\frac{1}{2}}\mathcal{T}_{\mu}^{\frac{1}{2}}\|^2 \le a_\gamma \log\frac{4}{\delta}, \tag{89}$$

where $b_{\gamma,\delta} = b_\gamma \log^2 \frac{4}{\delta}$. For every $\mathbf{x} \in \Omega_1$, according to Lemma 17, we know that there exists a subset $U_{\mathbf{x}}$ of $\mathbb{R}^{m\times n}$ with measure at least $1-\delta$, such that for all $\mathbf{G} \in U_{\mathbf{x}}$,

$$\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 \le 6\lambda, \tag{90}$$

provided that,

$$m \ge 100c_0' \log^\beta n\lambda^{-\gamma}\log^3\frac{4}{\delta}\left(1+10b_\gamma\right),$$

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 $1-3\delta$,

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \le (6\lambda+\mu)\, a_\gamma \log\frac{4}{\delta},$$

which leads to the desired result. ■

A.6. Estimating Projection Errors with Plain Nyström Subsampling

Proof of Lemma 20 As P is the projection operator onto $\overline{\operatorname{range}\{\mathcal{S}_{\tilde{\mathbf{x}}}^*\}}$ with $\tilde{\mathbf{x}} = \{x_1,\cdots,x_m\}$,

$$P = \mathcal{S}_{\tilde{\mathbf{x}}}^*(\mathcal{S}_{\tilde{\mathbf{x}}}\mathcal{S}_{\tilde{\mathbf{x}}}^*)^\dagger\mathcal{S}_{\tilde{\mathbf{x}}} \succeq \mathcal{S}_{\tilde{\mathbf{x}}}^*(\mathcal{S}_{\tilde{\mathbf{x}}}\mathcal{S}_{\tilde{\mathbf{x}}}^*+\mu I)^{-1}\mathcal{S}_{\tilde{\mathbf{x}}} = \mathcal{S}_{\tilde{\mathbf{x}}}^*\mathcal{S}_{\tilde{\mathbf{x}}}(\mathcal{S}_{\tilde{\mathbf{x}}}^*\mathcal{S}_{\tilde{\mathbf{x}}}+\mu I)^{-1} = \mathcal{T}_{\tilde{\mathbf{x}}}(\mathcal{T}_{\tilde{\mathbf{x}}}+\mu I)^{-1},$$

where for the last second equality, we used Lemma 9. Thus,

$$I-P \preceq I - \mathcal{T}_{\tilde{\mathbf{x}}}(\mathcal{T}_{\tilde{\mathbf{x}}}+\mu I)^{-1} = \mu(\mathcal{T}_{\tilde{\mathbf{x}}}+\mu I)^{-1}.$$

It thus follows that

$$\mathcal{T}_\mu^{\frac{1}{2}}(I-P)^{\frac{1}{2}}\mathcal{T}_\mu^{\frac{1}{2}} \preceq \mu\mathcal{T}_\mu^{\frac{1}{2}}(\mathcal{T}_{\tilde{\mathbf{x}}}+\mu I)^{-1}\mathcal{T}_\mu^{\frac{1}{2}}.$$

Using $\|A^*A\|^2 = \|A\|^2$ and the above,

$$\|(I-P)\mathcal{T}_\mu^{\frac{1}{2}}\|^2 = \|\mathcal{T}_\mu^{\frac{1}{2}}(I-P)\mathcal{T}_\mu^{\frac{1}{2}}\| \le \mu\|\mathcal{T}_\mu^{\frac{1}{2}}(\mathcal{T}_{\tilde{\mathbf{x}}}+\mu I)^{-1}\mathcal{T}_\mu^{\frac{1}{2}}\| = \mu\|(\mathcal{T}_{\tilde{\mathbf{x}}}+\mu I)^{-1/2}\mathcal{T}_\mu^{\frac{1}{2}}\|^2. \tag{91}$$

Thus,

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \le \|(I-P)\mathcal{T}_\mu^{\frac{1}{2}}\|^2 \le \mu\|(\mathcal{T}_{\tilde{\mathbf{x}}}+\mu I)^{-1/2}(\mathcal{T}+\mu I)^{1/2}\|^2.$$

Using Lemma 14 with $\mu = \frac{1\vee\log m^\gamma}{m}$, one can prove the desired result. ■

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 $S = \overline{\mathrm{range}\{\mathcal{S}_{\mathbf{x}}^* \mathbf{G}^\top\}}$, where each row $\frac{1}{\sqrt{m}}\mathbf{a}_j^\top$ of $\mathbf{G}$ is i.i.d. drawn according to

$$\mathbb{P}\left(\mathbf{a} = \frac{1}{\sqrt{q_i}}\mathbf{e}_i\right) = q_i, \quad i \in \{1, \cdots, n\}$$

Here $\{\mathbf{e}_i : i \in [n]\}$ is the standard basis of $\mathbb{R}^n$ and

$$q_i := q_i(\lambda) = \frac{\hat{l}_i(\lambda)}{\sum_j \hat{l}_j(\lambda)}.$$

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 $\mathbf{x} \subseteq H^n$, assume that for $\lambda \in [0,1]$, (57) holds for some $b_\gamma > 0$, $\gamma \in [0,1]$. Then there exists a subset $U_{\mathbf{x}}$ of $\mathbb{R}^{m\times n}$ with measure at least $1-\delta$, such that for all $\mathbf{G} \in U_{\mathbf{x}}$,*

$$\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 \leq 3\lambda, \tag{92}$$

provided that

$$m \geq 8 b_\gamma \lambda^{-\gamma} L^2 \log \frac{8 b_\gamma \lambda^{-\gamma}}{\delta}. \tag{93}$$

Proof If we choose $u = 0$ in the proof of Lemma 17, then $\mathcal{S}_{\mathbf{x}} = \mathcal{S}_1$ and $\mathcal{S}_2 = 0$. Similarly, $\mathcal{T}_{\mathbf{x}} = T_1$. In this case, (86) reads as

$$\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - W^*W)\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}} = V(\Sigma^2+\lambda I)^{-1/2}\Sigma(I - U^*\mathbf{G}^*\mathbf{G}U)\Sigma(\Sigma^2+\lambda I)^{-1/2}V^*.$$

Thus, using $V^*V = I$, $U^*U = I$ and U is of full column rank,

$$\begin{aligned}\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - W^*W)\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\| \leq& \|V\|\|U^*U(\Sigma^2+\lambda I)^{-1/2}\Sigma U^*(I-\mathbf{G}^*\mathbf{G})U\Sigma(\Sigma^2+\lambda I)^{-1/2}U^*U\| \\ \leq& \|U(\Sigma^2+\lambda I)^{-1/2}\Sigma U^*(I-\mathbf{G}^*\mathbf{G})U\Sigma(\Sigma^2+\lambda I)^{-1/2}U^*\|.\end{aligned}$$

Using $\mathbf{K} := \mathbf{K}_{\mathbf{x}\mathbf{x}} = \mathcal{S}_{\mathbf{x}}\mathcal{S}_{\mathbf{x}}^* = U\Sigma^2U^*$, we get

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - W^*W)\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\| \leq \|\left(\mathbf{K}(\mathbf{K}+\lambda I)^{-1}\right)^{1/2}(I-\mathbf{G}^*\mathbf{G})\left(\mathbf{K}(\mathbf{K}+\lambda I)^{-1}\right)^{1/2}\|.$$

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

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - W^*W)\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\| \leq \|\frac{1}{m}\sum_{i=1}^m(\mathbb{E}[\mathcal{X}_i] - \mathcal{X}_i)\|.$$

A simple calculation shows that

$$\begin{aligned}\|\mathcal{X}_i\| =& \mathbf{a}_i^* \left(\mathbf{K}(\mathbf{K}+\lambda I)^{-1}\right) \mathbf{a}_i \le \max_{j\in[n]} \frac{\left(\mathbf{K}(\mathbf{K}+\lambda I)^{-1}\right)_{jj}}{q_j} \\ =& \max_{j\in[n]} \frac{l_j(\lambda)}{q_j} = \max_{j\in[n]} \frac{l_j(\lambda)\sum_k \hat{l}_k(\lambda)}{\hat{l}_j(\lambda)} \le L^2 \sum_j l_j(\lambda) = L^2 \operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1}),\end{aligned}$$

and

$$\mathbb{E}[\mathcal{X}_i^2] = \mathbb{E}[\mathbf{a}_i^* \left(\mathbf{K}(\mathbf{K}+\lambda I)^{-1}\right) \mathbf{a}_i \mathcal{X}_i] \le L^2 \operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1})\mathbb{E}[\mathcal{X}_i] = L^2 \operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1})\mathbf{K}\mathbf{K}_\lambda^{-1}.$$

Thus,

$$\|\mathbb{E}[\mathcal{X}_i] - \mathcal{X}_i\| \le \mathbb{E}\|\mathcal{X}_i\| + \|\mathcal{X}_i\| \le 2L^2 \operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1}),$$

and

$$\mathbb{E}\Big[\big(\mathcal{X}_i - \mathbb{E}[\mathcal{X}_i]\big)^2\Big] \preceq \mathbb{E}[\mathcal{X}_i^2] \preceq L^2 \operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1})\mathbf{K}\mathbf{K}_\lambda^{-1}.$$

Letting $\mathcal{V} = L^2 \operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1})\mathbf{K}\mathbf{K}_\lambda^{-1}$, we have

$$\|\mathcal{V}\| \le L^2 \operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1}),$$

and

$$\frac{\operatorname{tr}(\mathcal{V})}{\|\mathcal{V}\|} = \frac{\operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1})}{\|\mathbf{K}\mathbf{K}_\lambda^{-1}\|} = \operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1})\left(1 + \frac{\lambda}{\|\mathbf{K}\|}\right).$$

Applying Lemma 22, noting that $\operatorname{tr}(\mathbf{K}\mathbf{K}_\lambda^{-1}) = \operatorname{tr}(\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\lambda}^{-1})$ and $\|\mathbf{K}\| = \|\mathcal{T}_{\mathbf{x}}\|$ as $\mathcal{T}_{\mathbf{x}} = \mathcal{S}_{\mathbf{x}}^*\mathcal{S}_{\mathbf{x}}$, we get that there exists a subset $U_{\mathbf{x}} \subseteq \mathbb{R}^{m\times n}$ with measure at least $1-\delta$ such that for all $\mathbf{G} \in U_{\mathbf{x}}$,

$$\|\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}(\mathcal{T}_{\mathbf{x}} - W^*W)\mathcal{T}_{\mathbf{x}\lambda}^{-\frac{1}{2}}\| \le \frac{4L^2 \operatorname{tr}(\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\lambda}^{-1})\beta}{3m} + \sqrt{\frac{2L^2 \operatorname{tr}(\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\lambda}^{-1})\beta}{m}}, \quad \beta = \log \frac{4\operatorname{tr}(\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\lambda}^{-1})(1+\lambda/\|\mathcal{T}_{\mathbf{x}}\|)}{\delta}.$$

If $\lambda \le \|\mathcal{T}_{\mathbf{x}}\|$, using Condition (57), we have

$$\beta \le \log \frac{4b_\gamma\lambda^{-\gamma}(1+\lambda/\|\mathcal{T}_{\mathbf{x}}\|)}{\delta} \le \log \frac{8b_\gamma\lambda^{-\gamma}}{\delta},$$

and, combining with (93),

$$\frac{4L^2 \operatorname{tr}(\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\lambda}^{-1})\beta}{3m} + \sqrt{\frac{2L^2 \operatorname{tr}(\mathcal{T}_{\mathbf{x}}\mathcal{T}_{\mathbf{x}\lambda}^{-1})\beta}{m}} \le \frac{2}{3}.$$

Thus,

$$\left\|\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}(\mathcal{T} - W^*W)\mathcal{T}_{\mathbf{x}\lambda}^{-1/2}\right\| \le \frac{2}{3}, \quad \forall \mathbf{G} \in U_{\mathbf{x}}.$$

Following from (83) and (84), one can prove (92) for the case $\lambda \le \|\mathcal{T}_{\mathbf{x}}\|$. The proof for the case $\lambda \ge \|\mathcal{T}_{\mathbf{x}}\|$ is trivial:

$$\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 \le \|I-P\|^2\|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 \le \|\mathcal{T}_{\mathbf{x}}\| \le \lambda.$$

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 Ξ be a closed subset of Euclidean space $\mathbb{R}^d$, the output space $Y \subseteq \mathbb{R}$. Let μ 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 μ. A reproducing kernel K is a symmetric function $K : \Xi \times \Xi \to \mathbb{R}$ such that $(K(u_i, u_j))_{i,j=1}^{\ell}$ is positive semidefinite for any finite set of points $\{u_i\}_{i=1}^{\ell}$ in Ξ. The kernel K defines a reproducing kernel Hilbert space (RKHS) $(\mathcal{H}_K, \|\cdot\|_K)$ as the completion of the linear span of the set $\{K_\xi(\cdot) := K(\xi, \cdot) : \xi \in \Xi\}$ with respect to the inner product $\langle K_\xi, K_u\rangle_K := K(\xi, u)$. For any $f \in \mathcal{H}_K$, the reproducing property holds: $f(\xi) = \langle K_\xi, f\rangle_K$.

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

$$K(x, x') = \begin{cases} (1-y)x, & x \le y; \\ (1-x)y, & x \ge y. \end{cases}$$

Then the kernel induces a Sobolev Space $H = \{f : X \to \mathbb{R} | f$ 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

Ahmed Alaoui and Michael W Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Advances in Neural Information Processing Systems*, pages 775–783, 2015.

Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(21):1–38, 2017.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28 (3):253–263, 2008.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

Gilles Blanchard and Nicole Mücke. Optimal rates for regularization of statistical inverse learning problems. *Foundations of Computational Mathematics*, 18(4):971–1013, 2018.

Andrea Caponnetto. Optimal learning rates for regularization operators in learning theory. *Technical report*, 2006.

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

Andrea Caponnetto and Yuan Yao. Adaptation for regularization operators in learning theory. 2006.

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

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

Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13(Dec):3475–3506, 2012.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithm. *arXiv 1702.07254v1*, 2017.

Junichi Fujii, Masatoshi Fujii, Takayuki Furuta, and Ritsuo Nakamoto. Norm inequalities equivalent to Heinz inequality. *Proceedings of the American Mathematical Society*, 118 (3):827–830, 1993.

L Lo Gerfo, Lorenzo Rosasco, Francesca Odone, Ernesto De Vito, and Alessandro Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

Alex Gittens and Michael W Mahoney. Revisiting the nyström method for improved large-scale machine learning. *Journal of Machine Learning Research*, 17(1):3977–4041, 2016.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Felix Krahmer and Rachel Ward. New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3): 1269–1281, 2011.

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar. Sampling techniques for the nystrom method. In *Artificial Intelligence and Statistics*, pages 304–311, 2009.

Junhong Lin and Volkan Cevher. Optimal rates of sketched-regularized algorithms for least-squares regression over Hilbert spaces. *arXiv preprint arXiv:1803.04371 (Proceedings of the 35th International Conference on Machine Learning)*, 2018a.

Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *arXiv preprint arXiv:1801.07226*, 2018b.

Junhong Lin and Lorenzo Rosasco. Optimal rates for learning with Nyström stochastic gradient methods. *arXiv preprint arXiv:1710.07797*, 2017a.

Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017b.

Junhong Lin, Alessandro Rudi, Lorenzo Rosasco, and Volkan Cevher. Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces. *Applied and Computational Harmonic Analysis*, 2018. URL `https://doi.org/10.1016/j.acha.2018.09.009`.

Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends in Machine Learning*, 3(2):123–224, 2011.

Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3): 277–289, 2008.

Stanislav Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *arXiv preprint arXiv:1112.5448*, 2011.

GL Myleiko, S Pereverzyev Jr, and SG Solodky. Regularized Nyström subsampling in regression and ranking problems under general smoothness assumptions. 2017.

Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

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

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

James O Ramsay. *Functional data analysis.* Wiley Online Library, 2006.

Alessandro Rudi, Raffaello Camoriano, and Lorenzo Rosasco. Less is more: Nystrom computational regularization. *arXiv preprint arXiv:1507.04717*, 2015.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

Bharath Sriperumbudur and Nicholas Sterge. Approximate kernel PCA using random features: Computational vs. statistical trade-off. *arXiv preprint arXiv:1706.06296*, 2017.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Ingo Steinwart, Don R Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *Conference On Learning Theory*, 2009.

Joel A Tropp. User-friendly tools for random matrices: An introduction. Technical report, DTIC Document, 2012.

Shusen Wang, Alex Gittens, and Michael W Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research*, 18(1):8039–8088, 2017.

Christopher KI Williams and Matthias Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, pages 661–667. MIT press, 2000.

Tianbao Yang, Yu-Feng Li, Mehrdad Mahdavi, Rong Jin, and Zhi-Hua Zhou. Nyström method vs random fourier features: A theoretical and empirical comparison. In *Advances in Neural Information Processing Systems*, pages 476–484, 2012.

Yun Yang, Mert Pilanci, and Martin J Wainwright. Randomized sketches for kernels: Fast and optimal nonparametric regression. *Annals of Statistics*, 45(3):991–1023, 2017.

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

Yuchen Zhang, John C Duchi, and Martin J Wainwright. Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. *Journal of Machine Learning Research*, 16:3299–3340, 2015.