Optimal Rates of Sketched-regularized Algorithms for Least-Squares Regression over Hilbert Spaces

Junhong Lin [1] **Volkan Cevher** [1]

Abstract

We investigate regularized algorithms combining with projection for least-squares regression problem over a Hilbert space, covering nonparametric regression over a reproducing kernel 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.

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,

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

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 (Cucker & Zhou, 2007; Steinwart & 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-regularized algorithm. Such algorithms can be viewed as solving an empirical, linear equation with the empirical covariance operator replaced by a regularized one, see (Caponnetto & Yao, 2006; Bauer et al., 2007; Gerfo et al., 2008; Lin et al., 2018) and references therein. Here, the regularization is used to control the complexity of the solution to against over-fitting and to achieve best generalization ability.

The function/estimator generated by classic regularized algorithm is in the subspace $\overline{span\{\mathbf{x}\}}$ 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. Such approaches have computational advantages in nonparametric regression with kernel methods (Williams & Seeger, 2000; Smola & Schölkopf, 2000). Typically, with a subsample/sketch dimension $m < n$, $S = \overline{span\{\tilde{x}_j : 1 \le j \le m\}}$ where $\tilde{x}_j$ is chosen randomly from the input set $\mathbf{x}$, or $S = \overline{span\{\sum_{j=1}^m G_{ij} x_j : 1 \le i \le m\}}$ where $\mathbf{G} = [G_{ij}]_{1 \le i \le m, 1 \le j \le n}$ is a general randomized matrix whose rows are drawn according to a distribution. The resulted algorithms are called Nyström regularized algorithm and sketched-regularized algorithm, respectively.

Our starting points of this paper are recent papers (Bach, 2013; Alaoui & Mahoney, 2015; Yang et al., 2015; Rudi et al., 2015; Myleiko et al., 2017) where convergence results on Nyström/sketched regularized algorithms for learning with kernel methods are given. 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 & Mahoney, 2015) for Nyström ridge regression and in (Yang et al., 2015) 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 & 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

[1] Laboratory for Information and Inference Systems, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. Correspondence to: Junhong Lin <junhong.lin@epfl.ch>, Volkan Cevher <volkan.cevher@epfl.ch>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

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 algorithms 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 case that 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 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 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 optimal rates 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 are the first ones with 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 our proof, we naturally integrate proof techniques from (Smale & Zhou, 2007; Caponnetto & De Vito, 2007; Rudi et al., 2015; Myleiko et al., 2017; Lin et al., 2018). Our novelties lie in a new estimates 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 4 for details.

The rest of the paper is organized as follows. Section 2 introduces some auxiliary notations and projected-regularized algorithms. Section 3 present assumptions and our main results, followed with simple discussions. Finally, Section 4 gives the proofs of our main results.

2. Learning with Projected-regularized Algorithms

In this section, we introduce some notations as well as auxiliary operators, and present projected-regularized algorithms.

2.1. Notations and Auxiliary Operators

Let $Z = H \times \mathbb{R}$, $\rho_X(\cdot)$ the induced marginal measure on H of ρ, and $\rho(\cdot|x)$ 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

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

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 $\|f\|_\rho = \left(\int_H |f(x)|^2 d\rho_X\right)^{\frac{1}{2}}$.

For a given bounded operator $L : L^2_{\rho_X} \to H$, $\|L\|$ denotes the operator norm of L, i.e., $\|L\| = \sup_{f \in L^2_{\rho_X}, \|f\|_\rho = 1} \|Lf\|_H$. Let $r \in \mathbb{N}_+$, 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):

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

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

$$\|\mathcal{S}_\rho \omega\|_\rho = \|\sqrt{\mathcal{T}} \omega\|_H. \tag{4}$$

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

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

We define the sampling operator $\mathcal{S}_\mathbf{x} : H \to \mathbb{R}^n$ by $(\mathcal{S}_\mathbf{x} \omega)_i = \langle \omega, x_i \rangle_H$, $i \in [n]$, where the norm $\|\cdot\|_{\mathbb{R}^n}$ in $\mathbb{R}^n$ is the Euclidean norm times $1/\sqrt{n}$. Its adjoint operator $\mathcal{S}^*_\mathbf{x} : \mathbb{R}^n \to H$, defined by $\langle \mathcal{S}^*_\mathbf{x} \mathbf{y}, \omega \rangle_H = \langle \mathbf{y}, \mathcal{S}_\mathbf{x} \omega \rangle_{\mathbb{R}^n}$ for $\mathbf{y} \in \mathbb{R}^n$ is thus given by $\mathcal{S}^*_\mathbf{x} \mathbf{y} = \frac{1}{n} \sum_{i=1}^n y_i x_i$. Moreover, we can define the empirical covariance operator $\mathcal{T}_\mathbf{x} : H \to H$ such that $\mathcal{T}_\mathbf{x} = \mathcal{S}^*_\mathbf{x} \mathcal{S}_\mathbf{x}$. Obviously, $\mathcal{T}_\mathbf{x} = \frac{1}{n} \sum_{i=1}^n \langle \cdot, x_i \rangle_H x_i$. By Assumption (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 & Zhou, 2007; Steinwart & Christmann, 2008), defined as,

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

A simple calculation shows that the following well-known fact holds (Cucker & Zhou, 2007; Steinwart & Christmann, 2008), for all $f \in L^2_{\rho_X}$, $\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_\rho$. Then it is easy to see that (7) is equivalent to $\inf_{f\in H_\rho} \|f - f_\rho\|^2_\rho$. 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 & Rosasco, 2017),

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

and

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

Note that f_H does not necessarily be in H_ρ, as indicated by a simple example constructed in the appendix.
Throughput this paper, S is a closed, finite-dimensional subspace of H, and P is the projection operator onto S.

2.2. Projected-regularized Algorithms

In this subsection, we demonstrate and introduce 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 $\widetilde{\mathcal{E}}_{\mathbf{z}}(\omega)$, $\widetilde{\mathcal{E}}_{\mathbf{z}}(\omega) = \frac{1}{n}\sum_{i=1}^n (\langle \omega, x_i\rangle_H - y_i)^2$. 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 is given by $\hat{\omega} = P\hat{\alpha}$, with $\hat{\alpha}$ satisfying $P\mathcal{T}_{\mathbf{x}}P\hat{\alpha} = P\mathcal{S}^*_{\mathbf{x}}\mathbf{y}$. Motivated by the classic (iterated) ridge regression, we replace $P\mathcal{T}_{\mathbf{x}}P$ with a regularized one, and thus leads to the following projected (iterated) ridge regression.

Algorithm 1. *The projected (iterated) ridge regression algorithm of order τ over the samples $\mathbf{z}$ and the subspace S is given by $f^{\mathbf{z}}_\lambda = \mathcal{S}_\rho \omega^{\mathbf{z}}_\lambda$, where* [1]

$$\omega^{\mathbf{z}}_\lambda = P\mathcal{G}_\lambda(P\mathcal{T}_{\mathbf{x}}P)P\mathcal{S}^*_{\mathbf{x}}\mathbf{y}, \;\; \mathcal{G}_\lambda(u) = \sum_{i=1}^{\tau} \lambda^{i-1}(\lambda + u)^{-i}. \tag{11}$$

Remark 1. *1) In this paper, we focus on projected ridge regression, but all the derived results 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*

$$\sup_{\lambda\in\Lambda} \sup_{u\in]0,\kappa^2]} |\mathcal{G}_\lambda(u)(u + \lambda)| \leq E. \tag{12}$$

and

$$\sup_{\alpha\in[0,\tau]} \sup_{\lambda\in\Lambda} \sup_{u\in]0,\kappa^2]} |1 - \mathcal{G}_\lambda(u)u|(u + \lambda)^\alpha \lambda^{-\alpha} \leq F. \tag{13}$$

2) A simple calculation shows that

$$\mathcal{G}_\lambda(u) = \frac{1 - q^\tau}{u} = \frac{\sum_{i=0}^{\tau-1} q^i}{u + \lambda}, \quad q = \frac{\lambda}{\lambda + u}. \tag{14}$$

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 classic ridge regression and the algorithm is projected ridge regression.
3) Another typical filter function studied in the literature is $\mathcal{G}_\lambda(u) = u^{-1}1_{\{u\geq\lambda\}}$, 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 number.

In the above, λ is a regularization parameter which needs to be well chosen in order to achieve best performance. Throughout this paper, we assume that $1/n \leq \lambda \leq 1$.
The performance of an estimator $f^{\mathbf{z}}_\lambda$ can be measured in terms of *excess risk* (*generalization error*), $\mathcal{E}(f^{\mathbf{z}}_\lambda) - \inf_{H_\rho} \mathcal{E} = \widetilde{\mathcal{E}}(\omega^{\mathbf{z}}_\lambda) - \inf_H \widetilde{\mathcal{E}}$, which is exactly $\|f^{\mathbf{z}}_\lambda - f_H\|^2_\rho$ according to (10). Assuming that $f_H \in H_\rho$, i.e., $f_H = \mathcal{S}_\rho\omega_*$ for some $\omega_* \in H$ (in this case, the solution with minimal H-norm for $f_H = \mathcal{S}_\rho\omega$ is denoted by ω_H), it can be measured in terms of H-norm, $\|\omega^{\mathbf{z}}_\lambda - \omega_H\|_H$, which is closely related to $\|\mathcal{L}^{-\frac{1}{2}}\mathcal{S}_\rho(\omega^{\mathbf{z}}_\lambda - \omega_H)\|_H = \|\mathcal{L}^{-\frac{1}{2}}(f^{\mathbf{z}}_\lambda - f_H)\|_\rho$, according to (5). In what follows, we will measure the performance of an estimator $f^{\mathbf{z}}_\lambda$ in terms of a broader class of norms, $\|\mathcal{L}^{-a}(f^{\mathbf{z}}_\lambda - f_H)\|_\rho$, where $a \in [0, \frac{1}{2}]$ is such that $\mathcal{L}^{-a}f_H$ is well defined. But one should keep in mind that all the derived results also hold if we replace $\|\mathcal{L}^{-a}(f^{\mathbf{z}}_\lambda - f_H)\|_\rho$ with $\|\mathcal{T}^{\frac{1}{2}-a}(\omega^{\mathbf{z}}_\lambda - \omega_H)\|_H$ in the attainable case, i.e., $f_H \in H_\rho$. We will report these results in a longer version of this paper. Convergence with respect to different norms has its strong backgrounds in convex optimization, inverse problems, and statistical learning theory. Particularly, convergence with respect to target function values and H-norm has been studied in convex optimization. Interestingly, convergence in H-norm can imply convergence in target function values (although the derived rate is not optimal), while the opposite is not true.

3. Convergence Results

In this section, we first introduce some basic assumptions and then present convergence results for projected-regularized algorithms. Finally, we give results for sketched/Nyström regularized algorithms.

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

3.1. Assumptions

In this subsection, we introduce three standard assumptions made in statistical learning theory (Steinwart & Christmann, 2008; Cucker & Zhou, 2007; Lin et al., 2018). The first assumption relates to a moment condition on the output value y.

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

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

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. Condition (15) implies that the regression function is bounded almost surely, using the Cauchy-Schwarz inequality.

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

Assumption 2. *f_H satisfies*

$$\int_H (f_H(x) - f_\rho(x))^2 x \otimes x d\rho_X(x) \preceq B^2 \mathcal{T}, \tag{16}$$

and the following Hölder source condition

$$f_H = \mathcal{L}^\zeta g_0, \quad \textit{with} \quad \|g_0\|_\rho \leq R. \tag{17}$$

Here, B, R, ζ are non-negative numbers.

Condition (16) 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 & Zhou, 2007; Caponnetto, 2006; Caponnetto & De Vito, 2007; Steinwart et al., 2009), for kernel-based non-parametric regression, it is satisfied with $B = 0$. Condition (17) characterizes the regularity of the target function f_H (Smale & 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 & Christmann, 2008). This means that the expected risk minimization (1) has at least one solution in H, which is referred to as the attainable case.

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*

$$\mathrm{tr}(\mathcal{T}(\mathcal{T} + \lambda I)^{-1}) \leq c_\gamma \lambda^{-\gamma}, \quad \textit{for all } \lambda > 0. \tag{18}$$

The left hand-side of (18) is called degrees of freedom (Zhang, 2005), or effective dimension (Caponnetto & 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 as the capacity independent setting. Assumption 3 with $\gamma \in [0, 1]$ allows to derive better rates. It is satisfied, e.g., if the eigenvalues of $\mathcal{T}$ satisfy a polynomial decaying condition $\sigma_i \sim i^{-1/\gamma}$, or with $\gamma = 0$ if $\mathcal{T}$ is finite rank.

3.2. Results for Projected-regularized Algorithms

We are now ready to state our first result as follows. Throughout this paper, C denotes a positive constant that depends only on $\kappa^2, c_\gamma, \gamma, \zeta\ B, M, Q, R, \tau$ 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 $\lambda = n^{\theta - 1}$ for some $\theta \in [0, 1]$, $\tau \geq \zeta$, and $a \in [0, \frac{1}{2} \wedge \zeta]$. Then the following holds with probability at least $1 - \delta$ ($0 < \delta < 1$). 1) If $\zeta \in [0, 1]$,*

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim \lambda^{-a} \log^2 \frac{3}{\delta} t_{\theta,n}^{1-a} \times \left(\lambda^\zeta + \frac{1}{\sqrt{n\lambda^\gamma}} + \lambda^{\zeta - 1} \left(\Delta_5 + \Delta_5^{1-a} \lambda^a \right) \right). \tag{19}$$

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

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

Here, Δ_5 is the projection error $\|(I - P)\mathcal{T}^{\frac{1}{2}}\|^2$ and

$$t_{\theta,n} = [1 \vee (\theta^{-1} \wedge \log n^\gamma)]. \tag{21}$$

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 bias. The second term depends on the sample size, and it arises from estimating variance. The third term depends on the projection error. Note that there is a trade-off between the bias and variance terms. Ignoring the projection error, solving this trade-off leads to the best choice on λ and the following results.

Corollary 2. *Under the assumptions and notations of Theorem 1, let $\lambda = n^{-\frac{1}{1 \vee (2\zeta + \gamma)}}$. Then the following holds with probability at least $1 - \delta$.*
1) If $2\zeta + \gamma \leq 1$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim n^{-(\zeta - a)} \left(1 + (\gamma \log n)^{1-a}\right)(1 + \lambda^{-1}\Delta_5) \log^2 \frac{3}{\delta}. \tag{22}$$

2) If $\zeta \in [0,1]$ *and* $2\zeta+\gamma>1$,

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

3) If $\zeta \geq 1$,

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

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 3. *Under the assumptions and notations of Theorem 1, let* $\lambda = n^{-\frac{1}{1\vee(2\zeta+\gamma)}}$ *and* $P=I$. *Then with probability at least* $1-\delta$,

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

The above result recovers the result derived in (Lin et al., 2018). The convergence rates are optimal as they match the mini-max rates with $\zeta \geq 1/2$ derived in (Caponnetto & De Vito, 2007; Blanchard & Mucke, 2016).

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 $\overline{range\{\mathcal{S}_{\mathbf{x}}^*\mathbf{G}^*\}}$, where $\mathbf{G}\in\mathbb{C}^{m\times n}$ is a sketch matrix whose rows are i.i.d drawn according to a distribution F. In this paper, we assume the distribution F satisfies the following two properties.

- **Isotropy property**: We say that F obeys the isotropy property if
$$\mathbb{E}[\mathbf{a}\mathbf{a}^*]=I, \quad \mathbf{a}\sim F. \tag{26}$$
- **Bounded property:** We assume that the random vector $\mathbf{a}\sim F$ is bounded almost surely: for some $\mu>0$,
$$\|\mathbf{a}\|_2 \leq \sqrt{n}\mu. \tag{27}$$

Examples for the above sketch mechanics include *subsampled orthogonal systems* (OS), *subsampled tight or continuous frames*, and *random convolutions*, etc, see (Candes & Plan, 2011) for further details. In this paper, we focus on OS sketches, which are based on randomly sampling the rows of a fixed orthonormal matrix $K\in\mathbb{R}^{n\times n}$. Examples of such matrices include the discrete Fourier transform (DFT) matrix, and the Hadamard matrix. 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.

Conditions (27) implies that $\mu\geq 1$, due to the isotropy property. Without loss of generality, we assume that $\mu=1$ throughout.

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

Corollary 4. *Under the assumptions of Theorem 1, let* $S=\overline{range\{\mathcal{S}_{\mathbf{x}}^*\mathbf{G}^*\}}$, *where* $\mathbf{G}\in\mathbb{C}^{m\times n}$ *is a randomized matrix whose rows are i.i.d drawn from the distribution* F. *Let* $\lambda = n^{-\frac{1}{1\vee(2\zeta+\gamma)}}$ *and*

$$m \gtrsim \begin{cases} n^\gamma\log^2\frac{3n^\gamma}{\delta}\log\frac{3}{\delta} & \text{if } 2\zeta+\gamma\leq 1,\\ n^{\frac{\gamma(\zeta-a)}{(1-a)(2\zeta+\gamma)}}\log\frac{3n^\gamma}{\delta}\log^2\frac{3}{\delta} & \text{if } \zeta\geq 1,\\ n^{\frac{\gamma}{2\zeta+\gamma}}\log\frac{3n^\gamma}{\delta}\log^2\frac{3}{\delta} & \text{otherwise}.\end{cases} \tag{28}$$

Then with confidence at least $1-\delta$, *for* $\zeta\leq 1$, *or* $\zeta>1$ *and* $a\leq\gamma/(2\zeta+\gamma-2)$, *the following holds*

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}}-f_H)\|_\rho \lesssim \log^3\frac{3}{\delta}\begin{cases} n^{-(\zeta-a)}\left(1+(\gamma\log n)^{2(1-a)}\right), & \text{if } 2\zeta+\gamma\leq 1,\\ n^{-\frac{\zeta-a}{2\zeta+\gamma}}, & \text{if } 2\zeta+\gamma>1.\end{cases} \tag{29}$$

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. Note that the minimal sketch dimension from the above is proportional to the effective dimension $\lambda^{-\gamma}$ up to a logarithmic factor for the case $\zeta\leq 1$.

Remark 2. *1) The bounded assumption* (27) *may be replaced with a high-probability bounded assumption as that in (Candes & Plan, 2011), which is satisfied for Gaussian sketches.*

2) Considering only the case $\zeta=1/2$ *and* $a=0$, *(Yang et al., 2015) provides optimal error bounds for sketched ridge regression within the fixed design setting.*

3) Letting $\zeta=1/2$, *the minimal sketch dimension from the above is smaller than* $O(n^{\frac{\gamma}{\gamma+1}}\log^4 n)$ *from (Yang et al., 2015) using OS sketches.*

3.4. Results for Nyström Regularized Algorithms

As a byproduct of the paper, using Corollary 2, we derive the following results for Nyström regularized algorithms.

Corollary 5. *Under the assumptions of Theorem 1, let* $S = \overline{span\{x_1, \cdots, x_m\}}$, $2\zeta + \gamma > 1$, *and* $\lambda = n^{-\frac{1}{2\zeta+\gamma}}$. *Then with probability at least* $1 - \delta$,

$$\|\mathcal{L}^{-a}(f_\lambda^{\mathbf{z}} - f_H)\|_\rho \lesssim n^{-\frac{\zeta-a}{2\zeta+\gamma}} \log^3 \frac{3}{\delta},$$

provided that

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

Remark 3. *1) In the above, we only consider the plain Nyström subsampling. Using the ALS Nyström subsampling (Drineas et al., 2012; Gittens & Mahoney, 2013; Alaoui & Mahoney, 2015) and the proof technique developed in this paper and (Rudi et al., 2015), we can further improve the projection dimension condition to* (28) *(possibly with an extra* $\log n$*). We will report this result in a longer version of this paper.*
2) 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* $\zeta \geq 1$.
3) 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).*

All the results stated in this section will be proved in the next section.

4. Proof

In this section, we prove the results stated in Section 3. We first give some deterministic estimates and an analytics result. We then give some probabilistic estimates. Applying the probabilistic estimates into the analytics result, we prove the results for projected-regularized algorithms. We finally estimate the projection errors and present the proof for sketched-regularized algorithms.

4.1. Deterministic Estimates

In this subsection, we introduce some deterministic estimates. For notational simplicity, throughout this paper, we denote

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

We define a deterministic vector ω_λ as follows,

$$\omega_\lambda = \mathcal{G}_\lambda(\mathcal{T})\mathcal{S}_\rho^* f_H. \tag{30}$$

The vector ω_λ is often called population function. We introduce the following lemma. The proof is essentially the same as that for Lemma 26 from (Lin & Cevher, 2018). We thus omit it.

Lemma 6. *Under Assumption 2, the following holds.*
1) For any $\zeta - \tau \leq a \leq \zeta$,

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda - f_H)\|_\rho \leq R\lambda^{\zeta-a}. \tag{31}$$

2)

$$\|\mathcal{T}^{a-1/2}\omega_\lambda\|_H \leq \tau R \cdot \begin{cases} \lambda^{\zeta+a-1}, & \text{if } -\zeta \leq a \leq 1-\zeta, \\ \kappa^{2(\zeta+a-1)}, & \text{if } a \geq 1-\zeta. \end{cases} \tag{32}$$

The above lemma provides some basic properties for the population function. It will be useful for the proof of our main results. The left hand-side of (31) is often called true bias.
Using the above lemma and some basic operator inequalities, we can prove the following analytic, deterministic result.

Proposition 7. *Under Assumption 2, let*

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

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

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

$$\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T} - \mathcal{T}_{\mathbf{x}})\| \leq \Delta_4,$$

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

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

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq \tau\lambda^{-a}\Delta_1^{1-a} \times \Big(\Delta_2 + 2(\tau+1)R\lambda^\zeta + \tau R\lambda^{\zeta-1}(\Delta_5 + \Delta_5^{1-a}\lambda^a)\Big). \tag{33}$$

2) If $\zeta \geq 1$,

$$\|\mathcal{L}^{-a}(\mathcal{S}_\rho\omega_\lambda^{\mathbf{z}} - f_H)\|_\rho \leq \tau\lambda^{-a}\Delta_1^{1-a} \times \Big(\Delta_2 + 3R\lambda^\zeta + \kappa^{2(\zeta-1)}R\big(\kappa\tau\Delta_4 + \tau\Delta_5 + \lambda(\Delta_3 + \Delta_5)^{(\zeta-1)\wedge 1} + \lambda^{\frac{1}{2}}\Delta_3^{(\zeta-\frac{1}{2})\wedge 1} + \Delta_5^{1-a}\lambda^a\big)\Big). \tag{34}$$

The above proposition is key to our proof. The proof of the above proposition for the case $\zeta \leq 1$ borrows ideas from (Smale & Zhou, 2007; Caponnetto & De Vito, 2007; Rudi et al., 2015; Myleiko et al., 2017; Lin et al., 2018), whereas the key step is an error decomposition from (Lin & Cevher, 2018). Our novelty lies in the proof for the case $\zeta \geq 1$, see the appendix for further details.

4.2. Proof for Projected-regularized Algorithms

To derive total error bounds from Proposition 7, it is necessary to develop probabilistic estimates for the random quantities Δ_1, Δ_2, Δ_3 and Δ_4. We thus introduce the following four lemmas.

Lemma 8. *(Lin et al., 2018) Under Assumption 3, let $\delta \in (0,1)$, $\lambda = n^{-\theta}$ for some $\theta \geq 0$, and*

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

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

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

and

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

Lemma 9. *Let $0 < \delta < 1/2$. It holds with probability at least $1-\delta$:*

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

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

Lemma 10. *Under Assumption 3, let $0 < \delta < 1/2$. It holds with probability at least $1-\delta$:*

$$\|\mathcal{T}_\lambda^{-\frac{1}{2}}(\mathcal{T} - \mathcal{T}_{\mathbf{x}})\|_{HS} \leq 2\kappa \left(\frac{2\kappa}{n\sqrt{\lambda}} + \sqrt{\frac{c_\gamma}{n\lambda^\gamma}} \right) \log \frac{2}{\delta}.$$

The proof of the above lemmas can be done simply applying concentration inequalities for sums of Hilbert-space-valued random variables. We refer to (Lin & Rosasco, 2017) for the proofs.

Lemma 11. *(Lin et al., 2018) Under Assumptions 1, 2 and 3, let ω_λ be given by (30). For all $\delta \in]0, 1/2[$, the following holds with probability at least $1-\delta$:*

$$\|\mathcal{T}_\lambda^{-1/2}[(\mathcal{T}_{\mathbf{x}}\omega_\lambda - \mathcal{S}_{\mathbf{x}}^*\mathbf{y}) - (\mathcal{T}\omega_\lambda - \mathcal{S}_\rho^* f_\rho)]\|_H \leq \left(\frac{C_1}{n\lambda^{\frac{1}{2}\vee(1-\zeta)}} + \sqrt{\frac{C_2\lambda^{2\zeta}}{n\lambda} + \frac{C_3}{n\lambda^\gamma}} \right) \log \frac{2}{\delta}. \tag{36}$$

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

With the above probabilistic estimates and the analytics result, Proposition 7, we are now ready prove results for projected-regularized algorithms.

Proof of Theorem 1. We use Proposition 7 to prove the result. We thus need to estimate Δ_1, Δ_2, Δ_3 and Δ_4. Following from Lemmas 8, 9, 10 and 11, with $n^{-1} \leq \lambda \leq 1$, we know that with probability at least $1-\delta$,

$$\Delta_1 \lesssim t_{\theta,n} \log \frac{3}{\delta}, \tag{37}$$

$$\Delta_2 \lesssim \left(\frac{1}{n\lambda^{\frac{1}{2}\vee(1-\zeta)}} + \lambda^\zeta + \frac{1}{\sqrt{n\lambda^\gamma}} \right) \log \frac{3}{\delta},$$

$$\Delta_3 \lesssim \frac{1}{\sqrt{n}} \log \frac{3}{\delta}, \tag{38}$$

$$\Delta_4 \lesssim \frac{1}{\sqrt{n\lambda^\gamma}} \log \frac{3}{\delta}.$$

The results thus follow by introducing the above estimates into (33) or (34), combining with a direct calculation and $1/n \leq \lambda \leq 1$. □

4.3. Proof for Sketched-regularized Algorithms

In order to use Corollary 2 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. We begin with the following concentration inequalities.

Lemma 12. *Let $0 < \delta < 1$ and $\lambda > 0$. For any given $\mathbf{x} \subseteq H^n$, there exists a subset $U_{\mathbf{x}}$ of $\mathbb{R}^{m\times n}$ with measure at least $1-\delta$, such that for all $\mathbf{G} \in U_{\mathbf{x}}$,*

$$\left\| (\mathcal{T}_{\mathbf{x}}+\lambda)^{-1/2}(\mathcal{T}_{\mathbf{x}} - m^{-1}\mathcal{S}_{\mathbf{x}}^*\mathbf{G}^*\mathbf{G}\mathcal{S}_{\mathbf{x}})(\mathcal{T}_{\mathbf{x}}+\lambda)^{-1/2} \right\| \leq \frac{4\mathcal{N}_{\mathbf{x}}(\lambda)\beta}{3m} + \sqrt{\frac{2\mathcal{N}_{\mathbf{x}}(\lambda)\beta}{m}},$$

where

$$\mathcal{N}_{\mathbf{x}}(\lambda) = \mathrm{tr}((\mathcal{T}_{\mathbf{x}}+\lambda)^{-1}\mathcal{T}_{\mathbf{x}}),$$
$$\beta = \log \frac{4\mathcal{N}_{\mathbf{x}}(\lambda)(1+\lambda/\|\mathcal{T}_{\mathbf{x}}\|)}{\delta}.$$

The above lemma can be proved using the concentration inequalities from (Tropp, 2012; Minsker, 2011). With the above lemma and Lemma 8, we can estimate $\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2$ as follows.

Lemma 13. *Let $0 < \delta < 1$ and $\theta \in [0,1]$. Given a fix $\mathbf{x} \in H^n$, assume that for $\lambda = n^{-\theta}$,*

$$\mathrm{tr}((\mathcal{T}_{\mathbf{x}}+\lambda)^{-1}\mathcal{T}_{\mathbf{x}}) \leq b_\gamma \lambda^{-\gamma} \tag{39}$$

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 $\mathbf{G} \in U_{\mathbf{x}}$,

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

provided that

$$m \geq 8b^\gamma n^{\theta\gamma}\log\frac{8b_\gamma n^{\theta\gamma}}{\delta}. \tag{40}$$

Under the condition (39), Lemma 13 provides an upper bound for $\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|$, which will be used to control the projection error using the following lemma.

Lemma 14. *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*

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

The left-hand side of (39) is called empirical effective dimension. It can be estimated as follows.

Lemma 15. *Under Assumption 3, let $\lambda = n^{-\theta}$ for some $\theta \in [0,1]$ and $0<\delta<1$. With confidence $1-\delta$,*

$$\begin{aligned}&\mathrm{tr}((\mathcal{T}_{\mathbf{x}}+\lambda)^{-1}\mathcal{T}_{\mathbf{x}})\\ &\leq 3(4\kappa^2+2\kappa\sqrt{c_\gamma}+c_\gamma)\log\frac{4}{\delta}a_{n,\delta/2,\gamma}(\theta)\lambda^{-\gamma},\end{aligned} \tag{41}$$

where $a_{n,\delta/2,\gamma}(\theta)$ is given as in Lemma 8.

The above lemma improves Proposition 1 of (Rudi et al., 2015). 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 and give the proof for sketched-regularized algorithms.

Proof of Corollary 4. Let $\lambda' = n^{-\theta'}$, with

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

Following from Corollary 2, Lemmas 8, 9 and 15, we know that there exists a subset V of Z^n with measure at least $1-4\delta$, such that for all $\mathbf{z} \in V$, (22) (or (23), or (24)), (37), (38), and (41) (with θ and λ replaced by θ' and λ' in (41), respectively) hold.
For any $\mathbf{z} \in V$, using Lemma 13 with

$$\begin{aligned}b_\gamma =&3(4\kappa^2+2\kappa\sqrt{c_\gamma}+c_\gamma)\log\frac{4}{\delta}a_{n,\delta/2,\gamma}(\theta')\\ \lesssim&(1\vee[(1-\theta')^{-1}\wedge\log n^\gamma]+\log\frac{3}{\delta})\log\frac{3}{\delta},\end{aligned}$$

we know that there exists a subset $U_{\mathbf{z}}$ of $\mathbb{R}^{m\times n}$ with measure at least $1-\delta$, such that for all $\mathbf{G} \in U_{\mathbf{z}}$,

$$\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 \lesssim \frac{1}{n^{\theta'}}, \tag{42}$$

provided $m \gtrsim n^{\theta'\gamma}b_\gamma\log\frac{3b_\gamma n^{\theta'\gamma}}{\delta}$, which is guaranteed by Condition (28). Using $\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2 = \|\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|$, and Lemma 14,

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

Introducing with (38), and (42), and noting that $\|\mathcal{T}^{\frac{1}{2}}(I-P)\mathcal{T}^{\frac{1}{2}}\| = \|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2$, we get

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \lesssim \frac{\log\frac{3}{\delta}}{\sqrt{n}} + \frac{1}{n^{\theta'}}.$$

Introduce the above into (24), one can prove the desired results for the case $\zeta \geq 1$.
Now consider the case $\zeta \leq 1$. Note that

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

Introducing with (37) and using a similar argument as that for (50),

$$\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \lesssim (\|(I-P)\mathcal{T}_{\mathbf{x}}^{\frac{1}{2}}\|^2+\lambda)t_{\theta,n}\log\frac{3}{\delta}.$$

Applying (42), $\|(I-P)\mathcal{T}^{\frac{1}{2}}\|^2 \lesssim (\frac{1}{n^{\theta'}}+\lambda)t_{\theta,n}\log\frac{3}{\delta}$. Introducing the above into (22), or (23), one can prove the desired results for the case $\zeta \leq 1$. □

The proof of Corollary 5 will be given in the appendix due to space limitation.

5. Conclusion

In this paper, we prove optimal statistical results with respect to variants of norms for sketched/Nyström regularized algorithms. Our contributions are mainly on theoretical aspects. First, our results for sketched-regularized algorithms generalize previous results (Yang et al., 2015) 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, which have not been studied before for both Nyström and sketched regularized algorithms. Third, our results provide the first 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 $\zeta = 1$. A future direction is to extend our analysis to learning with random features, see (Sriperumbudur & Sterge, 2017; Lin & Rosasco, 2018) and 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), and from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (grant agreement n 725594-time-data).

References

Alaoui, Ahmed and Mahoney, Michael W. Fast randomized kernel ridge regression with statistical guarantees. pp. 775–783, 2015.

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

Bach, Francis. On the equivalence between kernel quadrature rules and random feature expansions. *Arxiv*, 2015.

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

Blanchard, Gilles and Mucke, Nicole. Optimal rates for regularization of statistical inverse learning problems. *arXiv preprint arXiv:1604.04054*, 2016.

Candes, Emmanuel J and Plan, Yaniv. A probabilistic and RIPless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011.

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

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

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

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

Dicker, Lee H, Foster, Dean P, and Hsu, Daniel. Kernel ridge vs. principal component regression: Minimax bounds and the qualification of regularization operators. 2016.

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

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

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

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

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

Gittens, Alex and Mahoney, Michael W. Revisiting the nystrom method for improved large-scale machine learning. *arXiv preprint arXiv:1303.1849*, 2013.

Hansen, Frank. An operator inequality. *Mathematische Annalen*, 246(3):249–250, 1980.

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

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

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

Lin, Junhong and Rosasco, Lorenzo. Generalization properties of doubly stochastic learning algorithms. *Journal of Complexity*, 47:42–61, 2018.

Lin, Junhong, Rudi, Alessandro, Rosasco, Lorenzo, and Cevher, Volkan. Optimal rates for spectral-regularized algorithms with least-squares regression over hilbert spaces. *arXiv preprint arXiv:1801.06720*, 2018.

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

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

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

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

Rudi, Alessandro, Camoriano, Raffaello, and Rosasco, Lorenzo. Less is more: Nyström computational regularization. *Advances in Neural Information Processing Systems*, pp. 1657–1665, 2015.

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

Smola, Alex J and Schölkopf, Bernhard. Sparse greedy matrix approximation for machine learning. 2000.

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

Steinwart, Ingo and Christmann, Andreas. *Support Vector Machines*. Springer Science & Business Media, 2008.

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

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

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

Yang, Yun, Pilanci, Mert, and Wainwright, Martin J. Randomized sketches for kernels: Fast and optimal nonparametric regression. *arXiv preprint arXiv:1501.06195*, 2015.

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

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