Appl. Comput. Harmon. Anal. ••• (••••) •••–•••

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

www.elsevier.com/locate/acha

Distributed regularized least squares with flexible Gaussian kernels ☆

Ting Hu [a], Ding-Xuan Zhou [b]

[a] *School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*
[b] *School of Data Science and Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong*

ARTICLE INFO

Article history:
Received 26 December 2019
Received in revised form 22 November 2020
Accepted 22 March 2021
Available online xxxx
Communicated by Naoki Saito

Keywords:
Reproducing kernel Hilbert space
Flexible Gaussian kernels
Distributed learning
Semi-supervised learning
Sobolev space

ABSTRACT

We propose a distributed learning algorithm for least squares regression in reproducing kernel Hilbert spaces (RKHSs) generated by flexible Gaussian kernels, based on a divide-and-conquer strategy. Our study demonstrates that Gaussian kernels with flexible variances greatly improve the learning performance of distributed algorithms generated by a fixed Gaussian. Under some mild conditions, we establish sharp error bounds for the distributed algorithm with labeled data in which the variance of the Gaussian kernel serves as a tuning parameter. We show that with suitably chosen parameters our error rates can be almost mini-max optimal under the standard Sobolev smoothness condition on the target function. By utilizing additional information of unlabeled data for semi-supervised learning, we relax the restrictions on the number of data partition and the range of the Sobolev smoothness index.

1. Introduction

Distributed learning algorithms are commonly used in engineering and scientific communities for their capabilities to handle massive data. A divide-and-conquer strategy is a widely used approach in various distributed learning paradigms due to its simplicity and scalability. The basic procedure of the strategy begins with randomly partitioning the whole data set into disjoint subsets of equal size, produces a series of estimators by a base learning algorithm with each subset, and then averages the individual solutions together to get a global output. Its applications and theoretical analysis were investigated in a broad range of learning problems, such as classification [8], matrix factorization [13], perceptron [14] and conditional maximum entropy models [15]. Recently, the divide-and-conquer approach applied to kernel-based algorithms has been developed in many machine learning tasks [7,9–12,17,31]. The asymptotic behaviors of the averaged

☆ The work described in this paper was partially supported by National Natural Science Foundation of China (Projects 12071356), the Research Grants Council of Hong Kong (Project No. CityU 11306617) and Hong Kong Institute for Data Science.

E-mail addresses: tinghu@whu.edu.cn (T. Hu), mazhou@cityu.edu.hk (D.-X. Zhou).

https://doi.org/10.1016/j.acha.2021.03.008
1063-5203/

estimators were considered and some consistency results were derived. It was shown that the averaged estimators can retain mini-max optimal rates over the base algorithm working with the whole data set.

Gaussians are the most important and commonly used kernels in the design of kernel-based algorithms. The variances of Gaussians characterize the frequency range of the key function features and appropriate values of variances are essential to the learning power of the algorithms. However, distributed learning for regression with Gaussian kernels has not been fully considered in the existing literature. It remains a lack of theoretical understanding on the role of Gaussian variances in guaranteeing the effectiveness of distributed learning. Note that the RKHS induced by a single Gaussian kernel has a low capacity and allowing Gaussians with flexible variances improves learning abilities in terms of regularization error and approximation error. In the case of flexible Gaussian kernels, the variance of Gaussian is not a given constant and can be chosen to depend on the sample size, associated with smoothness conditions on target functions, the intrinsic dimension or some other priori information on the learning problems. That is one advantage of flexible Gaussians used in learning algorithms, see [27–29,22] and the references therein. In this paper, based on the divide-and-conquer strategy, we study the distributed regularized least squares with flexible Gaussian kernels. In our work, the variance of a Gaussian is not fixed and changes according to the sample size. It is therefore of great interest to investigate how to employ suitable variances for keeping the effective learning performance of distributed learning when the data size grows rapidly.

Let $\mathcal{X}$ denote an input space which is assumed to be a compact subset of $\mathbb{R}^d$, an output space $\mathcal{Y} \subset \mathbb{R}$ be a set of real numbers, ρ be an underlying Borel probability measure on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. The Gaussian kernel with standard deviation $\sigma > 0$ is the function on $\mathcal{X} \times \mathcal{X}$ given by

$$K_\sigma(x,u) := \exp\left\{-\frac{|x-u|^2}{\sigma^2}\right\}.$$

With K_σ, the reproducing kernel Hilbert space $\mathcal{H}_\sigma$ is induced by the completion of the linear span of the set of functions $\{K_\sigma(x,\cdot) : x \in \mathcal{X}\}$ with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}_\sigma}$. It has the *reproducing property*, that is, for any $f \in \mathcal{H}_\sigma$,

$$\langle f, K_\sigma(x,\cdot)\rangle_{\mathcal{H}_\sigma} = f(x), \quad x \in \mathcal{X}. \tag{1}$$

Given a data set $D = \{(x_i,y_i)\}_{i=1}^N \subset \mathcal{Z}$, the *regularized least squares* with the Gaussian RKHS $\mathcal{H}_\sigma$ can be stated as

$$f_D := f_{D,\lambda,\sigma} = \arg\min_{f\in\mathcal{H}_\sigma}\left\{\frac{1}{|D|}\sum_{(x,y)\in D}(f(x)-y)^2 + \lambda\|f\|^2_{\mathcal{H}_\sigma}\right\}. \tag{2}$$

Here $\lambda > 0$ is a regularization parameter and $|D| := N$ denotes the cardinality of D. When the data set D has a very large size N, we can apply the divide-and-conquer strategy to this scheme. Suppose that the whole data set D consists of m disjoint subsets $\{D_l\}_l$ of equal size, that is, $D = \bigcup_{l=1}^m D_l$, $|D_1| = \cdots, = |D_m| := n$ and $N = mn$. *Distributed learning with the regularized least squares and Gaussian kernels* considered in this paper takes the form of a weighted average of the local estimators $\{f_{D_l}\}_l$ as

$$\bar{f}_D = \sum_{l=1}^m \frac{|D_l|}{|D|} f_{D_l} = \frac{1}{m}\sum_{l=1}^m f_{D_l}, \tag{3}$$

where $\{f_{D_l}\}_l$ are produced by algorithm (2) with individual data subsets $\{D_l\}_l$.

When the distributed regularized least squares algorithm is applied with a fixed RKHS, it was shown in [31] that as long as the individual data size $|D|_l$ is not too small, mini-max optimal learning rates

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

can be obtained by a matrix analysis approach. In this approach, error bounds were established under some boundedness conditions on the normalized eigenfunctions of the integral operator associated with the kernel. Up to now, it is yet unknown when general kernels satisfy the boundedness condition in [31] for the eigenfunctions. Such conditions were successfully removed in the paper [11] with a novel integral operator approach and the optimal rate is achieved by means of effective dimensions of RKHSs and the regularity of target functions. Similar results were established for spectral algorithms [7,17], gradient descent algorithms [12], minimum error entropy principles [9] and the bias correct regularization kernel networks [7]. It should be noted that, either the boundedness assumption for eigenfunctions or some regularity requirement for target functions is necessary in earlier works, which is hardly satisfied when a fixed Gaussian kernel is used (to be discussed in Section 2). In addition, the obtained learning rate is constrained by the approximation error (or regularization error) of the RKHS. It has been pointed out in [29,18] that the fixed Gaussian RKHS $\mathcal{H}_\sigma$ has a poor approximation ability for target functions in Sobolev spaces. All these lead to the observation that analysis and results with these spaces are infeasible in many scenarios.

The purpose of this paper is to study the learning performance of (3) with flexible Gaussian RKHSs by estimating the L^2-error bounds in terms of the total data size N. Under some mild smoothness condition on the target function and weak eigenfunction assumption, the concrete rates can be almost optimal in the mini-max sense if the variance of K_σ is referred to as a tuning parameter in the learning process. Furthermore, by a semi-supervised approach, we improve the upper bound for data partition size m and the range of the smoothness parameter for the target function, which ensures the mini-max optimal rates in distributed learning.

The remainder of the paper is organized as follows. In Section 2, we first introduce some necessary notations and assumptions. We then state error bounds for distributed algorithm (3) in supervised learning and semi-supervised learning, respectively. Some discussions and comparisons with related work are also provided. Section 3 presents a bias-variance based decomposition for the learning error and some key lemmas that will be useful in the proof of our main results. The proofs of main results and necessary estimations are given in Sections 4 and 5. Some basic lemmas and proofs are postponed to the appendix.

2. Main results

We begin with some necessary notations and assumptions used in this paper. Our work is carried out in the setting of nonparametric regression. The probability measure ρ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ can be decomposed into the marginal distribution $\rho_{\mathcal{X}}$ on $\mathcal{X}$ and the conditional distributions $\rho(\cdot|x)$ for $x \in \mathcal{X}$. Denote $\|\cdot\|_\rho$ as the L^2-norm in the $L^2_{\rho_{\mathcal{X}}}$ space, which is defined by $\|f\|_\rho := \|f\|_{L^2_{\rho_{\mathcal{X}}}} = \left(\int_{\mathcal{X}} |f(x)|^2 d\rho_{\mathcal{X}}\right)^{\frac{1}{2}}$.

In regression analysis, the target function is the conditional mean $\mathbb{E}(Y|X = x)$ with $x \in \mathcal{X}$, that is, the regression function

$$f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x), \quad x \in \mathcal{X}.$$

Throughout the paper, the second moment condition is taken to describe the tail property of the output $\mathcal{Y}$, i.e., for some constant $B > 0$,

$$\mathbb{E}[y^2|x] := \int_{y\in\mathcal{Y}} y^2 d\rho(y|x) \le B^2, \quad \forall x \in \mathcal{X}. \tag{4}$$

It implies that f_ρ is bounded by B since $\|f_\rho\|_\infty \le \left(\int_{\mathcal{Y}} |y|^2 d\rho(y|x)\right)^{1/2} \le B$. The quality of the global estimator $\bar{f}_D$ for regression is measured by *the mean squared error* $\|\bar{f}_D - f_\rho\|^2_\rho$. In the following subsections, we state our main results in terms of the error bounds in supervised and semi-supervised learning, respectively.

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

2.1. Optimal rates for supervised learning

Recall that the Sobolev space $H^\alpha(\mathbb{R}^d)$ with index $\alpha > 0$ consists of all functions in $L^2(\mathbb{R}^d)$ such that the norm $\|f\|_{H^\alpha(\mathbb{R}^d)} := \left\{ \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \left(1 + |\omega|^2\right)^\alpha |\hat{f}(\omega)|^2 d\omega \right\}^{\frac{1}{2}}$ is finite where $\hat{f}$ is the Fourier transform of f. This paper aims at error analysis of the distributed algorithm (3) under the Sobolev smoothness condition on f_ρ. Let us state our first main result, whose proof will be provided in Section 4.

Theorem 1. *Assume that $\rho_{\mathcal{X}}$ has a density function with respect to the Lebesgue measure on $\mathcal{X}$ and the corresponding density is bounded away from 0 and ∞. Suppose that the regression function f_ρ equals the restriction onto $\mathcal{X}$ of some function in $H^\alpha(\mathbb{R}^d)$ for some $\alpha > d$. If $\xi > 0$ (which can be arbitrarily small), $\sigma = N^{-\frac{1}{2\alpha+d}}$, $\lambda = N^{-1}$ and $m \le N^{\frac{\alpha-d}{2\alpha+d}-\xi}$, then*

$$\mathbb{E}\left[\|\bar{f}_D - f_\rho\|_\rho^2\right] \le C N^{-\frac{2\alpha}{2\alpha+d}+\xi}. \tag{5}$$

To avoid superfluous notation, here and in the following, C denotes a constant independent of N or m which may be different at each occurrence.

It is known in [6] (Section 3.5) that when the distribution $\rho_{\mathcal{X}}$ is the normalized uniform measure on a domain $\mathcal{X}$, the i-th entropy number (cf. Definition 1) of the embedding $id : H^\alpha(\mathcal{X}) \to L^2_{\rho_{\mathcal{X}}}$ decays as $O\left(i^{-\frac{\alpha}{d}}\right)$. Recall that, when $\alpha > \frac{d}{2}$, the Sobolev space $H^\alpha(\mathcal{X})$ is continuously embedded into the space $L^\infty(\mathcal{X})$ of all bounded measurable functions on $\mathcal{X}$. Collecting these facts, we apply Lemma 4 in Appendix with $\Theta = H^\alpha(\mathcal{X}), \eta = \frac{d}{\alpha}$ and $\delta_N = N^{-\frac{2}{2+\eta}} = N^{-\frac{2\alpha}{2\alpha+d}}$ and get that when $f_\rho \in H^\alpha(\mathcal{X})$, the probability inequalities

$$\mathbb{P}\left\{D : \|\tilde{f}_D - f_\rho\|_\rho^2 \ge \delta\right\} \ge \begin{cases} c_0, & \text{if} \quad \delta \le N^{-\frac{2\alpha}{2\alpha+d}}, \\ c_1 \exp\{-c_2 \delta N\}, & \text{if} \quad \delta \ge N^{-\frac{2\alpha}{2\alpha+d}} \end{cases}$$

hold with some positive constants c_0, c_1, c_2 for any estimator $\tilde{f}_D$ based on the data set $D = \{(x_i, y_i)\}_{i=1}^N$. Together with $\mathbb{E}[f] = \int_0^\infty \mathbb{P}\{f \ge \delta\} d\delta$ for non-negative functions f, it yields

$$\inf_{\tilde{f}_D} \sup_{f_\rho \in H^\alpha(\mathcal{X})} \mathbb{E}\left[\|\tilde{f}_D - f_\rho\|_\rho^2\right] \ge c_0 \int_0^{N^{-\frac{2\alpha}{2\alpha+d}}} d\delta + c_1 \int_{N^{-\frac{2\alpha}{2\alpha+d}}}^\infty \exp\{-c_2 \delta N\} d\delta \ge c_0 N^{-\frac{2\alpha}{2\alpha+d}}$$

where the infimum ranges over all estimators $\tilde{f}_D$ based on D. This lower bound implies that our result (5) is nearly minimax-optimal.

Remark 1. A consequence of Theorem 1 is that the error bound for the classical (non-distributed) least squares regularization scheme (2) with flexible Gaussians can achieve learning rates of order $\mathbb{E}\left[\|f_D - f_\rho\|_\rho^2\right] = O\left(N^{-\frac{2\alpha}{2\alpha+d}+\xi}\right)$ when $\alpha > d$. It greatly improves the learning rates for Gaussian schemes derived in the literature [28,29] which are not optimal. The same learning rates are achieved for all $\alpha > 1$ in the work [5] by using an oracle probability inequality. But it requires an extra projection of f_D onto $[-B, B]$ and that the output space $\mathcal{Y}$ is supported on the interval $[-B, B]$ for some $B > 0$. We will show in Subsection 2.3 that by a semi-supervised approach the range $\alpha > d$ in Theorem 1 can be extended to $\alpha > 0$.

Theorem 1 suggests that the choice of σ that guarantees the optimal rate is data-dependent and changes with the data size N. Meanwhile, to keep the optimality, it establishes an upper bound for the number

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

m of data partitions in algorithm (3). It demonstrates that with flexible Gaussian kernels the learning performance of distributed learning with regularized least squares can be as good as that of one single machine which could process the whole data.

Gaussians with flexible variances are often taken in data analysis with convolutions. It would be interesting to study connections of our results to the recent work on deep learning with convolutional neural networks [23,33,34]. Meanwhile, stochastic gradient descent algorithm or other gradient descent analogues are widely used in training deep neural networks [24,30]. It is also worthwhile to see how to improve the performance of these algorithms by aid of flexible Gaussians.

2.2. Learning rates with a general condition on eigenfunctions

As mentioned in the introduction, for kernel-based distributed learning algorithms, considerable works were carried out by means of the eigensystems associated with the integral operator $L_{K_\sigma}: L^2_{\rho_\mathcal{X}} \to L^2_{\rho_\mathcal{X}}$, that is defined by

$$L_{K_\sigma}(f) = \int_{\mathcal{X}} f(x) K_\sigma(x,\cdot) d\rho_\mathcal{X}, \quad \forall f \in L^2_{\rho_\mathcal{X}}.$$

Denote the set of positive eigenvalues of L_{K_σ} as $\{\lambda_i\}_i$ arranged in a decreasing order, and a set of normalized eigenfunctions $\{\phi_i\}_i$ of L_{K_σ} in $L^2_{\rho_\mathcal{X}}$ corresponding to the eigenvalues $\{\lambda_i\}_i$. In this paper, we shall present error analysis for (3) in general situations by making use of special properties of eigenpairs $\{(\lambda_i, \phi_i)\}_i$ of L_{K_σ}. To this end, we first recall some basic properties on the eigenvalues $\{\lambda_i\}_i$.

Definition 1. Let E and F be Banach spaces and $L: E \to F$ be a bounded linear operator. Then the i-th entropy number $e_i(L), i \geq 1$, of L is defined by

$$e_i(L) := \inf\left\{\varepsilon > 0 : \exists x_1, \cdots, x_{2^{i-1}} \in L(B_E) \quad such \quad that \quad L(B_E) \subset \bigcup_{j=1}^{2^{i-1}} (x_j + \varepsilon B_F)\right\},$$

where B_E and B_F denote the closed unit balls of E and F, respectively.

Recall that for $p \in (0,\infty)$ and a decreasing, non-negative sequence $\{a_i\}_i$, the Lorentz (p,∞)-norm is defined by $\|\{a_i\}\|_{p,\infty} = \sup_{i\geq 1} i^{\frac{1}{p}} a_i$. The next lemma [19] shows that the eigenvalues $\{\lambda_i\}_i$ have the same asymptotic behavior as the squared L^2_μ-entropy numbers in terms of the Lorentz (p,∞)-norm.

Lemma 1. *Let $\mu = \rho_\mathcal{X}$ be a probability distribution on $\mathcal{X}$ and $\mathcal{H}_\sigma$ be the Gaussian RKHS over $\mathcal{X}$ with $\sigma \in (0,1]$. For each $0 < p < 1$, there exists a constant $c_p > 0$ only depending on p such that*

$$c_p \left\|e_i^2(id: \mathcal{H}_\sigma \to L^2_\mu)\right\|_{p,\infty} \leq \|\{\lambda_i\}\|_{p,\infty} \leq 4 \left\|e_i^2(id: \mathcal{H}_\sigma \to L^2_\mu)\right\|_{p,\infty}.$$

The lemma presents a simple way to characterize the decay rate of the eigenvalues $\{\lambda_i\}_i$ provided that the entropy numbers decreases polynomially fast. We then turn to the decay of the L^2_μ-entropy numbers that is usually used to measure the capacity of $\mathcal{H}_\sigma$. The following lemma can be found in [20] as Theorem 7.34.

Lemma 2. *Let $\mu = \rho_\mathcal{X}$ be a probability distribution on $\mathcal{X}$ and $\mathcal{H}_\sigma$ be the Gaussian RKHS over $\mathcal{X}$ with $\sigma \in (0,1]$. Then, for any $\varepsilon > 0$ and $0 < p < 1$, there exists a constant $c_{p,\varepsilon} > 0$ such that*

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

$$e_i(id:\mathcal{H}_\sigma\to L^2_\mu)\le c_{p,\varepsilon}\sigma^{-\frac{(1-p)(1+\varepsilon)d}{2p}}i^{-\frac{1}{2p}}$$

for all $i\ge 1$.

Consequently, since $0<1-p<1$, we can obtain that $\left\|e_i^2(id:\mathcal{H}_\sigma\to L^2_{\rho_\mathcal{X}})\right\|_{p,\infty}\le c^2_{p,\varepsilon}\sigma^{-\frac{(1+\varepsilon)d}{p}}$ and

$$\lambda_i\le 4c^2_{p,\varepsilon}\sigma^{-\frac{(1+\varepsilon)d}{p}}i^{-\frac{1}{p}},\quad \forall i\ge 1, 0<p<1. \tag{6}$$

Remark 2. The estimate (6) exhibits a quantitative relation between polynomial decays of $\{\lambda_i\}_i$ and the value of σ. It is known that the decay of the eigenvalues is essentially a smoothness condition for the kernel. As Gaussians are smooth kernels, their associated eigenvalues' decay actually should obey a bound of the exponential form, but the changing variance should be taken into consideration. This paper incorporates a bound (6) for $\{\lambda_i\}_i$ involving σ into the analysis. To our knowledge, there have been no results to reveal explicit relations between the value of σ and the exponential decay of $\{\lambda_i\}_i$.

Our second main result is based on an assumption regarding the boundedness of eigenfunctions $\{\phi_i\}_i$ using the fast decay of $\{\lambda_i\}_i$.

Assumption. There is a constant C_1 such that the eigenfunctions $\{\phi_i\}_i$ of L_{K_σ} satisfy

$$\sup_i \lambda_i^s\|\phi_i\|^2_\infty\le C_1\sigma^{-\nu},\ \textit{for some } 0<s<1 \textit{ and } \nu>0. \tag{7}$$

Let us give some comments about the above assumption. The assumption that the eigenfunctions ϕ_i are uniformly bounded appeared in the early literature on kernel methods, where it was even claimed that such a strong assumption holds for general Mercer kernels. A counterexample with a C^∞ kernel was presented in [32] which showed that smoothness of the Mercer kernel does not guarantee the uniform boundedness of the eigenfunctions. Therefore, it is not appropriate to assume only the uniform boundedness. Recall that the sequence $\{\lambda_i\}_i$ has an exponential decay for a fixed σ but the decay becomes worse when σ is smaller. Assumption (7) exhibits that the increase of the L^∞-norm bounds for the eigenfunctions can be very fast, which is comparable to an exponential rate. It is thus considerably weaker than that of the uniformly boundedness assumption. We also note that assumption (7) coincides with the one used in [16] for general Mercer kernels. They remarked that the example given in [32] with a C^∞ kernel without uniformly bounded eigenfunctions satisfies such a weaker boundedness condition. In addition, [21] derived the optimal rate for the classical least squares algorithm (2) by an interpolation condition between $\mathcal{H}_\sigma$ and L^∞, that is, for some $0<s<1$, there holds

$$\|f\|_\infty\le C\|f\|^s_{\mathcal{H}_\sigma}\|f\|^{1-s}_{L^2_{\rho_\mathcal{X}}},\quad \forall f\in\mathcal{H}_\sigma, \tag{8}$$

where C is a constant independent of f. Obviously, (7) is weaker than the above assumption. In this work, the eigenpairs (λ_i,ϕ_i) generated from Theorem 1 is a special case satisfying (7).

With these preliminaries, we can state our second main result, the general error bounds for the distributed algorithm (3) involving the regularized generalization error, which measures the approximation ability of RKHSs with flexible Gaussians.

Definition 2. With the kernel K_σ and the regularization parameter $\lambda>0$, the *regularized generalization error* is defined as

$$\mathcal{D}(\sigma,\lambda)=\arg\min_{f\in\mathcal{H}_\sigma}\left\{\|f-f_\rho\|^2_\rho+\lambda\|f\|^2_{\mathcal{H}_\sigma}\right\}=\|f^\sigma_\lambda-f_\rho\|^2_\rho+\lambda\|f^\sigma_\lambda\|^2_{\mathcal{H}_\sigma} \tag{9}$$

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

where f_λ^σ is the minimizer of $\|f-f_\rho\|_\rho^2+\lambda\|f\|_{\mathcal{H}_\sigma}^2$ over $\mathcal{H}_\sigma$.

Remark 3. Due to the low capacity of $\mathcal{H}_\sigma$, it was shown in [29] that in general, the regularization error of a fixed Gaussian RKHS only obeys a logarithmic decay as $O\left((\log(1/\lambda))^{-\theta}\right)$ for some $\theta>0$ under the Sobolev smoothness condition on f_ρ. However, most results in the literature on distributed learning or other kernel based methods required a nice approximation ability of the RKHS, such as a polynomial decay $O\left(\lambda^\theta\right)$ with $\theta>0$. That is one reason why such results cannot be applied to the single Gaussian RKHS directly. It is then shown in a series of papers [27–29,22] that with flexible Gaussian RKHSs, the approximation power of the regularization scheme was greatly improved. It should be clear from this point that the learning schemes generated by Gaussian kernels with flexible variances have advantages over those with fixed Gaussians.

Theorem 2. *Assume that (4) and (7) hold. If* $\frac{N}{m\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}\geq C_0N^\eta$ *for some constant* $C_0>0$ *(independent of* N*) and* $\eta>0$*, then*

$$\mathbb{E}\left[\|\bar{f}_D-f_\rho\|_\rho^2\right]\leq C\left[(\log N)^2\left(\frac{m\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{N}+\frac{m^2\mathcal{S}^2(\sigma,\lambda)}{N^2}\right)+\frac{1}{m}+1\right]\times\left\{\mathcal{D}(\sigma,\lambda)+\frac{m\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+\mathcal{N}(\sigma,\lambda)\right]}{N}\right\},\tag{10}$$

where $\mathcal{N}(\sigma,\lambda):=\sum_{i=1}^\infty\frac{1}{1+\lambda/\lambda_i}$, $\mathcal{S}(\sigma,\lambda):=\sum_{i=1}^\infty\frac{\|\phi_i\|_\infty^2}{1+\lambda/\lambda_i}$ *are the two kernel-related quantities, and* C *is a constant independent of* N*,* σ *or* m*.*

The first quantity $\mathcal{N}(\sigma,\lambda)$ is the *effective dimension* [2] of $\mathcal{H}_\sigma$, which measures the complexity of the RKHS with respect to $\rho_{\mathcal{X}}$ and is actually the trace of the operator $L_{K_\sigma}(L_{K_\sigma}+\lambda I)^{-1}$ as $Tr\left(L_{K_\sigma}(L_{K_\sigma}+\lambda I)^{-1}\right)$. The second quantity $\mathcal{S}(\sigma,\lambda)$ involves the tail behavior of $\{\phi_i\}$ that is characterized by their L_∞-norms. Under assumption (7), we show that even if $\|\phi_i\|_\infty$ increases exponentially fast, the series in $\mathcal{S}(\sigma,\lambda)$ is still convergent. As we see, σ and λ are tuning parameters for achieving good learning rates in the learning process. To demonstrate the explicit learning rates of algorithm (3), a general case with suitably chosen parameters and the range of m is covered as follows.

Theorem 3. *Assume that (4) and (7) hold. Let the regression function* f_ρ *equal the restriction onto* $\mathcal{X}$ *of some function in* $H^\alpha(\mathbb{R}^d)$ *for some* $\alpha>\frac{d(1+s)+\nu}{2(1-s)}$ *and* $\frac{d\rho_{\mathcal{X}}}{dx}\in L^\infty(\mathcal{X})$*. Take* $\sigma=N^{-\frac{1}{2\alpha+d}}$ *and* $\lambda=N^{-1}$*. If*

$$m\leq N^{\frac{1}{2}\left[1-\frac{2d+\nu}{2\alpha+d}-s\right]-\xi},\tag{11}$$

then

$$\mathbb{E}\left[\|\bar{f}_D-f_\rho\|_\rho^2\right]\leq CN^{-\frac{2\alpha}{2\alpha+d}+\xi}\tag{12}$$

where ξ *is a fixed positive number which can be arbitrarily small and* C *is a constant independent of* N *or* m*.*

Remark 4. For the distributed regularized least squares associated with a general RKHS, [31] derived the optimal rate when f_ρ lies in the RKHS, but under the much stronger condition than (7), that $\{\phi_i\}$ are uniformly bounded. Meanwhile, when f_ρ is outside the RKHS, they claimed that the error rate is controlled by the approximation error in a ball with radius $R\geq1$, that is, $\inf_{\|f\|_{\mathcal{H}_\sigma}\leq R}\|f-f_\rho\|_\rho$. It has been shown in [18] that, for target functions in Sobolev spaces which are not C^∞, the error has a logarithmic convergence

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

rate in the form of $(\log R)^{-\theta}$. It implies that under the Sobolev smoothness condition on f_ρ the learning rate obtained in [31] is only in the logarithmic form of $(\log N)^{-\theta}$.

Another line of research on distributed kernel-based learning algorithms has been built on the effective dimensions of RKHSs and the regularity of target functions [7,10,11,17]. The former is similar to previous work on kernel methods [2,11,17] and can be characterized by the decay of eigenvalues $\{\lambda_i\}$ of L_{K_σ}. The latter, the regularity of f_ρ, imposes a strong smoothness condition on f_ρ for Gaussian RKHSs, that f_ρ lies in the range space $L^r_{K_\sigma}(L^2_{\rho_\mathcal{X}})$ for some $r > 0$, requiring $f_\rho \in C^\infty$. This prevents their results from being applied in general situations.

It should be noted that, condition $\frac{d\rho_\mathcal{X}}{dx} \in L^\infty(\mathcal{X})$ is not necessary to obtain (12) in Theorem 3. Instead of it, if the regularization error $\mathcal{D}(\sigma,\lambda)$ decays as $O\left(\lambda^{\frac{2\alpha}{2\alpha+d}}\right)$ with some suitably chosen $\lambda = \lambda(\sigma)$, error rate (12) still holds apart from a slightly different leading constant. In particular, we know from the past work [27,28] that for $f_\rho \in H^\alpha(\mathcal{X})$, $\mathcal{D}(\sigma,\lambda) = O\left(\sigma^{2\alpha} + \lambda\sigma^{-d}\right)$ if $\frac{d\rho_\mathcal{X}}{dx} \in L^\infty(\mathcal{X})$. Taking the trade-off $\lambda = \lambda(\sigma) = \sigma^{2\alpha+d}$, we find (12) holds.

It turns out that the error rate of $\mathcal{D}(\sigma,\lambda)$ plays an important role in deriving the explicit error rate for $\|\bar{f}_D - f_\rho\|^2_\rho$. As shown in Theorem 2, sharper bounds on $\mathcal{D}(\sigma,\lambda)$ will lead to better learning error. More precisely, the error estimate for $\mathcal{D}(\sigma,\lambda)$ dominates the choice of $\lambda = \lambda(\sigma)$ in our learning scheme. To our knowledge, in existing works, the estimates of $\mathcal{D}(\sigma,\lambda)$ for $f_\rho \in H^\alpha(\mathcal{X})$ satisfy the bounds of the form $O\left(\sigma^{h(\alpha)} + \lambda\sigma^{-d}\right)$ where $h(\alpha)$ is a function of the smoothness index α. Using these estimates, the best choice in algorithm (3) should be $\lambda = \sigma^{h(\alpha)+d}$ that ensures a sharp bound $O\left(\lambda^{\frac{h(\alpha)}{h(\alpha)+d}}\right)$ for $\mathcal{D}(\sigma,\lambda)$. The explicit forms of $h(\alpha)$ can be achieved according to the behavior of the approximation error estimated from the priori knowledge on the distribution ρ or other priori conditions on practical problems, such as condition $\frac{d\rho_\mathcal{X}}{dx} \in L^\infty(\mathcal{X})$ mentioned above. For more estimates of $\mathcal{D}(\sigma,\lambda)$, one can refer to the papers [5,27–29].

In Theorem 3, the order of the learning rate (12) is obtained by requiring $\alpha > \frac{(1+s)d+\nu}{2(1-s)}$. By tracing the proof of Theorem 3 in Section 4, we find that if α is less than the lower bound, the corresponding error for (3) is worse than (12) and far from the optimality even in the non-distributed case ($m = 1$). Besides, the upper bound (11) of m to obtain the sharp error rate is restricted by assumption (7) for eigenfunctions. We will address the issues in the next subsection.

2.3. Optimal rates for semi-supervised learning

As shown in Theorems 1 and 3, some restrictions on the range of m and the smoothness parameter α are required in order to obtain almost optimal rates. In this subsection, we show how to relax the restrictions by a semi-supervised approach [3,4].

Let unlabeled data $D^* = \{x_i^*\}_{i=1}^{N^*}$ with $|D^*| = N^*$ be drawn independently according to $\rho_\mathcal{X}$. In the divide-and-conquer strategy, D^* are partitioned equally into m subsets, i.e.

$$D^* = \bigcup_{l=1}^{m} D_l^*, \ \textit{with}\ |D_l^*| = n^*, \ n^* = \frac{N^*}{m}.$$

We construct a new data set $\tilde{D} = \bigcup_{l=1}^m \tilde{D}_l$ with $|\tilde{D}| := \tilde{N}$ by

$$\tilde{D}_l = D_l \bigcup D_l^* = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{\tilde{n}} \ \textit{with}\ \tilde{n} := n + n^*$$

where

$$(\tilde{x}_i, \tilde{y}_i) = \begin{cases} (x_i, \frac{\tilde{n}}{n} y_i), & \text{if} \quad 1 \le i \le n, \ (x_i, y_i) \in D_l, \\ (x^*_{i-n}, 0), & \text{if} \quad n+1 \le i \le \tilde{n}, \ x^*_{i-n} \in D_l^*. \end{cases}$$

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

Given the new set $\tilde{D}$, we get the averaged output $\bar{f}_{\tilde{D}}$ from the base algorithm (2) as follows

$$\bar{f}_{\tilde{D}} = \frac{1}{m}\sum_{l=1}^{m} f_{\tilde{D}_l}. \tag{13}$$

Now we can state our last main result on improved results in a semi-supervised learning framework.

Theorem 4. *Define $\bar{f}_{\tilde{D}}$ by (13) with $\tilde{D}$. Assume that (4) and (7) hold. Suppose that the regression function f_ρ equals the restriction onto $\mathcal{X}$ of some function in $H^\alpha(\mathbb{R}^d)$ for some $\alpha > 0$ and $\frac{d\rho_{\mathcal{X}}}{dx} \in L^\infty(\mathcal{X})$. Let $\tilde{N} \geq N^{\frac{2d+\nu}{2\alpha+d}+s+2\xi}$ with $\xi > 0$. Take $\sigma = N^{-\frac{1}{2\alpha+d}}$ and $\lambda = N^{-1}$. If*

$$m \leq \min\left\{\tilde{N}^{\frac{1}{2}} N^{-\frac{2d+\nu}{2(2\alpha+d)}-\frac{s}{2}-\xi}, \tilde{N}^{\frac{2}{3}} N^{-\frac{2d+2\nu}{3(2\alpha+d)}-\frac{2s}{3}-\xi}\right\}, \tag{14}$$

then

$$\mathbb{E}\left[\|\bar{f}_{\tilde{D}} - f_\rho\|_\rho^2\right] \leq C N^{-\frac{2\alpha}{2\alpha+d}+\xi},$$

where C is a constant independent of N or m.

Compared with Theorem 3, the upper bound (14) of m is enlarged as more unlabeled data are allowed in distributed learning. Meanwhile, provided that the number of unlabeled data is large enough, that is $\tilde{N} \geq N^{\frac{2d+\nu}{2\alpha+d}+s+2\xi}$ with $\xi > 0$, the range of α that ensures the order $O\left(N^{-\frac{2\alpha}{2\alpha+d}+\xi}\right)$ can be extended from $\alpha > \frac{d(1+s)+\nu}{2(1-s)}$ to $\alpha > 0$ for any $m \geq 1$ satisfying (14). In the next corollary, the improvements are demonstrated for Theorem 1.

Corollary 1. *Assume that the density function of $\rho_{\mathcal{X}}$ exists and is bounded away from 0 and ∞. Suppose that the regression function f_ρ equals the restriction onto $\mathcal{X}$ of some function in $H^\alpha(\mathbb{R}^d)$ for some $\alpha > 0$. Let $\tilde{N} \geq N^{\frac{3d}{2\alpha+d}+\xi}$ with $\xi > 0$. If $\sigma = N^{-\frac{1}{2\alpha+d}}$, $\lambda = N^{-1}$ and $m \leq \tilde{N}^{\frac{1}{2}} N^{-\frac{3d}{2(2\alpha+d)}-\xi}$, then*

$$\mathbb{E}\left[\|\bar{f}_{\tilde{D}} - f_\rho\|_\rho^2\right] \leq C N^{-\frac{2\alpha}{2\alpha+d}+\xi}$$

where C is a constant independent of N or m.

3. Error bounds for the bias-variance decomposition

Recall (9). Then a natural error decomposition can be derived as

$$\mathbb{E}\left[\|\bar{f}_D - f_\rho\|_\rho^2\right] \leq 2\mathbb{E}\left[\|\bar{f}_D - f_\lambda^\sigma\|_\rho^2\right] + 2\|f_\lambda^\sigma - f_\rho\|_\rho^2 \leq 2\mathbb{E}\left[\|\bar{f}_D - f_\lambda^\sigma\|_\rho^2\right] + 2\mathcal{D}(\sigma,\lambda). \tag{15}$$

Our key analysis is about the first term $\mathbb{E}\left[\|\bar{f}_D - f_\lambda^\sigma\|_\rho^2\right]$ since the approximation error $\mathcal{D}(\sigma,\lambda)$ has been studied well in [5,28,29]. To present the analysis, we need a bias-variance decomposition as follows.

Proposition 1. *If $|D_1| = \cdots = |D_m| = n$ with $N = mn$, then*

$$\mathbb{E}\left[\|\bar{f}_D - f_\lambda^\sigma\|_\rho^2\right] \leq \frac{1}{m^2}\sum_{l=1}^{m}\mathbb{E}\left[\|f_{D_l} - f_\lambda^\sigma\|_\rho^2\right] + \frac{1}{m}\sum_{l=1}^{m}\left\|\mathbb{E}\left[f_{D_l} - f_\lambda^\sigma\right]\right\|_\rho^2. \tag{16}$$

The proposition enables us to conduct our analysis by estimating the **variance** term $\mathbb{E}\left[\|f_{D_l} - f_\lambda^\sigma\|_\rho^2\right]$ and **bias** term $\left\|\mathbb{E}\left[f_{D_l} - f_\lambda^\sigma\right]\right\|_\rho^2$, respectively. The rest of this subsection is devoted to bounding them achieved by special features of Gaussians and the matrix analysis approach in [31].

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

3.1. Some probabilistic estimates

To bound the variance and bias terms in (16), we need some probabilistic estimates which will be proved in the appendix. To simplify the notations in the following, denote $\mathbb{E}^*[\cdot]$ as the conditional expectation $\mathbb{E}[\cdot|x_1, x_2, \cdots]$ and $\|\cdot\|_2$ as the ℓ^2-norm for a sequence in ℓ^2. To make it easy to follow our presentation, we summarize some notations that are repeatedly used in the proof in Table 1 given in Appendix.

Proposition 2. *Under Assumption (4), there holds*

$$\mathbb{E}^*[\|f_{D_l}\|^2_{\mathcal{H}_\sigma}] \le B^2/\lambda \tag{17}$$

and

$$\mathbb{E}^*[\|f_{\tilde{D}_l}\|^2_{\mathcal{H}_\sigma}] \le \tilde{B}^2/\lambda, \text{ with } \tilde{B}^2 := B^2\tilde{N}/N = B^2\tilde{n}/n. \tag{18}$$

Next, we state some estimates derived by the matrix analysis approach. Let u be a positive integer. Denote $\Phi = [\phi_j(x_i)]_{i=1\,j=1}^{n\quad u} \in \mathbb{R}^{n\times u}$ and $Q = \left(I + \lambda\, diag\left(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_u}\right)\right)^{\frac{1}{2}} \in \mathbb{R}^{u\times u}$.

Proposition 3. *Let $u \in \mathbb{N}$. The following probability inequality holds for any $t > 0$,*

$$\mathbb{P}\left\{\left\|Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1}\right\| \ge t\right\} \le 2u\exp\left\{-\frac{nt^2/2}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda) + 2\mathcal{S}(\sigma,\lambda)t/3}\right\}, \tag{19}$$

and the expectation for the second moment of $\left\|Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1}\right\|$ is bounded as

$$\mathbb{E}\left[\left\|Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1}\right\|^2\right] \le 64\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{n} + \frac{64\mathcal{S}^2(\sigma,\lambda)}{n^2}\right](\log(u))^2. \tag{20}$$

Proposition 4. *Let $u \in \mathbb{N}$ and $\mathbf{a^1} = (\langle f_\lambda^\sigma, \phi_i\rangle_\rho)_{i=1}^u \in \mathbb{R}^u$. The following bounds hold*

$$\left\|\lambda Q^{-1} diag\left(\frac{1}{\lambda_1}, \cdots, \frac{1}{\lambda_u}\right)\mathbf{a^1}\right\|_2^2 \le \lambda\|f_\lambda^\sigma\|^2_{\mathcal{H}_\sigma}, \tag{21}$$

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v}\right\|_2^2\right] \le 2Tr(L_{K_\sigma})\beta_u\left(B^2/\lambda + \|f_\lambda^\sigma\|^2_{\mathcal{H}_\sigma}\right)/\lambda \tag{22}$$

with $\mathbf{v} = \left(\sum_{j=u+1}^{\infty}\langle f_{D_l} - f_\lambda^\sigma, \phi_j\rangle_\rho\phi_j(x_i)\right)_{i=1}^n \in \mathbb{R}^n$, $\beta_u = \sum_{j=u+1}^{\infty}\lambda_j\|\phi_j\|^2_\infty$ *and* $Tr(L_{K_\sigma}) := \sum_j \lambda_j$, *and*

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon\right\|_2^2\right] \le 2\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda) + 4B^2\mathcal{N}(\sigma,\lambda)\right]/n, \tag{23}$$

where $\epsilon = (f_\lambda^\sigma(x_i) - y_i)_{i=1}^n \in \mathbb{R}^n$.

3.2. Bounding the variance term

Now we can estimate the variance term (16).

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

Proposition 5. ***(variance bound)*** *Define* $\{f_{D_l}\}$ *by (2) and assume* $0 < \lambda \leq 1$. *Then we have*

$$\mathbb{E}\left[\|f_{D_l} - f_\lambda^\sigma\|_\rho^2\right] \leq \Delta + 2\lambda_{u+1}\left(B^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right), \tag{24}$$

where u *is an integer, and*

$$\Delta := 12\lambda\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 + \frac{24\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda) + 4B^2\mathcal{N}(\sigma,\lambda)\right]}{n} + 8\left(3Tr(L_{K_\sigma})\beta_u/\lambda + u\exp\left\{-\frac{n/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda) + \mathcal{S}(\sigma,\lambda)/3}\right\}\right)\left(B^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right). \tag{25}$$

Proof. Denote $E = f_{D_l} - f_\lambda^\sigma$. Then there exists a coefficient sequence $\mathbf{e} = \{e_i\}_i \in \ell^2$ such that E can be expanded as $E = \sum_{i=1}^\infty e_i\phi_i$ with $e_i = \langle E, \phi_i\rangle_\rho$.

To prove (24), we only need to estimate $\mathbb{E}[\|\mathbf{e}\|_2^2]$ by the fact that

$$\|E\|_\rho^2 = \|f_{D_l} - f_\lambda^\sigma\|_\rho^2 = \|\mathbf{e}\|_2^2 = \sum_i |e_i|^2.$$

Note that $\{\sqrt{\lambda_i}\phi_i\}$ forms an orthonormal basis of $\mathcal{H}_\sigma$, so we also get

$$\|E\|_{\mathcal{H}_\sigma}^2 = \|f_{D_l} - f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 = \sum_{i=1}^\infty \frac{e_i^2}{\lambda_i}. \tag{26}$$

Fixing the integer u, we decompose the vector $\mathbf{e}$ into two parts $\mathbf{e^1} = \{e_1, \cdots, e_u\}$ and $\mathbf{e^2} = \{e_{u+1}, e_{u+2}, \cdots\}$. It implies that $\mathbb{E}[\|\mathbf{e}\|_2^2] = \mathbb{E}[\|\mathbf{e^1}\|_2^2] + \mathbb{E}[\|\mathbf{e^2}\|_2^2]$. First, we consider $\mathbb{E}[\|\mathbf{e^2}\|_2^2]$. Since $\{\lambda_i\}$ is arranged in a decreasing order, by (26),

$$\begin{aligned}\mathbb{E}[\|\mathbf{e^2}\|_2^2] = \mathbb{E}\left[\sum_{i=u+1}^\infty |e_i|^2\right] &\leq \lambda_{u+1}\mathbb{E}\left[\sum_{i=u+1}^\infty \frac{e_i^2}{\lambda_i}\right] \leq \lambda_{u+1}\mathbb{E}[\|E\|_{\mathcal{H}_\sigma}^2] \\ &\leq 2\lambda_{u+1}(\mathbb{E}[\|f_{D_l}\|_{\mathcal{H}_\sigma}^2] + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2).\end{aligned} \tag{27}$$

Next, we turn to estimate $\mathbb{E}[\|\mathbf{e^1}\|_2^2]$. By the definition of f_{D_l}, we know that

$$\frac{1}{n}\sum_{i=1}^n \left(f_{D_l}(x_i) - y_i\right) K_\sigma(x_i, \cdot) + \lambda f_{D_l} = 0.$$

Recall $\epsilon_i = f_\lambda^\sigma(x_i) - y_i, i = 1, \cdots, n$. It follows that

$$\frac{1}{n}\sum_{i=1}^n E(x_i)K_\sigma(x_i,\cdot) + \frac{1}{n}\sum_{i=1}^n \epsilon_i K_\sigma(x_i,\cdot) + \lambda E = -\lambda f_\lambda^\sigma.$$

Let $f_\lambda^\sigma = \sum_{i=1}^\infty a_i\phi_i$ with $\mathbf{a} = \{a_i\}_i \in \ell^2$ in the basis $\{\phi_i\}_i$. Computing the $\mathcal{H}_\sigma$ inner products of both sides of the above equality with ϕ_k, we obtain that

$$\frac{1}{n}\sum_{i=1}^n E(x_i)\phi_k(x_i) + \frac{1}{n}\sum_{i=1}^n \epsilon_i\phi_k(x_i) + \lambda\frac{e_k}{\lambda_k} = -\lambda\frac{a_k}{\lambda_k}. \tag{28}$$

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

Recall $\mathbf{v} = [v_1, \cdots, v_n]^{\mathrm{T}}$ by $v_i = \sum_{j=u+1}^{\infty} e_j \phi_j(x_i), i = 1, \cdots, n$. Recall the matrix $\Phi = (\Phi_{ij})_{i,j} \in \mathbb{R}^{n\times u}$ with $\Phi_{ij} = (\phi_j(x_i))$ for $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, u\}$. Recall that $[E(x_i)]_i = \left[\sum_{j=1}^{u} e_j \phi_j(x_i) + \sum_{j=u+1}^{\infty} e_j \phi_j(x_i)\right]_i = \Phi \mathbf{e^1} + \mathbf{v}$. Applying (28) with $k = 1, \cdots, u$, we get that

$$\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi + \lambda M^{-1}\right)\mathbf{e^1} = -\lambda M^{-1}\mathbf{a^1} - \frac{1}{n}\Phi^{\mathrm{T}}\mathbf{v} - \frac{1}{n}\Phi^{\mathrm{T}}\epsilon, \tag{29}$$

where $\mathbf{a^1} = [a_1, \cdots, a_u]^{\mathrm{T}}$ denotes the vector formed by the first u terms of the sequence $\mathbf{a} = \{a_l\}_l$, $\epsilon = [\epsilon_1, \cdots, \epsilon_n]^{\mathrm{T}} \in \mathbb{R}^n$, and $M = diag(\lambda_1, \cdots, \lambda_u) \in \mathbb{R}^{u\times u}$.

In addition, since $Q = \left(I + \lambda M^{-1}\right)^{\frac{1}{2}}$, one gets

$$\frac{1}{n}\Phi^{\mathrm{T}}\Phi + \lambda M^{-1} = I + \lambda M^{-1} + \frac{1}{n}\Phi^{\mathrm{T}}\Phi - I = Q\left(I + Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1}\right)Q.$$

Putting this decomposition into (29) yields that

$$\left(I + Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1}\right)Q\mathbf{e^1} = -\lambda Q^{-1}M^{-1}\mathbf{a^1} - \frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v} - \frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon. \tag{30}$$

Denote the event $A := \left\{\left\|Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1}\right\| \le \frac{1}{2}\right\}$. If the event A happens, then

$$\left\|I + Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1}\right\| \ge \frac{1}{2}$$

and

$$\begin{aligned}\|Q\mathbf{e^1}\|_2^2 &\le 4\left\|\lambda Q^{-1}M^{-1}\mathbf{a^1} + \frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v} + \frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon\right\|_2^2 \\ &\le 12\left\|\lambda Q^{-1}M^{-1}\mathbf{a^1}\right\|_2^2 + 12\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v}\right\|_2^2 + 12\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon\right\|_2^2.\end{aligned} \tag{31}$$

Note that $\|\mathbf{e^1}\|_2^2 \le \|Q\mathbf{e^1}\|_2^2$. So we estimate $\mathbb{E}[\|\mathbf{e^1}\|_2^2]$ by bounding $\mathbb{E}[\|Q\mathbf{e^1}\|_2^2]$ as

$$\begin{aligned}\mathbb{E}[\|Q\mathbf{e^1}\|_2^2] &= \mathbb{E}\left[\mathbf{I}(A)\|Q\mathbf{e^1}\|_2^2\right] + \mathbb{E}\left[\mathbf{I}(A^c)\|Q\mathbf{e^1}\|_2^2\right] \\ &= \mathbb{E}\left[\mathbf{I}(A)\|Q\mathbf{e^1}\|_2^2\right] + \mathbb{E}\left[\mathbf{I}(A^c)\mathbb{E}^*\left[\|Q\mathbf{e^1}\|_2^2\right]\right] \\ &:= I_1 + I_2,\end{aligned}$$

where $\mathbf{I}(\cdot)$ denotes the indicator function.

For I_1, using the estimate (31), we have that

$$I_1 \le 12\left\|\lambda Q^{-1}M^{-1}\mathbf{a^1}\right\|_2^2 + 12\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v}\right\|_2^2\right] + 12\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon\right\|_2^2\right].$$

For I_2, we find

$$\mathbb{E}^*\left[\left\|Q\mathbf{e^1}\right\|_2^2\right] = \mathbb{E}^*\left[\sum_{k=1}^{u}(1 + \lambda/\lambda_k)e_k^2\right] = \mathbb{E}^*\left[\left(\sum_{k=1}^{u} e_k^2 + \lambda\sum_{k=1}^{u} e_k^2/\lambda_k\right)\right]$$

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

$$\leq \mathbb{E}^*\left[\left(\|E\|_\rho^2+\lambda\|E\|_{\mathcal{H}_\sigma}^2\right)\right] \leq (1+\lambda)\mathbb{E}^*\left[\|E\|_{\mathcal{H}_\sigma}^2\right]$$
$$\leq 2(1+\lambda)\mathbb{E}^*\left[\|f_{D_l}\|_{\mathcal{H}_\sigma}^2+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right] \leq 4\left(\mathbb{E}^*\left[\|f_{D_l}\|_{\mathcal{H}_\sigma}^2\right]+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right).$$

It follows from (17) that

$$I_2 \leq 4\mathbb{P}\left(A^c\right)\left(B^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right).$$

Based on the above bounds for I_1 and I_2, we have that

$$\mathbb{E}[\|Q\mathbf{e^1}\|_2^2] \leq 12\left\|\lambda Q^{-1}M^{-1}\mathbf{a^1}\right\|_2^2+12\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v}\right\|_2^2\right]$$
$$+12\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon\right\|_2^2\right]+4\mathbb{P}\left(A^c\right)\left(B^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right). \tag{32}$$

Applying the probability inequality (19) with $t=\frac{1}{2}$, one gets

$$\mathbb{P}\left(A^c\right) \leq 2u\exp\left\{-\frac{n/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\}.$$

Plugging the above probability bound for $\mathbb{P}\left(A^c\right)$ into (32), together with the estimates (21), (22) and (23), we obtain that

$$\mathbb{E}[\|Q\mathbf{e^1}\|_2^2] \leq 12\lambda\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2+24Tr(L_{K_\sigma})\beta_u\left(B^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right)/\lambda+\frac{24\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+4B^2\mathcal{N}(\sigma,\lambda)\right]}{n}$$
$$+8u\exp\left\{-\frac{n/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\}\left(B^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right):=\Delta. \tag{33}$$

Recall the fact

$$\mathbb{E}\left[\|f_{D_l}-f_\lambda^\sigma\|_\rho^2\right]=\mathbb{E}[\|\mathbf{e}\|_2^2] \leq \mathbb{E}[\|Q\mathbf{e^1}\|_2^2]+\mathbb{E}[\|\mathbf{e^2}\|_2^2].$$

This together with the estimate (27) for $\mathbb{E}[\|\mathbf{e^2}\|_2^2]$ yields the conclusion (24).

The proof is complete. □

3.3. Bounding the bias term

In this subsection we estimate the bias term in (16).

Proposition 6. (bias bound) *Define f_{D_l} by (2) with D_l. Then we have*

$$\left\|\mathbb{E}\left[f_{D_l}-f_\lambda^\sigma\right]\right\|_\rho^2 \leq \left(4Tr(L_{K_\sigma})\beta_u/\lambda+2\lambda_{u+1}\right)\left(B^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right)$$
$$+128\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{n}+\frac{64\mathcal{S}^2(\sigma,\lambda)}{n^2}\right](\log(u))^2\Delta, \tag{34}$$

where Δ and β_u are defined in (22) and (25), respectively.

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

Proof. Recall that $E := f_{D_l} - f_\lambda^\sigma = \sum_{i=1}^\infty e_i\phi_i$. Then

$$\|\mathbb{E}\left[f_{D_l} - f_\lambda^\sigma\right]\|_\rho^2 = \left\|\mathbb{E}\left[\sum_{i=1}^u e_i\phi_i\right]\right\|_\rho^2 + \left\|\mathbb{E}\left[\sum_{i=u+1}^\infty e_i\phi_i\right]\right\|_\rho^2 = \left\|\mathbb{E}\mathbf{e^1}\right\|_2^2 + \left\|\mathbb{E}\mathbf{e^2}\right\|_2^2. \tag{35}$$

The second term on the right-hand side of (35) can be bounded by (17) as

$$\begin{aligned}\left\|\mathbb{E}\mathbf{e^2}\right\|_2^2 &\le \mathbb{E}\left[\|\mathbf{e^2}\|_2^2\right] \le \lambda_{u+1}\mathbb{E}\left[\sum_{i=u+1}^\infty e_i^2/\lambda_i\right] \\ &\le \lambda_{u+1}\mathbb{E}\left[\|f_{D_l} - f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right] \le 2\lambda_{u+1}(B^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2).\end{aligned} \tag{36}$$

Hence, the key in our estimate is to bound the first term $\left\|\mathbb{E}\mathbf{e^1}\right\|_2^2$ on the right-hand side of (35). Notice that the expression (30) can be rewritten as

$$Q\mathbf{e^1} = -\lambda Q^{-1}M^{-1}\mathbf{a^1} - \frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v} - \frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon - Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}.$$

By the definition of f_λ^σ, we know that

$$\mathbb{E}\left[(f_\lambda^\sigma(x) - y)K_\sigma(x,\cdot)\right] = L_{K_\sigma}(f_\lambda^\sigma - f_\rho) = -\lambda f_\lambda^\sigma.$$

Taking the RKHS inner products of both sides of the above equality with the basis ϕ_k, we see

$$\mathbb{E}\left[(f_\lambda^\sigma(x) - y)\phi_k(x)\right] = -\frac{\lambda a_k}{\lambda_k}.$$

Recalling that $\epsilon_i = f_\lambda^\sigma(x_i) - y_i$, then

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\epsilon_i\phi_k(x_i)\right] = -\frac{\lambda a_k}{\lambda_k}.$$

Applying the above relation for $k = 1, \cdots, u$, we get

$$\mathbb{E}\left[\frac{1}{n}\Phi^{\mathrm{T}}\epsilon\right] = -\lambda M^{-1}\mathbf{a^1}.$$

As a consequence, by Jensen's inequality and $\left\|\mathbb{E}\mathbf{e^1}\right\|_2^2 \le \left\|\mathbb{E}Q\mathbf{e^1}\right\|_2^2$, we obtain that

$$\begin{aligned}\left\|\mathbb{E}\mathbf{e^1}\right\|_2^2 &\le \left\|\mathbb{E}\left[\lambda Q^{-1}M^{-1}\mathbf{a^1} + \frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v} + \frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon + Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}\right]\right\|_2^2 \\ &= \left\|\mathbb{E}\left[\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v} + Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}\right]\right\|_2^2 \\ &\le 2\left\|\mathbb{E}\left[\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v}\right]\right\|_2^2 + 2\left\|\mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}\right]\right\|_2^2 \\ &\le 2\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v}\right\|_2^2\right] + 2\left\|\mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}\right]\right\|_2^2.\end{aligned} \tag{37}$$

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

For the second term $\left\|\mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}\right]\right\|_2^2$ above, by Cauchy-Schwarz inequality, we have

$$\left\|\mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}\right]\right\|_2^2 \le \left(\mathbb{E}\left[\left\|Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}\right\|_2\right]\right)^2$$
$$\le \mathbb{E}\left[\left\|Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1}\right\|^2\right]\mathbb{E}\left[\left\|Q\mathbf{e^1}\right\|_2^2\right].$$

Combining this with (20) implies

$$\left\|\mathbb{E}\left[Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)\mathbf{e^1}\right]\right\|_2^2 \le 64\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{n} + \frac{64\mathcal{S}^2(\sigma,\lambda)}{n^2}\right](\log(u))^2\mathbb{E}\left[\left\|Q\mathbf{e^1}\right\|_2^2\right].$$

This together with (37) and (36) yields

$$\|\mathbb{E}\left[f_{D_l} - f_\lambda^\sigma\right]\|_\rho^2 \le 2\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v}\right\|_2^2\right]$$
$$+ 128\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{n} + \frac{64\mathcal{S}^2(\sigma,\lambda)}{n^2}\right](\log(u))^2\mathbb{E}\left[\left\|Q\mathbf{e^1}\right\|_2^2\right] + 2\lambda_{u+1}(B^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2). \tag{38}$$

The term $\mathbb{E}\left[\left\|Q\mathbf{e^1}\right\|_2^2\right]$ can be bounded by (33). Finally, we put the estimate (22) into (38) to draw our conclusion (35). □

4. Proofs of main results in supervised learning

This section is devoted to proving our main results stated in Subsections 2.1 and 2.2. To this end, we first estimate the quantities $\mathcal{S}(\sigma,\lambda), \mathcal{N}(\sigma,\lambda), \beta_u$ and $Tr(L_{K_\sigma})$ as follows.

Proposition 7. *For $p \in (0,1)$, $\varepsilon > 0$, we have*

$$\mathcal{N}(\sigma,\lambda) \le C_2\sigma^{-(1+\varepsilon)d}\lambda^{-p}, \tag{39}$$
$$Tr(L_{K_\sigma}) \le C_3\sigma^{-\frac{(1+\varepsilon)d}{p}}, \tag{40}$$

where the constants C_2 and C_3 are independent of λ or σ (given explicitly in the proof). Under assumption (7), for $0 < \lambda \le 1, \sigma > 0$ and $u \in \mathbb{N}$,

$$\mathcal{S}(\sigma,\lambda) \le C_4\lambda^{-(p+s)}\sigma^{-(1+\varepsilon)d-\nu}, \tag{41}$$
$$\beta_u \le C_5\sigma^{-\frac{(1-s)(1+\varepsilon)d+p\nu}{p}}u^{-\frac{1-p-s}{p}}, \tag{42}$$

where the constants C_4, C_5 are independent of λ, σ or u (given explicitly in the proof).

Proof. For $\mathcal{N}(\sigma,\lambda)$ and $Tr(L_{K_\sigma})$, we calculate by using (6)

$$\mathcal{N}(\sigma,\lambda) = \sum_{i\ge1}\frac{1}{1+\lambda/\lambda_i} = \sum_{i\ge1}\frac{\lambda_i}{\lambda_i+\lambda} \le \sum_{i\ge1}\frac{4c_{p,\varepsilon}^2\sigma^{-\frac{(1+\varepsilon)d}{p}}i^{-\frac{1}{p}}}{4c_{p,\varepsilon}^2\sigma^{-\frac{(1+\varepsilon)d}{p}}i^{-\frac{1}{p}}+\lambda}$$
$$\le \int_0^\infty \frac{4c_{p,\varepsilon}^2\sigma^{-\frac{(1+\varepsilon)d}{p}}t^{-\frac{1}{p}}}{4c_{p,\varepsilon}^2\sigma^{-\frac{(1+\varepsilon)d}{p}}t^{-\frac{1}{p}}+\lambda}dt \le \left(4c_{p,\varepsilon}^2\sigma^{-\frac{(1+\varepsilon)d}{p}}\right)^p\lambda^{-p}\int_0^\infty\frac{1}{1+t^{\frac{1}{p}}}dt$$

$$\leq \frac{(4c_{p,\varepsilon}^2)^p}{1-p}\sigma^{-(1+\varepsilon)d}\lambda^{-p} := C_2\sigma^{-(1+\varepsilon)d}\lambda^{-p},$$

and

$$Tr(L_{K_\sigma}) \leq 4c_{p,\varepsilon}^2\sigma^{-\frac{(1+\varepsilon)d}{p}}\sum_{i\geq 1} i^{-\frac{1}{p}} \leq \frac{4c_{p,\varepsilon}^2}{1-p}\sigma^{-\frac{(1+\varepsilon)d}{p}} := C_3\sigma^{-\frac{(1+\varepsilon)d}{p}}.$$

For $\mathcal{S}(\sigma,\lambda)$ and β_u, we get by assumption (7) and (6),

$$\begin{aligned}
\mathcal{S}(\sigma,\lambda) &= \sum_{i=1}^{\infty}\frac{\|\phi_i\|_\infty^2}{1+\lambda/\lambda_i} \leq C_1\sigma^{-\nu}\sum_{i=1}^{\infty}\frac{\lambda_i^{-s}}{1+\lambda/\lambda_i} = C_1\sigma^{-\nu}\lambda^{-s}\sum_{i=1}^{\infty}\frac{(\lambda/\lambda_i)^s}{1+\lambda/\lambda_i}\\
&\leq C_1\sigma^{-\nu}\lambda^{-s}\sum_{i=1}^{\infty}\frac{1}{(1+\lambda/\lambda_i)^{1-s}} \leq C_1\sigma^{-\nu}\lambda^{-s}\sum_{i=1}^{\infty}\frac{1}{(1+4^{-1}c_{p,\varepsilon}^{-2}\sigma^{\frac{(1+\varepsilon)d}{p}}\lambda i^{\frac{1}{p}})^{1-s}}\\
&\leq C_1\sigma^{-\nu}\lambda^{-s}\int_0^{\infty}\frac{1}{(1+4^{-1}c_{p,\varepsilon}^{-2}\sigma^{\frac{(1+\varepsilon)d}{p}}\lambda t^{\frac{1}{p}})^{1-s}}dt\\
&= C_1\sigma^{-\nu}\lambda^{-s}\left(4^p c_{p,\varepsilon}^{2p}\sigma^{-(1+\varepsilon)d}\lambda^{-p}\right)\int_0^{\infty}\frac{1}{\left(1+t^{\frac{1}{p}}\right)^{1-s}}dt\\
&\leq C_1 4^p c_{p,\varepsilon}^{2p}\left(\frac{1-s}{1-p-s}\right)\lambda^{-(p+s)}\sigma^{-(1+\varepsilon)d-\nu} := C_4\lambda^{-(p+s)}\sigma^{-(1+\varepsilon)d-\nu},
\end{aligned}$$

and for $u\in\mathbb{N}$,

$$\begin{aligned}
\beta_u &\leq C_1\sigma^{-\nu}\sum_{j=u+1}^{\infty}\lambda_j^{1-s} \leq C_1 4^{1-s}c_{p,\varepsilon}^{2(1-s)}\sigma^{-\frac{(1-s)(1+\varepsilon)d+p\nu}{p}}\sum_{j=u+1}^{\infty} i^{-\frac{1-s}{p}}\\
&\leq C_1 4^{1-s}c_{p,\varepsilon}^{2(1-s)}\left(\frac{p}{1-p-s}\right)\sigma^{-\frac{(1-s)(1+\varepsilon)d+p\nu}{p}}u^{-\frac{1-p-s}{p}}\\
&:= C_5\sigma^{-\frac{(1-s)(1+\varepsilon)d+p\nu}{p}}u^{-\frac{1-p-s}{p}}.
\end{aligned}$$

The proof is complete. □

Proof of Theorem 3. By checking the proof of Theorem 6 in [28] when $\frac{d\rho_X}{dx}\in L^\infty$, we know that $\mathcal{D}(\sigma,\lambda)\leq c(\sigma^{2\alpha}+\lambda\sigma^{-d})$ for some constant c independent of σ and λ. The choice of $\sigma = N^{-\frac{1}{2\alpha+d}}$ and $\lambda = N^{-1}$ yields that $\mathcal{D}(\sigma,\lambda)\leq 2cN^{-\frac{2\alpha}{2\alpha+d}}$ and $\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 \leq \mathcal{D}(\sigma,\lambda)/\lambda \leq 2cN^{\frac{d}{2\alpha+d}}$. We first estimate the bias bound (34).

Take the integer $u = N^t$ with t being the smallest integer greater than or equal to $\frac{(2-s)(1+\varepsilon)d+p(\nu+2\alpha)+2p(2\alpha+d)}{(2\alpha+d)(1-p-s)}$, then by (6), (40) and (42),

$$(4Tr(L_{K_\sigma})\beta_u/\lambda + 2\lambda_{u+1})\left(B^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right) \leq C_6 N^{-\frac{2\alpha}{2\alpha+d}}$$

with $C_6 = 16(C_3C_5 + 2c_{p,\varepsilon}^2)(B^2+2c)$.

We proceed to the quantity Δ defined in (25). With (41) and (40), we also see that

$$4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda) + \mathcal{S}(\sigma,\lambda)/3 \leq 8C_2C_4\sigma^{-2(1+\varepsilon)d-\nu}\lambda^{-2p-s} \leq 8C_2C_4N^{\frac{2d+\nu}{2\alpha+d}+s}N^{\frac{2\varepsilon d}{2\alpha+d}+2p}.$$

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

Thus,

$$\begin{aligned} & u\exp\left\{-\frac{n/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\}\left(B^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right) \\ & \le (B^2+2c)N^{t+1}\exp\left\{-\frac{N/(8m)}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\} \\ & \le (B^2+2c)N^{t+1}\exp\left\{-\frac{N^{1-\frac{2d+\nu}{2\alpha+d}-s}N^{-\frac{2\varepsilon d}{2\alpha+d}-2p}}{64C_2C_4m}\right\}. \end{aligned}$$

Denote $\xi := \frac{\varepsilon d}{2\alpha+d}+p$ and take the values of p and ε to be small enough such that $\xi < \frac{1}{2}\left(1-\frac{2d+\nu}{2\alpha+d}-s\right)$. This together with the restriction (11) of m yields that

$$\exp\left\{-\frac{N^{1-\frac{2d+\nu}{2\alpha+d}-s}N^{-\frac{2\varepsilon d}{2\alpha+d}-2p}}{64C_2C_4m}\right\} = \exp\left\{-\frac{N^{1-\frac{2d+\nu}{2\alpha+d}-s}N^{-2\xi}}{64C_2C_4m}\right\} \le \exp\left\{-\frac{N^{\frac{1}{2}\left(1-\frac{2d+\nu}{2\alpha+d}-s\right)-\xi}}{64C_2C_4}\right\}.$$

It implies that we can find a constant C_7 (depending on $\alpha, d, s, \nu, p, \varepsilon$) independent of N such that

$$\begin{aligned} & u\exp\left\{-\frac{n/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\}\left(B^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right) \\ & \le (B^2+2c)N^{t+1}\exp\left\{-\frac{N^{\frac{1}{2}\left(1-\frac{2d+\nu}{2\alpha+d}-s\right)-\xi}}{64C_2C_4}\right\} \le C_7mN^{-\frac{2\alpha}{2\alpha+d}}. \end{aligned}$$

So, we get that

$$\left(3Tr(L_{K_\sigma})\beta_u/\lambda+u\exp\left\{-\frac{n/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\}\right)\left(B^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right) \le (C_6+C_7)mN^{-\frac{2\alpha}{2\alpha+d}}.$$

Meanwhile,

$$\begin{aligned} \frac{24\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+4B^2\mathcal{N}(\sigma,\lambda)\right]}{n} &= \frac{24m\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+4B^2\mathcal{N}(\sigma,\lambda)\right]}{N} \\ &\le 24m\left(2cC_4N^{-\frac{2\alpha}{2\alpha+d}+s+\frac{d+\nu}{2\alpha+d}-1}N^{p+\frac{\varepsilon d}{2\alpha+d}}+4B^2C_2N^{-\frac{2\alpha}{2\alpha+d}}N^{p+\frac{\varepsilon d}{2\alpha+d}}\right) \\ &\le (48cC_4+96B^2C_2)mN^{-\frac{2\alpha}{2\alpha+d}+\xi+\max\left\{\frac{d+\nu}{2\alpha+d}+s-1,0\right\}}. \end{aligned}$$

Then, the quantity Δ in (25) is bounded as

$$\begin{aligned} \Delta &\le m\left[\frac{12\mathcal{D}(\sigma,\lambda)}{m}+(48cC_4+96B^2C_2)N^{-\frac{2\alpha}{2\alpha+d}+\xi+\max\left\{\frac{d+\nu}{2\alpha+d}+s-1,0\right\}}+8(C_6+C_7)N^{-\frac{2\alpha}{2\alpha+d}}\right] \\ &\le \left(24c+48cC_4+96B^2C_2+8(C_6+C_7)\right)mN^{-\frac{2\alpha}{2\alpha+d}+\xi+\max\left\{\frac{d+\nu}{2\alpha+d}+s-1,0\right\}} \\ &:= C_8mN^{-\frac{2\alpha}{2\alpha+d}+\xi+\max\left\{\frac{d+\nu}{2\alpha+d}+s-1,0\right\}}. \end{aligned} \tag{43}$$

Using (41), (39) and the restriction (11) again, we have

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

$$\begin{aligned}
&\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{n}+\frac{64\mathcal{S}^2(\sigma,\lambda)}{n^2}\right](\log u)^2\\
&\le t^2\left[\frac{C_2C_4m\sigma^{-2(1+\varepsilon)d-\nu}\lambda^{-2p-s}}{N}+\frac{64C_2^2m^2\sigma^{-2(1+\varepsilon)d-2\nu}\lambda^{-2p-2s}}{N^2}\right](\log N)^2\\
&= t^2\left[\frac{C_2C_4mN^{\frac{2(1+\varepsilon)d+\nu}{2\alpha+d}+2p+s}}{N}+\frac{64C_2^2m^2N^{\frac{2(1+\varepsilon)d+2\nu}{2\alpha+d}+2p+2s}}{N^2}\right](\log N)^2\\
&= t^2\left[\frac{C_2C_4mN^{\frac{2d+\nu}{2\alpha+d}+s+2\xi}}{N}+\frac{64C_2^2m^2N^{\frac{2d+2\nu}{2\alpha+d}+2s+2\xi}}{N^2}\right](\log N)^2\\
&\le (C_2C_4+64C_2^2)t^2/m.
\end{aligned}$$

Putting the above bound and (43) into the bound (34), we have that

$$\begin{aligned}
\|\mathbb{E}\left[f_{D_l}-f_\lambda^\sigma\right]\|_\rho^2 &\le C_6N^{-\frac{2\alpha}{2\alpha+d}}+128(C_2C_4+64C_2^2)C_8t^2N^{-\frac{2\alpha}{2\alpha+d}+\xi+\max\left\{\frac{d+\nu}{2\alpha+d}+s-1,0\right\}}\\
&\le \left(C_6+128(C_2C_4+64C_2^2)C_8t^2\right)N^{-\frac{2\alpha}{2\alpha+d}+\xi+\min\left\{\frac{d+\nu}{2\alpha+d}+s-1,0\right\}}. \qquad (44)
\end{aligned}$$

Also, collecting the above analysis, with the bound (24), we have that

$$\mathbb{E}\left[\|f_{D_l}-f_\lambda^\sigma\|_\rho^2\right] \le (C_8+C_6)mN^{-\frac{2\alpha}{2\alpha+d}+\xi+\max\left\{\frac{d+\nu}{2\alpha+d}+s-1,0\right\}}. \qquad (45)$$

Combining Proposition 1 with (44) and (45), we get that

$$\mathbb{E}\left[\|\bar{f}_D-f_\lambda^\sigma\|_\rho^2\right] \le \left(2C_6+C_8+128(C_2C_4+64C_2^2)C_8t^2\right)N^{-\frac{2\alpha}{2\alpha+d}+\xi+\max\left\{\frac{d+\nu}{2\alpha+d}+s-1,0\right\}}.$$

This together with (15) yields our statement of Theorem 3 by taking

$$C = 2(2C_6+C_8+128(C_2C_4+64C_2^2)C_8t^2)+4c.$$

The proof is complete. □

Proof of Theorem 2. Following the above proof, we apply the bounds (24) and (34) to prove the statement in Theorem 2. We first consider the variance bound (24). The restriction $\frac{N}{m\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)} \ge C_0N^\eta$ for $\eta>0$ implies that $\exp\left\{-\frac{n/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\} = \exp\left\{-\frac{N/(8m)}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\} = O\left(\exp\{-C'N^\eta\}\right)$ for some constant $C'>0$ (independent of N). We also note that the quantities parameterized by u on the right-hand side of (24) decrease as the value of u increases. We can choose $u=N^t$ with t being the smallest integer greater than or equal to $\frac{\log\left(\sigma^{(p+s-2)(1+\varepsilon)d}\lambda^{p((p+s-2))}n^p/\mathcal{D}^p(\sigma,\lambda)\right)}{(1-p-s)\log N}$. So, the leading term of the bound (24) is

$$12\lambda\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2+\frac{24\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+4B^2\mathcal{N}(\sigma,\lambda)\right]}{n}.$$

By the similar idea, we can also get that the leading term of (34) is

$$128\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{n}+\frac{64\mathcal{S}^2(\sigma,\lambda)}{n^2}\right](\log N)^2\left\{12\lambda\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2+\frac{24\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+4B^2\mathcal{N}(\sigma,\lambda)\right]}{n}\right\}.$$

Collecting the above analysis, we get

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

$$\mathbb{E}\left[\|\bar{f}_D - f_\lambda^\sigma\|_\rho^2\right]$$
$$\leq C\left\{\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{n} + \frac{\mathcal{S}^2(\sigma,\lambda)}{n^2}\right](\log N)^2 + \frac{1}{m}\right\}\left\{\lambda\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 + \frac{[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda) + \mathcal{N}(\sigma,\lambda)]}{n}\right\}$$

where C is a constant independent of N or m. This bound with (15) and $\lambda\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 \leq \mathcal{D}(\sigma,\lambda)$ yields our statement in Theorem 2. □

Next, we prove Theorem 1 by Theorem 3. To check assumption (7) in Theorem 3, we need the following lemma.

Lemma 3. *Assume that $\rho_{\mathcal{X}}$ has a probability density function away from 0 and ∞. Let $\{(\lambda_i,\phi_i)\}_{i\geq 1}$ be the normalized eigenpairs of the operator $L_{K_\sigma}: L^2_{\rho_{\mathcal{X}}} \to L^2_{\rho_{\mathcal{X}}}$. Then for all $i \geq 1$ and any $r \in \mathbb{N}$, there holds*

$$\|\phi_i\|_\infty^2 \leq C_1'\sigma^{-d}\lambda_i^{-\frac{d}{2r}}, \tag{46}$$

where C_1' is a constant independent of λ_i, σ (to be given in the proof).

Proof. When $\rho_{\mathcal{X}}$ is the normalized uniform distribution on $\mathcal{X}$, we know by [1] (page 230) that for all $0 < s < 1$,

$$[L^2_{\rho_{\mathcal{X}}}, W_2^r(\mathcal{X})]_{s,1} = B^{sr}_{2,1}(\mathcal{X})$$

where $W_2^r(\mathcal{X})$ denotes the Sobolev space of integer order, $B^{sr}_{2,1}(\mathcal{X})$ a Besov space and $[L^2_{\rho_{\mathcal{X}}}, W_2^r(\mathcal{X})]_{s,1}$ an interpolation space between $L^2_{\rho_{\mathcal{X}}}$ and $W_2^r(\mathcal{X})$. It implies that for some constant c' depending on r, d, s, there holds for any $f \in W_2^r(\mathcal{X})$

$$\|f\|_{B^{sr}_{2,1}(\mathcal{X})} \leq c'\|f\|^s_{W_2^r(\mathcal{X})}\|f\|^{1-s}_{L^2_{\rho_{\mathcal{X}}}}. \tag{47}$$

Since the Gaussian kernel K_σ is C^∞ and $\mathcal{X}$ is compact, $\mathcal{H}_\sigma$ can be embedded into the Sobolev space $W_2^r(\mathcal{X})$ with an arbitrarily fixed $r \in \mathbb{N}$. It implies that the relation (47) holds for any $f \in \mathcal{H}_\sigma$. Moreover, take $s = \frac{d}{2r}$, $B^{sr}_{2,1}(\mathcal{X})$ can be continuously embedded into $\ell_\infty(\mathcal{X})$ ([1], Theorem 7.34). This together with (47) yields that for any $f \in \mathcal{H}_\sigma$,

$$\|f\|_\infty \leq c\|f\|^{\frac{d}{2r}}_{W_2^r(\mathcal{X})}\|f\|^{1-\frac{d}{2r}}_{L^2_{\rho_{\mathcal{X}}}} \tag{48}$$

where c is a constant depending only on r, d.
Furthermore, by Theorem 4.48 in [20], we know that there exists a constant $c_{r,d,\mathcal{X}}$ depending on r, d and the volume of $\mathcal{X}$ such that for any $f \in \mathcal{H}_\sigma$,

$$\|f\|_{W_2^r(\mathcal{X})} \leq c_{r,d,\mathcal{X}}\sigma^{-r}\|f\|_{\mathcal{H}_\sigma}. \tag{49}$$

Note that $\|\phi_i\|_{L^2_{\rho_{\mathcal{X}}}} = 1$ for $i \geq 1$ and $\{\sqrt{\lambda_i}\phi_i\}_i$ forms an orthonormal basis in $\mathcal{H}_\sigma$. Putting the above bounds (48) and (49) together with $f = \phi_i$, we have

$$\|\phi_i\|_\infty^2 \leq c^2\left(\|\phi_i\|^2_{W_2^r(\mathcal{X})}\right)^{\frac{d}{2r}} \leq c^2c^2_{r,d,\mathcal{X}}\sigma^{-d}\left(\|\phi_i\|^2_{\mathcal{H}_\sigma}\right)^{\frac{d}{2r}} = c^2c^2_{r,d,\mathcal{X}}\sigma^{-d}\lambda_i^{-\frac{d}{2r}},\ \forall\ i \geq 1.$$

By taking $C_1' := c^2c^2_{r,d,\mathcal{X}}$, we can get the bound (46) when $\rho_{\mathcal{X}}$ is the normalized uniform distribution on $\mathcal{X}$. Note that the above proof is also applicable to a distribution whose probability density function is away from 0 and ∞.

The proof is complete. □

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

We are now in a position to prove Theorem 1.

Proof of Theorem 1. According to Lemma 3, we know when $\rho_{\mathcal{X}}$ has a probability density function away from 0 and ∞, assumption (7) holds with $s = \frac{d}{2r}$ and $\nu = d$ for any $r \in \mathbb{N}$. Here let r be large enough. Then we can get the statement of Theorem 1 by applying Theorem 3.

The proof is complete. □

5. Proofs in semi-supervised learning

This section is devoted to proving the main results stated in Subsection 2.3, which shows the improved performance of distributed regularized least squares using only unlabeled data. According to (15) and Proposition 1 again, we need to estimate the bias bound $\left\|\mathbb{E}\left[f_{\tilde{D}_l} - f_\lambda^\sigma\right]\right\|_\rho^2$ and the variance bound $\mathbb{E}\left[\left\|f_{\tilde{D}_l} - f_\lambda^\sigma\right\|_\rho^2\right]$ with the semi-supervised data set $\tilde{D}_l$. At a high-level, the proofs in semi-supervised learning are similar to those in supervised learning. So, we only outline some necessary proof procedures and estimates that are used in the proofs.

As in supervised learning, we denote $\tilde{E} = f_{\tilde{D}_l} - f_\lambda^\sigma = \tilde{E} = \sum_{i=1}^\infty \tilde{e}_i \phi$ with $\tilde{\mathbf{e}} = \{\tilde{e}_i\}_i \in \ell^2$. Fix an integer $u \in \mathbb{N}$ to be determined, and decompose the vector $\tilde{\mathbf{e}}$ into two parts $\tilde{\mathbf{e}}^{\mathbf{1}} = \{\tilde{e}_1, \cdots, \tilde{e}_u\}$ and $\tilde{\mathbf{e}}^{\mathbf{2}} = \{\tilde{e}_{u+1}, \tilde{e}_{u+2}, \cdots\}$. It implies that

$$\mathbb{E}\left[\left\|f_{\tilde{D}_l} - f_\lambda^\sigma\right\|_\rho^2\right] = \mathbb{E}[\|\tilde{\mathbf{e}}\|_2^2] = \mathbb{E}[\|\tilde{\mathbf{e}}^{\mathbf{1}}\|_2^2] + \mathbb{E}[\|\tilde{\mathbf{e}}^{\mathbf{2}}\|_2^2].$$

Define the vector $\tilde{\mathbf{v}} = [\tilde{v}_1, \cdots, v_n]^{\mathrm{T}}$ by $\tilde{v}_i = \sum_{j=u+1}^\infty \tilde{e}_j \phi_j(x_i), i = 1, \cdots, \tilde{n}$. Let the matrix $\tilde{\Phi} = (\tilde{\Phi}_{ij})_{i,j} \in \mathbb{R}^{\tilde{n}\times u}$ with $\tilde{\Phi}_{ij} = (\phi_j(x_i))$ for $i \in \{1, \cdots, \tilde{n}\}$ and $j \in \{1, \cdots, u\}$. Denote $\tilde{\epsilon}_i = f_\lambda^\sigma(\tilde{x}_i) - \tilde{y}_i, i = 1, \cdots, \tilde{n}$ and $\tilde{\epsilon} = [\tilde{\epsilon}_1, \cdots, \tilde{\epsilon}_{\tilde{n}}]^T$.

By tracing the proof in Proposition 5, we know from (27) and (32) that

$$\mathbb{E}[\|\tilde{\mathbf{e}}^{\mathbf{2}}\|_2^2] \le 2\lambda_{u+1}(\tilde{B}^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2) \tag{50}$$

and

$$\begin{aligned}\mathbb{E}[\|\tilde{\mathbf{e}}^{\mathbf{1}}\|_2^2] \le \mathbb{E}[\|Q\tilde{\mathbf{e}}^{\mathbf{1}}\|_2^2] \le 12\left\|\lambda Q^{-1}M^{-1}\mathbf{a}^{\mathbf{1}}\right\|_2^2 + 12\mathbb{E}\left[\left\|\frac{1}{\tilde{n}}Q^{-1}\tilde{\Phi}^{\mathrm{T}}\tilde{\mathbf{v}}\right\|_2^2\right] \\ + 12\mathbb{E}\left[\left\|\frac{1}{\tilde{n}}Q^{-1}\tilde{\Phi}^{\mathrm{T}}\tilde{\epsilon}\right\|_2^2\right] + 4\mathbb{P}\left(\tilde{A}^c\right)\left(\tilde{B}^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right),\end{aligned} \tag{51}$$

where $\tilde{A} := \left\{\left\|Q^{-1}\left(\frac{1}{\tilde{n}}\tilde{\Phi}^{\mathrm{T}}\tilde{\Phi} - I\right)Q^{-1}\right\| \le \frac{1}{2}\right\}$.

We now present a proposition regarding the terms on the right-hand side of (51).

Proposition 8. *The following bounds hold*

$$\mathbb{P}\left(\tilde{A}^c\right) \le 2u\exp\left\{-\frac{\tilde{n}/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda) + \mathcal{S}(\sigma,\lambda)/3}\right\}, \tag{52}$$

$$\mathbb{E}\left[\left\|\frac{1}{\tilde{n}}Q^{-1}\tilde{\Phi}^{\mathrm{T}}\tilde{\mathbf{v}}\right\|_2^2\right] \le 2Tr(L_{K_\sigma})\beta_u\left(\tilde{B}^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right)/\lambda, \tag{53}$$

and

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

$$\mathbb{E}\left[\left\|\frac{1}{\tilde{n}}Q^{-1}\tilde{\Phi}^{\mathrm{T}}\tilde{\epsilon}\right\|_2^2\right] \le \frac{2}{\tilde{n}}\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+2B^2\mathcal{N}(\sigma,\lambda)\right]+\frac{4B^2\mathcal{N}(\sigma,\lambda)}{n}. \tag{54}$$

Proof. The bounds (52) and (53) can be obtained respectively by following the bound of $\mathbb{P}\left(A^c\right)$ in the proof of Proposition 5 and the estimate in (22). But for bound (54), we should notice that $\tilde{\epsilon}_i \neq \epsilon_i$ when $1 \le i \le n$. So, we cannot derive (54) directly by the proof of (23). Note by (63) in the appendix that

$$\mathbb{E}\left[\left\|\frac{1}{\tilde{n}}Q^{-1}\tilde{\Phi}^{\mathrm{T}}\tilde{\epsilon}\right\|_2^2\right] = \frac{1}{\tilde{n}^2}\sum_{k=1}^{u}\sum_{i=1}^{\tilde{n}}\frac{\mathbb{E}[\phi_k^2(\tilde{x}_i)\tilde{\epsilon}_i^2]}{1+\lambda/\lambda_k}.$$

When $1 \le i \le n$, we have

$$\begin{aligned}
\frac{\mathbb{E}[\phi_k^2(\tilde{x}_i)\tilde{\epsilon}_i^2]}{1+\lambda/\lambda_k} &\le 2\left(\frac{\mathbb{E}\left[\phi_k^2(\tilde{x}_i)(f_\lambda^\sigma(\tilde{x}_i)-f_\rho(\tilde{x}_i))^2\right]}{1+\lambda/\lambda_k}+\frac{\mathbb{E}\left[\phi_k^2(x_i)(f_\rho(\tilde{x}_i)-\tilde{y}_i)^2\right]}{1+\lambda/\lambda_k}\right)\\
&\le 2\left(\frac{\|\phi_k\|_\infty^2\mathbb{E}\left[(f_\lambda^\sigma(\tilde{x}_i)-f_\rho(\tilde{x}_i))^2\right]}{1+\lambda/\lambda_k}+\frac{2\mathbb{E}\left[\phi_k^2(\tilde{x}_i)\mathbb{E}^*\left[|f_\rho(\tilde{x}_i)|^2+|\tilde{y}_i|^2\right]\right]}{1+\lambda/\lambda_k}\right)\\
&\le 2\left(\frac{\|\phi_k\|_\infty^2\mathbb{E}\left[(f_\lambda^\sigma(\tilde{x}_i)-f_\rho(\tilde{x}_i))^2\right]}{1+\lambda/\lambda_k}+\frac{2B^2+2\mathbb{E}\left[\phi_k^2(\tilde{x}_i)\mathbb{E}^*\left[|\tilde{y}_i|^2\right]\right]}{1+\lambda/\lambda_k}\right)\\
&\le 2\left(\frac{\|\phi_k\|_\infty^2\mathcal{D}(\sigma,\lambda)}{1+\lambda/\lambda_k}+\frac{2B^2(1+\tilde{n}^2/n^2)}{1+\lambda/\lambda_k}\right).
\end{aligned}$$

When $n+1 \le i \le \tilde{n}$, by $\tilde{y}_i = 0$, we have that

$$\frac{\mathbb{E}[\phi_k^2(\tilde{x}_i)\tilde{\epsilon}_i^2]}{1+\lambda/\lambda_k} \le 2\left(\frac{\|\phi_k\|_\infty^2\mathcal{D}(\sigma,\lambda)}{1+\lambda/\lambda_k}+\frac{2B^2}{1+\lambda/\lambda_k}\right).$$

So,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{\tilde{n}}Q^{-1}\tilde{\Phi}^{\mathrm{T}}\tilde{\epsilon}\right\|_2^2\right] \le \frac{2}{\tilde{n}^2}\sum_{k=1}^{u}\sum_{i=1}^{\tilde{n}}\left[\frac{\|\phi_k\|_\infty^2\mathcal{D}(\sigma,\lambda)}{1+\lambda/\lambda_k}+\frac{2B^2}{1+\lambda/\lambda_k}\right]+\frac{2}{\tilde{n}^2}\sum_{k=1}^{u}\sum_{i=1}^{n}\left(\frac{2B^2\tilde{n}^2/n^2}{1+\lambda/\lambda_k}\right)\\
&=\frac{2}{\tilde{n}}\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+2B^2\mathcal{N}(\sigma,\lambda)\right]+\frac{2}{n^2}\sum_{k=1}^{u}\sum_{i=1}^{n}\left(\frac{2B^2}{1+\lambda/\lambda_k}\right)\\
&=\frac{2}{\tilde{n}}\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+2B^2\mathcal{N}(\sigma,\lambda)\right]+\frac{4B^2\mathcal{N}(\sigma,\lambda)}{n}.
\end{aligned}$$

The proof is complete. □

With this proposition in place, we can get the variance bound in semi-supervised learning.

Proposition 9. *Define $f_{\tilde{D}_l}$ by (2) with $\tilde{D}_l$. Then we have*

$$\mathbb{E}\left[\left\|f_{\tilde{D}_l}-f_\lambda^\sigma\right\|_\rho^2\right] \le \tilde{\Delta}+2\lambda_{u+1}\left[\tilde{n}B^2/(n\lambda)+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right], \tag{55}$$

where u is an integer, and

$$\tilde{\Delta} := 12\lambda\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2 + \frac{24\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda)+2B^2\mathcal{N}(\sigma,\lambda)\right]}{\tilde{n}} + \frac{48B^2\mathcal{N}(\sigma,\lambda)}{n}$$
$$+ 8\left(3Tr(L_{K_\sigma})\beta_u/\lambda + u\exp\left\{-\frac{\tilde{n}/8}{4\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)+\mathcal{S}(\sigma,\lambda)/3}\right\}\right)\left[\tilde{n}B^2/(n\lambda)+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right]. \quad (56)$$

Proof. Notice that $\mathbb{E}\left[\left\|f_{\tilde{D}_l,\lambda}-f_\lambda^\sigma\right\|_\rho^2\right] \le \mathbb{E}[\|Q\tilde{\mathbf{e}}^{\mathbf{1}}\|_2^2]+\mathbb{E}[\|\tilde{\mathbf{e}}^{\mathbf{2}}\|_2^2]$. Combining (50), (51) with Proposition 8 and (21) yields the conclusion (55). □

Next we get the bias bound in semi-supervised learning.

Proposition 10. *Define $\bar{f}_{\tilde{D}_l}$ by (2) with $\tilde{D}_l$. Then we have*

$$\left\|\mathbb{E}\left[f_{\tilde{D}_l}-f_\lambda^\sigma\right]\right\|_\rho^2 \le \left(4Tr(L_{K_\sigma})\beta_u/\lambda+2\lambda_{u+1}\right)\left[\tilde{n}B^2/(n\lambda)+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right]$$
$$+ 128\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{\tilde{n}}+\frac{64\mathcal{S}^2(\sigma,\lambda)}{\tilde{n}^2}\right](\log(u))^2\tilde{\Delta}, \quad (57)$$

where $\tilde{\Delta}$ is defined in Proposition 9.

Proof. Notice that

$$\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}}\mathbb{E}\left[\tilde{\epsilon}_i\phi_k(\tilde{x}_i)\right] = \mathbb{E}\left[(f_\lambda^\sigma(x)-y)\phi_k(x)\right] = -\frac{\lambda a_k}{\lambda_k}.$$

Then we get

$$\mathbb{E}\left[\frac{1}{\tilde{n}}\tilde{\Phi}^T\tilde{\epsilon}\right] = -\lambda M^{-1}\mathbf{a}^{\mathbf{1}}.$$

Following the same proof procedures as those in the proof of Proposition 6, we have by (38)

$$\left\|\mathbb{E}\left[f_{\tilde{D}_l}-f_\lambda^\sigma\right]\right\|_\rho^2 \le 2\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\tilde{\Phi}^{\mathrm{T}}\tilde{\mathbf{v}}\right\|_2^2\right]$$
$$+ 128\left[\frac{\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda)}{\tilde{n}}+\frac{64\mathcal{S}^2(\sigma,\lambda)}{\tilde{n}^2}\right](\log(u))^2\mathbb{E}\left[\left\|Q\tilde{\mathbf{e}}^{\mathbf{1}}\right\|_2^2\right]+2\lambda_{u+1}(\tilde{B}^2/\lambda+\|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2).$$

This together with (51) and (53) yields the conclusion. □

Based on Proposition 10 and Proposition 9, we now prove Theorem 4.

Proof of Theorem 4. Here we choose $u=\tilde{N}^t$ with t being the smallest integer greater than or equal to $\frac{(2-s)(1+\varepsilon)d+p(\nu+2\alpha)+(1+p+s)(2\alpha+d)}{(2\alpha+d)(1-p-s)}$. The remainder of the proof is very similar to that of Theorem 3. We omit it for simplicity. This completes the proof. □

Proof of Corollary 1. Note by (46) that assumption (7) holds with $\nu=d$ and an arbitrarily small $s>0$ in this case. Then the statement of Corollary 1 can be derived directly from Theorem 4. □

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

Appendix A. Useful Lemmas

In this appendix, we give some useful lemmas in proving our main results. The first one is regarding the mini-max optimal rate in the regression setting, which can be found as Theorem 2.2 in [25] and Theorem 13 in [21].

Lemma 4. *Let Θ be a subset of $L^2_{\rho_{\mathcal{X}}}$ such that all $f \in \Theta$ is uniformly bounded. If for some $\eta \in (0,1)$, the entropy numbers satisfy*

$$e_i(\Theta, L^2_{\rho_{\mathcal{X}}}) = O\left(i^{-\frac{1}{\eta}}\right),$$

then there exist constants $c_0 > 0, c_1, c_2 > 0$ and a sequence $\{\delta_n\}$ with $\delta_n \sim n^{-\frac{2}{2+\eta}}$ such that when $f_\rho \in \Theta$, $\delta > 0$ and $n \geq 1$, there holds

$$\mathbb{P}\left\{D : \|f_D - f_\rho\|^2_{L^2_{\rho_{\mathcal{X}}}} \geq \delta\right\} \geq \begin{cases} c_0, & \text{if} \quad \delta \leq \delta_n, \\ c_1 \exp\{-c_2 \delta n\}, & \text{if} \quad \delta \geq \delta_n, \end{cases}$$

where f_D is a prediction function based on a given data set D.

The second lemma found in [26] provides a matrix concentration inequality that is used to bound spectral norms of sums of independent, random, symmetric matrices.

Lemma 5. *Consider a finite sequence $\{X_k\}_k$ of independent, random, symmetric matrices with dimension u. Assume that*

$$\mathbb{E}[X_k] = 0, \quad \|X_k\| \leq R \quad \text{almost} \quad \text{surely},$$

and the norm of the total variance $\Sigma^2 := \left\| \sum_k \mathbb{E}\left[X_k^2\right] \right\| < \infty$, then the following inequality holds for every $t > 0$,

$$\begin{aligned} \mathbb{P}\left\{ \left\| \sum_k X_k \right\| \geq t \right\} &\leq 2u \exp\left\{ -\frac{t^2/2}{\Sigma^2 + Rt/3} \right\} \\ &\leq \begin{cases} 2u \exp\left\{-\frac{t^2}{8\Sigma^2}\right\}, & \text{for} \quad t \leq \Sigma^2/R, \\ 2u \exp\left\{-\frac{t}{8R}\right\}, & \text{for} \quad t \geq \Sigma^2/R. \end{cases} \end{aligned} \tag{58}$$

With the help of the matrix concentration inequality above, we can develop an upper bound for the expectation of the second moment $\left\|\sum_k X_k\right\|^2$, that is, if $u \geq 4e$,

$$\mathbb{E}\left[\left\| \sum_k X_k \right\|^2 \right] \leq 16(\Sigma^2 + 64R^2)(\log(u))^2. \tag{59}$$

Proof of the expectation estimate (59). Using the formula $\mathbb{E}[\xi] = \int_0^\infty \mathbb{P}\{\xi > t\} dt$ with $\xi = \left\|\sum_k X_k\right\|^2$, we get by (58) that

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

$$\mathbb{E}\left[\left\|\sum_k X_k\right\|^2\right] = \int_0^\infty \mathbb{P}\left\{\left\|\sum_k X_k\right\|^2 > t\right\} dt = \int_0^\infty \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt$$

$$= \left(\int_{\left\{t^{\frac{1}{2}} \le \Sigma^2/R\right\}} + \int_{\left\{t^{\frac{1}{2}} \ge \Sigma^2/R\right\}}\right) \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt$$

$$:= I_1 + I_2.$$

First, we estimate I_1 as

$$I_1 = \int_{\left\{t^{\frac{1}{2}} \le \Sigma^2/R\right\}} \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt$$

$$= \int_{\{8\Sigma^2 \log(2u) \le t \le (\Sigma^2/R)^2\}} \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt + \int_{\{t < 8\Sigma^2 \log(2u)\}} \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt$$

$$\le \int_{\{8\Sigma^2 \log(2u) \le t \le (\Sigma^2/R)^2\}} \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt + 8\Sigma^2 \log(2u).$$

Obviously, if $8\Sigma^2 \log(2u) \ge \left(\Sigma^2/R\right)^2$, we have $I_1 \le 8\Sigma^2 \log(2u)$. Otherwise,

$$\int_{\{8\Sigma^2 \log(2u) \le t \le (\Sigma^2/R)^2\}} \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt \le 2u \int_{\{8\Sigma^2 \log(2u) \le t \le (\Sigma^2/R)^2\}} \exp\left\{-\frac{t}{8\Sigma^2}\right\} dt$$

$$\le 2u \int_{\{t \ge 8\Sigma^2 \log(2u)\}} \exp\left\{-\frac{t}{8\Sigma^2}\right\} dt = 8\Sigma^2.$$

It follows that

$$I_1 \le 8\Sigma^2 \log(2u) + 8\Sigma^2 = 8\Sigma^2 \log(2eu). \tag{60}$$

For I_2, we have

$$I_2 = \int_{\left\{t^{\frac{1}{2}} \ge \Sigma^2/R\right\}} \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt$$

$$\le \int_{\left\{t^{\frac{1}{2}} \ge \Sigma^2/R, t^{\frac{1}{2}} \ge 16R\log(4u)\right\}} \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt + \int_{\left\{t^{\frac{1}{2}} \le 16R\log(4u)\right\}} \mathbb{P}\left\{\left\|\sum_k X_k\right\| > t^{\frac{1}{2}}\right\} dt$$

$$\le 2u \int_{\left\{t^{\frac{1}{2}} \ge \Sigma^2/R, t^{\frac{1}{2}} \ge 16R\log(4u)\right\}} \exp\left\{-\frac{t^{\frac{1}{2}}}{8R}\right\} dt + \left(16R\log(4u)\right)^2$$

$$\leq 2u \int\limits_{\left\{t^{\frac{1}{2}} \geq 16R\log(4u)\right\}} \exp\left\{-\frac{t^{\frac{1}{2}}}{8R}\right\} dt + \Big(16R\log(4u)\Big)^2$$

$$= 4u \int\limits_{\{t \geq 16R\log(4u)\}} \exp\left\{-\frac{t}{8R}\right\} t dt + \Big(16R\log(4u)\Big)^2$$

$$= 4u\left[-8Rte^{-\frac{t}{8R}}\right]_{16R\log(4u)}^{\infty} + 4u \cdot 8R \int\limits_{16R\log(4u)}^{\infty} e^{-\frac{t}{8R}} dt + \Big(16R\log(4u)\Big)^2$$

$$\leq (16R)^2 + \Big(16R\log(4u)\Big)^2 \leq \Big(16R\log(4eu)\Big)^2. \tag{61}$$

Based on the above estimates (60) and (61) for I_1 and I_2, by $u \geq 4e$, we get

$$\mathbb{E}\left[\left\|\sum_k X_k\right\|^2\right] \leq 8\Sigma^2\log(2eu) + \Big(16R\log(4eu)\Big)^2 \leq (16\Sigma^2)\log(u) + (32R)^2(\log(u))^2$$
$$\leq 16(\Sigma^2 + 64R^2)(\log(u))^2.$$

Then the proof is complete. □

Appendix B. Proofs of estimates in Subsection 3.1

In this appendix we prove estimates stated in subsection 3.1.

Proof of Proposition 2. Note that

$$\lambda\|f_{D_l}\|_{\mathcal{H}_\sigma}^2 \leq \frac{1}{|D_l|}\sum_{(x,y)\in D_l}(f_{D_l}(x)-y)^2 + \lambda\|f_{D_l}\|_{\mathcal{H}_\sigma}^2 \leq \frac{1}{|D_l|}\sum_{(x,y)\in D_l} y^2.$$

Taking the conditional expectation $\mathbb{E}^*$ on the both sides of the above inequality, by (4), we have

$$\mathbb{E}^*[\lambda\|f_{D_l}\|_{\mathcal{H}_\sigma}^2] \leq \mathbb{E}^*[y^2] \leq B^2.$$

Also,

$$\mathbb{E}^*\left[\lambda\|f_{\tilde{D}_l}\|_{\mathcal{H}_\sigma}^2\right] \leq \mathbb{E}^*\left[\frac{1}{|\tilde{D}_l|}\sum_{(\tilde{x},\tilde{y})\in\tilde{D}_l}\tilde{y}^2\right] \leq \frac{\tilde{N}}{N}B^2.$$

So the stated estimates follow. The proof is complete. □

Proof of Proposition 3. For each $1 \leq i \leq n$, we denote $\pi_i = (\phi_1(x_i), \cdots, \phi_u(x_i))^{\mathrm{T}} \in \mathbb{R}^u$. Then we get that

$$Q^{-1}\left(\frac{1}{n}\Phi^{\mathrm{T}}\Phi - I\right)Q^{-1} = \frac{1}{n}\sum_{i=1}^{n} Q^{-1}\left(\pi_i\pi_i^{\mathrm{T}} - I\right)Q^{-1}.$$

We derive the bound (19) by Lemma 5. Let the random matrix sequence $\{X_i\}_{i=1}^n$ be

$$X_i := \frac{1}{n}Q^{-1}\left(\pi_i\pi_i^{\mathrm{T}} - I\right)Q^{-1}, \quad 1 \leq i \leq n.$$

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

It is easy to check that $\{X_i\}_{i=1}^n \subset R^{u\times u}$ are independent, symmetric matrixes and $\mathbb{E}[X_i] = 0, 1 \le i \le n$. Since $\pi_i\pi_i^{\mathrm{T}}$ is rank one and Q is diagonal, we have that

$$\frac{1}{n}\left\|Q^{-1}\pi_i\pi_i^{\mathrm{T}}Q^{-1}\right\| = \frac{1}{n}\pi_i^{\mathrm{T}}\left(I+\lambda M^{-1}\right)\pi_i = \frac{1}{n}\sum_{k=1}^{u}\frac{\phi_k(x_i)^2}{1+\lambda/\lambda_k}.$$

It tells us that

$$\|X_i\| \le \frac{1}{n}\left(\left\|Q^{-1}\pi_i\pi_i^{\mathrm{T}}Q^{-1}\right\| + \frac{1}{1+\lambda/\lambda_1}\right) \le \frac{2}{n}\sum_{k=1}^{u}\frac{\phi_k(x_i)^2}{1+\lambda/\lambda_k} \le \frac{2}{n}\mathcal{S}(\sigma,\lambda) := R.$$

We also find that for each $1 \le i \le n$,

$$\mathbb{E}\left[\|X_i^2\|\right] \le \mathbb{E}\left[\|X_i\|^2\right] \le \frac{4}{n^2}\mathbb{E}\left[\left(\sum_{k=1}^{u}\frac{\phi_k(x_i)^2}{1+\lambda/\lambda_k}\right)^2\right] \le \frac{4}{n^2}\sum_{k=1}^{\infty}\frac{\|\phi_k\|_\infty^2}{1+\lambda/\lambda_k}\mathbb{E}\left[\sum_{k=1}^{u}\frac{\phi_k(x_i)^2}{1+\lambda/\lambda_k}\right]$$
$$= \frac{4}{n^2}\sum_{k=1}^{\infty}\frac{\|\phi_k\|_\infty^2}{1+\lambda/\lambda_k}\left(\sum_{k=1}^{u}\frac{1}{1+\lambda/\lambda_k}\right) \le \frac{4}{n^2}\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda).$$

As a consequence, the total variance is bounded by

$$\left\|\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\right]\right\| \le \sum_{i=1}^{n}\left\|\mathbb{E}\left[X_i^2\right]\right\| \le \sum_{i=1}^{n}\mathbb{E}\left[\|X_i^2\|\right] \le \frac{4}{n}\mathcal{S}(\sigma,\lambda)\mathcal{N}(\sigma,\lambda) := \Sigma^2.$$

Using the first inequality of (58), we can get the conclusion (19). □

Proof of Proposition 4. The first bound (21) can be found in [31]. We now derive the second inequality (22). Observe that

$$\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\mathbf{v} = (M+\lambda I)^{-\frac{1}{2}}\left(\frac{1}{n}M^{1/2}\Phi^{\mathrm{T}}\mathbf{v}\right).$$

Since the matrix $(M+\lambda I)^{-\frac{1}{2}}$ is positive definite, its operator norm is bounded as

$$\left\|(M+\lambda I)^{-\frac{1}{2}}\right\| = \sup_{1\le j\le u}\frac{1}{\sqrt{\lambda_j+\lambda}} \le \frac{1}{\sqrt{\lambda}}. \tag{62}$$

For the estimate of $\left\|\frac{1}{n}M^{1/2}\Phi^{\mathrm{T}}\mathbf{v}\right\|_2^2$, let $\Phi_k = (\phi_k(x_1),\cdots,\phi_k(x_n))^{\mathrm{T}} \in \mathbb{R}^n$, then

$$\left\|\frac{1}{n}M^{1/2}\Phi^{\mathrm{T}}\mathbf{v}\right\|_2^2 = \frac{1}{n^2}\sum_{k=1}^{u}\lambda_k\left(\Phi_k^{\mathrm{T}}\mathbf{v}\right)^2 \le \frac{1}{n^2}\sum_{k=1}^{u}\lambda_k\|\Phi_k\|_2^2\|\mathbf{v}\|_2^2.$$

Notice that

$$\mathbb{E}\left[\|\Phi_k\|_2^2\|\mathbf{v}\|_2^2\right] = \mathbb{E}\left[\|\Phi_k\|_2^2\,\mathbb{E}^*\left[\|\mathbf{v}\|_2^2\right]\right].$$

We first bound the conditional expectation $\mathbb{E}^*\left[\|\mathbf{v}\|_2^2\right]$. We find

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

$$v_i^2 \le \left(\sum_{j=u+1}^{\infty} e_j\phi_j(x_i)\right)^2 \le \left(\sum_{j=u+1}^{\infty} e_j^2/\lambda_j\right)\left(\sum_{j=u+1}^{\infty}\lambda_j\phi_j^2(x_i)\right)$$
$$\le \|E\|_{\mathcal{H}_\sigma}^2\left(\sum_{j=u+1}^{\infty}\lambda_j\|\phi_j\|_\infty^2\right) \le 2\left(\|f_{D_l}\|_{\mathcal{H}_\sigma}^2 + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right)\left(\sum_{j=u+1}^{\infty}\lambda_j\|\phi_j\|_\infty^2\right).$$

This together with (17) yields

$$\mathbb{E}^*\left[\|\mathbf{v}\|_2^2\right] = \mathbb{E}^*\left[\sum_{i=1}^{n}|v_i|^2\right] \le 2n\left(B^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right)\left(\sum_{j=u+1}^{\infty}\lambda_j\|\phi_j\|_\infty^2\right).$$

Notice that $\mathbb{E}[\|\Phi_k\|_2^2] = \mathbb{E}\left[\sum_{i=1}^n \phi_k^2(x_i)\right] = n$. Based on the above estimates, we get

$$\mathbb{E}\left[\left\|\frac{1}{n}M^{1/2}\phi^{\mathrm{T}}\mathbf{v}\right\|_2^2\right] \le 2\left(\sum_{k=1}^{u}\lambda_k\right)\left(B^2/\lambda + \|f_\lambda^\sigma\|_{\mathcal{H}_\sigma}^2\right)\left(\sum_{j=d+1}^{\infty}\lambda_j\|\phi_j\|_\infty^2\right).$$

Then the conclusion (22) follows from (62) and $\sum_{k=1}^u \lambda_k \le Tr(L_{K_\sigma})$.

Finally, we turn to $\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon\right\|_2^2\right]$. Noting that the diagonal entries of Q^{-1} is $1/\sqrt{1+\lambda/\lambda_k}$, then

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon\right\|_2^2\right] = \frac{1}{n^2}\sum_{k=1}^{u}\sum_{i=1}^{n}\frac{\mathbb{E}[\phi_k^2(x_i)\epsilon_i^2]}{1+\lambda/\lambda_k}. \tag{63}$$

Decompose $\epsilon_i = f_\lambda^\sigma(x_i) - y_i$ as $\epsilon_i = f_\lambda^\sigma(x_i) - f_\rho(x_i) + f_\rho(x_i) - y_i$. This together with (4) yields that

$$\frac{\mathbb{E}[\phi_k^2(x_i)\epsilon_i^2]}{1+\lambda/\lambda_k} \le 2\left(\frac{\mathbb{E}\left[\phi_k^2(x_i)(f_\lambda^\sigma(x_i)-f_\rho(x_i))^2\right]}{1+\lambda/\lambda_k} + \frac{\mathbb{E}\left[\phi_k^2(x_i)(f_\rho(x_i)-y_i)^2\right]}{1+\lambda/\lambda_k}\right)$$
$$\le 2\left(\frac{\|\phi_k\|_\infty^2\mathbb{E}\left[(f_\lambda^\sigma(x_i)-f_\rho(x_i))^2\right]}{1+\lambda/\lambda_k} + \frac{2\mathbb{E}\left[\phi_k^2(x_i)\mathbb{E}^*\left[|f_\rho(x_i)|^2+|y_i|^2\right]\right]}{1+\lambda/\lambda_k}\right)$$
$$\le 2\left(\frac{\|\phi_k\|_\infty^2\mathbb{E}\left[(f_\lambda^\sigma(x_i)-f_\rho(x_i))^2\right]}{1+\lambda/\lambda_k} + \frac{4B^2}{1+\lambda/\lambda_k}\right)$$
$$\le 2\left(\frac{\|\phi_k\|_\infty^2\mathcal{D}(\sigma,\lambda)}{1+\lambda/\lambda_k} + \frac{4B^2}{1+\lambda/\lambda_k}\right),$$

where the last inequality is derived from the fact

$$\mathbb{E}\left[(f_\lambda^\sigma(x_i)-f_\rho(x_i))^2\right] = \int_{\mathcal{X}}(f_\lambda^\sigma(x)-f_\rho(x))^2\,d\rho_{\mathcal{X}} \le \mathcal{D}(\sigma,\lambda).$$

As a consequence,

$$\mathbb{E}\left[\left\|\frac{1}{n}Q^{-1}\Phi^{\mathrm{T}}\epsilon\right\|_2^2\right] \le \frac{2}{n^2}\sum_{k=1}^{u}\sum_{i=1}^{n}\left(\frac{\|\phi_k\|_\infty^2\mathcal{D}(\sigma,\lambda)}{1+\lambda/\lambda_k} + \frac{4B^2}{1+\lambda/\lambda_k}\right)$$
$$= \frac{2}{n}\left[\mathcal{S}(\sigma,\lambda)\mathcal{D}(\sigma,\lambda) + 4B^2\mathcal{N}(\sigma,\lambda)\right].$$

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

Table 1
List of notations.

Notation	Meaning of the notation
D	the labeled data set $D = \{(x_i, y_i)\}_{i=1}^N$
D^*	the unlabeled data set $D^* = \{x_i^*\}_{i=1}^{N^*}$
$\tilde{D}$	the union of D and D^* with $\tilde{D} = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{\tilde{N}}$ given in subsection 2.3
D_l /D_l^* /$\tilde{D}_l$	the l-th subset of D /D^*/$\tilde{D}$
n /n^* /$\tilde{n}$	the size of D_l /D_l^*/$\tilde{D}_l$
$f_{D_l}/f_{\tilde{D}_l}$	the l-th output function by regularized least squares with $D_l/\tilde{D}_l$
$\bar{f}_D/\ \bar{f}_{\tilde{D}}$	the global output of distributed least squares with $D/\tilde{D}$
$\epsilon/\tilde{\epsilon}$	the noise vector $\epsilon = \{\epsilon_i\}_{i=1}^n$ with $\epsilon_i = f_\lambda^\sigma(x_i) - y_i$/ $\tilde{\epsilon} = \{\tilde{\epsilon}_i\}_{i=1}^{\tilde{n}}$ with $\tilde{\epsilon}_i = f_\lambda^\sigma(\tilde{x}_i) - \tilde{y}_i$
u	the positive integer that is to be determined
ϕ_j	the j-th eigenfunction of the operator L_{K_σ}
Φ	$n \times u$ matrix with entries $\Phi_{ij} = \phi_j(x_i), i = 1, \cdots, n, j = 1, \cdots, u$
$\tilde{\Phi}$	$\tilde{n} \times u$ matrix with entries $\tilde{\Phi}_{ij} = \phi_j(\tilde{x}_i), i = 1, \cdots, \tilde{n}, j = 1, \cdots, u$
M	diagonal matrix $M = diag(\lambda_1, \cdots, \lambda_u)$
Q	diagonal matrix $Q = (I + \lambda M)^{\frac{1}{2}}$
$e_i/\tilde{e}_i$	basis coefficients of $f_{D_l} - f_\lambda^\sigma$/ $f_{\tilde{D}_l} - f_\lambda^\sigma$ in $f_{D_l} - f_\lambda^\sigma = \sum_j e_j\phi_j$ /$f_{\tilde{D}_l} - f_\lambda^\sigma = \sum_j \tilde{e}_j\phi_j$
$\mathbf{e^1}/\tilde{\mathbf{e}}^\mathbf{1}$	the u dimension vector $\mathbf{e^1} = [e_1, \cdots, e_u]^T/\tilde{\mathbf{e}}^\mathbf{1} = [\tilde{e}_1, \cdots, \tilde{e}_u]^T$
$\mathbf{e^2}/\tilde{\mathbf{e}}^\mathbf{2}$	the tail vector $\mathbf{e^2} = [e_{u+1}, \cdots]^T$ /$\tilde{\mathbf{e}}^\mathbf{2} = [\tilde{e}_{u+1}, \cdots]^T$
$\mathbf{a^1}$	vector $\{a_i\}_{i=1}^u$ consisting of the first u basis coefficients of $f_\lambda^\sigma = \sum_j a_j\phi_j$ in $L^2_{\rho_\mathcal{X}}$
$\mathbf{v}/\tilde{\mathbf{v}}$	vector $\mathbf{v} = [v_1, \cdots, v_n]^T, v_i = \sum_{j=u+1}^\infty e_j\phi_j(x_i)/\tilde{\mathbf{v}} = [\tilde{v}_1, \cdots, \tilde{v}_n]^T$, $\tilde{v}_i = \sum_{j=u+1}^\infty \tilde{e}_j\phi_j(x_i)$
β_u	the tail sum $\beta_u = \sum_{j=u+1}^\infty \lambda_j \|\phi_j\|_\infty^2$
$Tr(L_{K_\sigma})$	the kernel trace $\sum_j \lambda_j$.
$\mathcal{N}(\sigma, \lambda)$	the effective dimension of $\mathcal{H}_\sigma$, given as the sum $\sum_k \frac{1}{1+\lambda/\lambda_k}$
$\mathcal{S}(\sigma, \lambda)$	the sum $\sum_k \frac{\|\phi_k\|_\infty^2}{1+\lambda/\lambda_k}$

The proof is complete. □

References

[1] R.A. Adams, J.F. Fournier, Sobolev Spaces, Elsevier, 2003.
[2] A. Caponnetto, E. De Vito, Optimal rates for the regularized least-squares algorithm, Found. Comput. Math. 7 (3) (2007) 331–368.
[3] A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning theory, Anal. Appl. 8 (02) (2010) 161–183.
[4] X. Chang, S.-B. Lin, D.-X. Zhou, Distributed semi-supervised learning with kernel ridge regression, J. Mach. Learn. Res. 18 (1) (2017) 1493–1514.
[5] M. Eberts, I. Steinwart, Optimal regression rates for SVMs using Gaussian kernels, Electron. J. Stat. 7 (2013) 1–42.
[6] D.E. Edmunds, H. Triebel, Function Spaces, Entropy Numbers, Differential Operators, Cambridge University Press, 2008.
[7] Z.-C. Guo, L. Shi, Q. Wu, Learning theory of distributed regression with bias corrected regularization kernel network, J. Mach. Learn. Res. 18 (1) (2017) 4237–4261.
[8] C.-J. Hsieh, S. Si, I. Dhillon, A divide-and-conquer solver for kernel support vector machines, in: International Conference on Machine Learning, 2014, pp. 566–574.
[9] T. Hu, Q. Wu, D.-X. Zhou, Distributed kernel gradient descent algorithm for minimum error entropy principle, Appl. Comput. Harmon. Anal. 20 (1) (2020) 229–256.
[10] J. Lin, A. Rudi, L. Rosasco, V. Cevher, Optimal rates for spectral algorithms with least-squares regression over Hilbert spaces, Appl. Comput. Harmon. Anal. 48 (3) (2020) 868–890.
[11] S.-B. Lin, X. Guo, D.-X. Zhou, Distributed learning with regularized least squares, J. Mach. Learn. Res. 18 (1) (2017) 3202–3232.
[12] S.-B. Lin, D.-X. Zhou, Distributed kernel-based gradient descent algorithms, Constr. Approx. 47 (2) (2018) 249–276.
[13] L.W. Mackey, M.I. Jordan, A. Talwalkar, Divide-and-conquer matrix factorization, in: Advances in Neural Information Processing Systems, 2011, pp. 1134–1142.
[14] R. McDonald, K. Hall, G. Mann, Distributed training strategies for the structured perceptron, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, 2010, pp. 456–464.
[15] R. Mcdonald, M. Mohri, N. Silberman, D. Walker, G.S. Mann, Efficient large-scale distributed training of conditional maximum entropy models, in: Advances in Neural Information Processing Systems, 2009, pp. 1231–1239.
[16] S. Mendelson, J. Neeman, Regularization in kernel learning, Ann. Stat. 38 (1) (2010) 526–565.
[17] N. Mücke, G. Blanchard, Parallelizing spectrally regularized kernel algorithms, J. Mach. Learn. Res. 19 (1) (2018) 1069–1097.

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008

[18] S. Smale, D.-X. Zhou, Estimating the approximation error in learning theory, Anal. Appl. 1 (01) (2003) 17–41.
[19] I. Steinwart, Oracle inequalities for support vector machines that are based on random entropy numbers, J. Complex. 25 (5) (2009) 437–454.
[20] I. Steinwart, A. Christmann, Support Vector Machines, Springer, 2008.
[21] I. Steinwart, D.R. Hush, C. Scovel, Optimal rates for regularized least squares regression, in: COLT, 2009.
[22] I. Steinwart, C. Scovel, Fast rates for support vector machines using Gaussian kernels, Ann. Stat. 35 (2) (2007) 575–607.
[23] I. Steinwart, P. Thomann, N. Schmid, Learning with hierarchical Gaussian kernels, Technical report, Fakultät für Mathematik und Physik, Universität Stuttgart, 2016, arXiv:1612.00824, 2016.
[24] I. Sutskever, J. Martens, G. Dahl, G. Hinton, On the importance of initialization and momentum in deep learning, in: International Conference on Machine Learning, 2013, pp. 1139–1147.
[25] V. Temlyakov, Optimal estimators in learning theory, Banach Cent. Publ. 72 (2006) 341.
[26] J.A. Tropp, User-friendly tail bounds for sums of random matrices, Found. Comput. Math. 12 (4) (2012) 389–434.
[27] B. Wang, T. Hu, Unregularized online algorithms with varying Gaussians. Accepted by Constructive Approximation.
[28] D.-H. Xiang, D.-X. Zhou, Classification with Gaussians and convex loss, J. Mach. Learn. Res. 10 (2009) 1447–1468.
[29] Y. Ying, D.-X. Zhou, Learnability of Gaussians with flexible variances, J. Mach. Learn. Res. 8 (2007) 249–276.
[30] S. Zhang, A.E. Choromanska, Y. LeCun, Deep learning with elastic averaging SGD, in: Advances in Neural Information Processing Systems, 2015, pp. 685–693.
[31] Y. Zhang, J. Duchi, M. Wainwright, Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates, J. Mach. Learn. Res. 16 (1) (2015) 3299–3340.
[32] D.-X. Zhou, The covering number in learning theory, J. Complex. 18 (3) (2002) 739–767.
[33] D.X. Zhou, Deep distributed convolutional neural networks: universality, Anal. Appl. 16 (06) (2018) 895–919.
[34] D.X. Zhou, Universality of deep convolutional neural networks, Appl. Comput. Harmon. Anal. 48 (2) (2020) 787–794.

Please cite this article in press as: T. Hu, D.-X. Zhou, Distributed regularized least squares with flexible Gaussian kernels, Appl. Comput. Harmon. Anal. (2021), https://doi.org/10.1016/j.acha.2021.03.008