Electronic version of an article published as Analysis and Applications, vol. 17, no. 5, 2019, p. 837-851, https://doi.org/10.1142/S0219530519400037,

On semi-supervised learning with summary statistics

Huihui Qin[1] and Xin Guo[2]

[1]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. huihui.qin@connect.polyu.hk

[2]Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. x.guo@polyu.edu.hk

June 24, 2019

Abstract

Nowadays the extensive collection and analyzing of data is stimulating widespread privacy concerns, and therefore is increasing tensions between the potential sources of data and researchers. A privacy-friendly learning framework can help to ease the tensions, and to free up more data for research. We propose a new algorithm, LESS (Learning with Empirical feature-based Summary statistics from Semi-supervised data), which uses only summary statistics instead of raw data for regression learning. The selection of empirical features serves as a trade-off between prediction precision and the protection of privacy. We show that LESS achieves the minimax optimal rate of convergence, in terms of the size of the labeled sample. LESS extends naturally to the applications where data are separately held by different sources. Compared with existing literature on distributed learning, LESS removes the restriction of minimum sample size on single data sources.

Keywords: distributed learning, semi-supervised learning, empirical features, summary statistics, privacy protection

Mathematics Subject Classification 2010: 68T05, 68Q32, 41A25

1 Introduction

Many reproducing kernel-based machine learning algorithms are designed without considering privacy issues. In particular, under the structural risk minimization scheme, as pointed out by the representer theorem, the whole input part of training data, which may contain private information, has to be shipped along with the predicted function. Privacy concern would restrict the application of such algorithms. On the other hand, usually there are unlabeled data available with the same marginal distribution as the training data. For example, these unlabeled data could be produced by sampling from the estimated density, or be obtained from public domain without privacy issues [38, 28]. In this paper, we study the

methodology for masking the sensitive private information in training data, with the help of unlabeled data.

Semi-supervised learning is a big class of machine learning problems where unlabeled data are used in addition to the data points with labels, e.g., for classification or regression. In recent years, unlabeled data are observed helpful for capturing the underlying manifold structures of data distribution [12, 3], relaxing the requirement on single-source minimum sample size in distributed learning [26, 16], and improving the convergence under weak regularity assumptions of the regression function [16]. In this paper, unlabeled data (possibly also including the input part of the labeled data) are used to build empirical features first. Then, we use the empirical features to construct summary statistics, based on which we introduce a new algorithm, **LESS** (Learning with Empirical feature-based Summary statistics from Semi-supervised data), of which the main advantages are summarized below.

- LESS achieves the minimax optimal convergence rate, in terms of the size of labeled sample.
- With the help of unlabeled data, LESS has an automatic generalization to distributed learning, where the restriction on single-source minimum sample size is completely removed.
- The summary statistics we adopt provide a protocol for communicating data with privacy. Unlike classical kernel-based algorithms, LESS collects only the summary statistics, instead of the private raw data, for the centralized learning process.

Consider a regression learning problem with an input space X, which is a compact metric space, and an output space $Y \subset \mathbb{R}$. Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be a sample drawn independently from $(Z = X \times Y, \rho)$, where ρ is an unknown Borel probability measure such that the marginal distribution ρ_X on X is nondegenerate, i.e., $\rho_X(A) > 0$ for any measurable set A that has an interior point. The target of the regression problem is to learn the regression function $f_\rho : X \to \mathbb{R}$,

$$f_\rho(x) = \int_Y y d\rho(y|x),$$

from the sample $\mathbf{z}$, where $\rho(y|x)$ is the conditional distribution of ρ at x.

There is a large literature of the kernel methods for machine learning. See [32, 31, 13, 1, 35, 25], and the reference therein. Let $K : X \times X \to \mathbb{R}$ be a Mercer kernel. That is, K is a function which is symmetric, continuous, and positive, where positive means $\sum_{i,j=1}^l c_i c_j K(u_i, u_j) \geq 0$ for any integer $1 \leq l < \infty$, any coefficients $c_1, \ldots, c_l \in \mathbb{R}$, and any elements $u_1, \ldots, u_l \in X$. Let $(\mathcal{H}_K, \langle \cdot, \cdot \rangle_K, \|\cdot\|_K)$ be the reproducing kernel Hilbert space generated by K. The classical kernel-based regularized least squares algorithm is defined by

$$f_\lambda^{\mathbf{z}} = \arg \min_{f \in \mathcal{H}_K} \left\{ \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\}, \tag{1}$$

where $\lambda > 0$ is the regularization parameter. Kernel-based learning algorithms usually have the flaws in privacy protection. For example, by the well-known representer theorem [35], $f_\lambda^{\mathbf{z}}$ in (1) takes the form

$$f_\lambda^{\mathbf{z}} = \sum_{i=1}^{m} c_i K_{x_i}, \tag{2}$$

where $c_1, \ldots, c_m \in \mathbb{R}$ are the coefficients determined by (1), and for any $x, u \in X$, the function $K_x : X \to \mathbb{R}$ is defined by $K_x(u) = K(x, u)$. It is easy to see that to ship $f_\lambda^{\mathbf{z}}$, the unlabeled part $\mathbf{x} = \{x_i\}_{i=1}^m$ of the sample $\mathbf{z}$ must be shipped together. We put a discussion in Section 3. In this paper, we try to solve this problem on privacy, by introducing the empirical feature-based summary statistics.

We assume that there is another sample $\mathbf{u} = \{u_i\}_{i=1}^n$, drawn independently from ρ_X without labels. For applications, the sample $\mathbf{u}$ may come from some openly accessible sources, for example those with the privacy expired. Note that we do not assume independence between $\mathbf{u}$ and $\mathbf{x}$. In particular, a part, or even all of $\mathbf{x}$ could just be put into $\mathbf{u}$. This inclusion is sometimes useful, and is covered by our analysis.

Define $L_K^{\mathbf{u}} : \mathcal{H}_K \to \mathcal{H}_K$ as an operator by

$$L_K^{\mathbf{u}} f = \frac{1}{n} \sum_{i=1}^{n} f(u_i) K_{u_i}, \tag{3}$$

where $|\mathbf{u}| = n$ is the size of $\mathbf{u}$. By the reproducing property [1, 14] that for any $f \in \mathcal{H}_K$ and $u \in X$, $\langle f, K_u \rangle_K = f(u)$, one has that for any $f, g \in \mathcal{H}_K$,

$$\langle L_K^{\mathbf{u}} f, g \rangle_K = \frac{1}{n} \sum_{i=1}^{n} f(u_i) g(u_i) = \langle f, L_K^{\mathbf{u}} g \rangle_K .$$

In particular, $\langle L_K^{\mathbf{u}} f, f \rangle_K = \frac{1}{n} \sum_{i=1}^n f(u_i)^2 \geq 0$. So $L_K^{\mathbf{u}}$ is a positive semi-definite operator with rank (i.e., the dimension of its image) at most n. Therefore, we can write $\{(\lambda_i^{\mathbf{u}}, \phi_i^{\mathbf{u}})\}_i$ as the eigensystem of $L_K^{\mathbf{u}}$ with $\lambda_1^{\mathbf{u}} \geq \lambda_2^{\mathbf{u}} \geq \ldots \geq \lambda_n^{\mathbf{u}} \geq 0 = \lambda_{n+1}^{\mathbf{u}} = \ldots$. The zero eigenvalues are counted purposely to make $\{\phi_i^{\mathbf{u}}\}_i$ an orthonormal basis of $\mathcal{H}_K$. Similarly, we define $L_K^{\mathbf{x}}$ and $\{(\lambda_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_i$ for the input part $\mathbf{x}$ of the sample $\mathbf{z}$ by substituting $\mathbf{u}$ with $\mathbf{x}$, and n with $m = |\mathbf{x}|$ in (3).

Algorithm LESS. The sample dependent functions $\phi_i^{\mathbf{u}}$'s are referred to as empirical features (so are $\phi_i^{\mathbf{x}}$'s). These functions are studied in literature [17, 39, 40] as powerful tools for regression, classification, and nonlinear dimension reduction. Let $1 \leq N \leq n$ be an integer. Consider the summary statistic $\mathbf{d} = (d_1, \ldots, d_N)^T$, defined by

$$d_i = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{m} \sum_{j=1}^{m} y_j K_{x_j} \right\rangle_K, \quad 1 \leq i \leq N. \tag{4}$$

The superscripts $\mathbf{u}$ and $\mathbf{z}$ of $\mathbf{d}$ and d_i's are dropped to avoid heavy notation. The summary statistic $\mathbf{d}$ is then used to build the output function of LESS,

$$f_\lambda^{\mathbf{u},\mathbf{z}} = (L_K^{\mathbf{u}} + \lambda I)^{-1} \sum_{i=1}^{N} d_i \phi_i^{\mathbf{u}} = \sum_{i=1}^{N} \frac{d_i}{\lambda_i^{\mathbf{u}} + \lambda} \phi_i^{\mathbf{u}}, \tag{5}$$

where $\lambda > 0$ is the regularization parameter, and in this paper, I denotes the identity operator, with its domain inferred from the context. Here, recall that $\phi_i^{\mathbf{u}}$ is an eigenfunction of $L_K^{\mathbf{u}}$, $L_K^{\mathbf{u}}\phi_i^{\mathbf{u}} = \lambda_i^{\mathbf{u}}\phi_i^{\mathbf{u}}$. We have $(L_K^{\mathbf{u}} + \lambda I)^{-1}\phi_i^{\mathbf{u}} = \frac{1}{\lambda_i^{\mathbf{u}}+\lambda}\phi_i^{\mathbf{u}}$.

We see that by the introduction of the empirical features $\phi_i^{\mathbf{u}}$'s, the training sample $\mathbf{z}$ is encoded into $\mathbf{d}$, instead of directly shipped along the predicted function $f_\lambda^{\mathbf{u},\mathbf{z}}$. From the statistic $\mathbf{d}$, it is even not trivial to recover the sample size m! Of course, a safer design could be achieved by adding noise to $\mathbf{d}$, which we leave as future work.

LESS for distributed learning. The summary statistics $\mathbf{d}$ provides an automatic and unified way for distributed learning. In fact, suppose that instead of (4), the sample $\mathbf{z}$ is stored separately in ℓ sources $\mathbf{z} = \mathbf{z}_1 \cup \mathbf{z}_2 \cup \ldots \cup \mathbf{z}_\ell$ without overlapping, then one defines $\mathbf{d}^J = (d_1^J, \ldots, d_N^J)^T$ by

$$d_i^J = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{|\mathbf{z}_J|} \sum_{(x,y)\in\mathbf{z}_J} yK_x \right\rangle_K, \quad 1 \le J \le \ell, \quad 1 \le i \le N. \tag{6}$$

Again, one may centralize the summary statistics $\mathbf{d}^J$'s without directly collecting the private data sets $\mathbf{z}_J$'s. More importantly, the weighted average of $\mathbf{d}^J$'s is exactly $\mathbf{d}$,

$$\mathbf{d} = \sum_{J=1}^{\ell} \frac{|\mathbf{z}_J|}{|\mathbf{z}|}\mathbf{d}^J. \tag{7}$$

So, without any configuration, LESS can be directly applied to distributed learning problems, where data are separately held by different sources as privacy. From (7), we see that the sizes of different data subsets have no effect on the learning process (5). In another way of saying, our analysis on LESS applies automatically to this distributed design (6).

The rest of the paper is organized as follows. We first give our main results in Section 2. Comparisons and discussions, as well as the details of implementations are put in Section 3. Proofs are placed in Section 4.

2 Main Results

In this section, we formulate the main assumptions and our main results.

Write $(L^2_{\rho_X}, \|\cdot\|_\rho)$ the Hilbert space of square-integrable functions on X with respect to the measure ρ_X. Define $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$ by

$$f \mapsto \int_X f(x)K_x d\rho_X(x).$$

Since K is continuous and X is compact, L_K is compact. It is easy to verify that L_K is positive semi-definite. Furthermore, L_K is of trace class (hence Hilbert-Schmidt), and since ρ_X is nondegenerate, $\|L_K^{1/2} f\|_K = \|f\|_\rho$ for any $f \in L^2_{\rho_X}$. Denote $\kappa = \max\{1, \sup_{x\in X}\sqrt{K(x,x)}\}$. We have $\mathsf{Trace}(L_K) \le \kappa^2$. See [13, 14] for detailed proofs. So we write

$$\lambda_1 \ge \lambda_2 \ge \ldots \ge 0,$$

as all the eigenvalues of L_K, and $\phi_1, \phi_2, \ldots$ the corresponding eigenfunctions, normalized in $\mathcal{H}_K$. For $\lambda > 0$, write

$$\mathcal{N}(\lambda) = \mathsf{Trace}(L_K(L_K + \lambda I)^{-1})$$

the effective dimension of L_K [37, 8, 7]. The following assumption (A1) characterizes the capacity of the hypothesis space $\mathcal{H}_K$, and is widely adopted in learning theory literature [26, 25, 6].

(A1) *There exist some constants $0 < C_1 < \infty$ and $0 < s \leq 1$ such that $\mathcal{N}(\lambda) \leq C_1 \lambda^{-s}$ for any $0 < \lambda < \infty$.*

The following assumption (A2) characterizes the regularity of the regression function.

(A2) *There exists some $g_\rho \in L^2_{\rho_X}$ and $1/2 \leq r \leq 1$ such that $f_\rho = L_K^r g_\rho$.*

Note that Assumption (A2) implies $f_\rho \in \mathcal{H}_K$.

(A3) *$\int_Z y^2 d\rho(x,y) < \infty$, and that there exist two constants $0 < \sigma, M < \infty$, such that*

$$\int_Y \left(\exp\left\{ \frac{|y - f_\rho(x)|}{M} \right\} - \frac{|y - f_\rho(x)|}{M} - 1 \right) d\rho(y|x) \leq \frac{\sigma^2}{2M^2},$$

for ρ_X-almost all $x \in X$.

In particular, (A3) holds with $\sigma = \sqrt{2(e^2 - 3)}M$, when $|y| \leq M$ almost surely. For more discussions on (A3), see [26, 2, 8].

From the design (4) and (5), we see that intuitively, one needs sufficient coordinates for $\mathbf{d}$ to guarantee the convergence. In particular, we characterize the requirement by the following assumption (A4).

(A4) *N is large enough (meaning that enough empirical features are used), so that $\lambda_{N+1} \leq \kappa^2 \lambda$.*

Theorem 2.1. *Assume (A1), (A2), (A3), and $n \geq m$. For any $0 < \delta < 1$, one has with confidence at least $1 - \delta$ that*

$$\begin{aligned} \left\| f_\lambda^{\mathbf{u},\mathbf{z}} - f_\rho \right\|_\rho \leq & \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda} + 2 \right) \left(\frac{M + \sigma}{\kappa} + \|f_\rho\|_K + \|g_\rho\|_\rho \right) \mathcal{B}_{m,\lambda} \log^3 \frac{10}{\delta} \\ & + \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda} + 2 \right)^r \left(\lambda + \frac{4\kappa^2}{\sqrt{n}} + \lambda_{N+1} \right)^r \|g_\rho\|_\rho \log^{3r} \frac{10}{\delta} + \|g_\rho\|_\rho \lambda^r, \end{aligned}$$

where

$$\mathcal{B}_{n,\lambda} = \frac{2\kappa^2}{n\sqrt{\lambda}} + 2\kappa \sqrt{\frac{\mathcal{N}(\lambda)}{n}}, \tag{8}$$

and $\mathcal{B}_{m,\lambda}$ is similarly defined by substituting n with m.

We cite from [16, Lemma B.1] the following lemma, which is standard, and the proof can also be found in [25] and [19, Lemma 11].

Lemma 2.2. *Let R be a nonnegative random variable. Let $\alpha, \beta, \gamma > 0$. If for any $0 < \delta < 1$, one has with confidence at least $1 - \delta$ that $R \leq \alpha \log^{\gamma} \frac{\beta}{\delta}$, then for any $\mu > 0$,*

$$(\mathbb{E}[R^{\mu}])^{1/\mu} \leq \alpha \left[\beta \Gamma(\mu\gamma + 1)\right]^{1/\mu},$$

where $\Gamma(t) = \int_0^{\infty} e^{-u} u^{t-1} du$ is the Gamma function.

Corollary 2.3. *Assume (A1), (A2), (A3), (A4), and $n \geq \max\{m, m^{\frac{2}{2r+s}}\}$. Let $\lambda = m^{-\frac{1}{2r+s}}$. For any $0 < \delta < 1$, one has with confidence at least $1 - \delta$ that*

$$\left\| f_{\lambda}^{\mathbf{u},\mathbf{z}} - f_{\rho} \right\|_{\rho} \leq C_2 m^{-\frac{r}{2r+s}} \log^3 \frac{10}{\delta}, \tag{9}$$

where C_2 is a constant independent of m, n, or δ, and it is given at the end of the proof. Moreover, for any $\mu > 0$, Lemma 2.2 gives

$$\left[\mathbb{E}(\left\| f_{\lambda}^{\mathbf{u},\mathbf{z}} - f_{\rho} \right\|_{\rho}^{\mu})\right]^{1/\mu} \leq C_2 \left[10\Gamma(3\mu + 1)\right]^{1/\mu} m^{-\frac{r}{2r+s}}. \tag{10}$$

Remark 2.4. *Recall that $1 \leq N \leq n$. With the assumption $n \geq \max\{m, m^{\frac{2}{2r+s}}\}$ and the setting $\lambda = m^{-\frac{1}{2r+s}}$, it is always possible to find some $N \leq n$ that satisfies Assumption (A4). In fact, since the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots$ of L_K are arranged in non-increasing order, $\lambda_n \leq \frac{1}{n}\mathsf{Trace}(L_K) \leq \frac{\kappa^2}{n} \leq \kappa^2 m^{-\frac{1}{2r+s}} = \kappa^2 \lambda$.*

Remark 2.5. *It is well understood [8, 33, 2] that when $1/2 \leq r \leq 1$, the minimax optimal learning rate for learning algorithms that have only the access to $\mathbf{z}$ and with output functions in $\mathcal{H}_K$, is $O(m^{-\frac{r}{2r+s}})$. The bounds (9) and (10) in Corollary 2.3 match this rate.*

3 Discussions and Comparisons

3.1 Details for the implementations

Recall $m = |\mathbf{x}|$. Define the sampling operator $S_{\mathbf{x}} : \mathcal{H}_K \to \mathbb{R}^m$,

$$f \mapsto (f(x_i))_{i=1}^m.$$

It is straightforward to see that the adjoint operator $S_{\mathbf{x}}^T : \mathbb{R}^m \to \mathcal{H}_K$ is defined by

$$(c_i)_{i=1}^m \mapsto \sum_{i=1}^m c_i K_{x_i}.$$

Let $\mathbb{K}$ be the Gram matrix of the Mercer kernel K on $\mathbf{x}$, $\mathbb{K} = (K(x_i, x_j))_{i,j=1}^m$. Then,

$$\frac{1}{m}\mathbb{K} = \frac{1}{m} S_{\mathbf{x}} S_{\mathbf{x}}^T, \qquad L_K^{\mathbf{x}} = \frac{1}{m} S_{\mathbf{x}}^T S_{\mathbf{x}}. \tag{11}$$

So the eigenvalues of $\frac{1}{m}\mathbb{K}$, counting multiplicity, are $\lambda_1^{\mathbf{x}}, \ldots, \lambda_m^{\mathbf{x}}$, which are the first m eigenvalues of $L_K^{\mathbf{x}}$. Since $\mathbb{K}$ is positive semi-definite, we have the following eigen-decomposition

$$\frac{1}{m}\mathbb{K} = U\Lambda U^T, \qquad \Lambda = \mathrm{diag}\{\lambda_1^{\mathbf{x}}, \ldots, \lambda_m^{\mathbf{x}}\},$$

where $U = [U_1, \ldots, U_m]$ is an orthogonal matrix. Some simple linear algebra shows that if $\lambda_i^{\mathbf{x}} = 0$, then $\langle \phi_i^{\mathbf{x}}, L_K^{\mathbf{x}} \phi_i^{\mathbf{x}} \rangle_K = 0$, so $S_{\mathbf{x}} \phi_i^{\mathbf{x}} = 0$, which means $\phi_i^{\mathbf{x}}$ is perpendicular to the linear space spanned by $\{K_x : x \in \mathbf{x}\}$. In this case we do not have a representation of $\phi_i^{\mathbf{x}}$ with $\{K_x : x \in \mathbf{x}\}$. When $\lambda_i^{\mathbf{x}} > 0$, from $S_{\mathbf{x}}^T U_i = \frac{1}{\lambda_i^{\mathbf{x}}} S_{\mathbf{x}}^T (\frac{1}{m}\mathbb{K} U_i) = \frac{1}{\lambda_i^{\mathbf{x}}} L_K^{\mathbf{x}} (S_{\mathbf{x}}^T U_i)$, and $\left\| S_{\mathbf{x}}^T U_i \right\|_K^2 = m \left\langle U_i, \frac{1}{m}\mathbb{K} U_i \right\rangle_{\mathbb{R}^m} = m\lambda_i^{\mathbf{x}}$, we can take

$$\phi_i^{\mathbf{x}} = \frac{1}{\sqrt{m\lambda_i^{\mathbf{x}}}} S_{\mathbf{x}}^T U_i, \qquad U_i = \frac{1}{\sqrt{m\lambda_i^{\mathbf{x}}}} S_{\mathbf{x}} \phi_i^{\mathbf{x}}.$$

For two samples $\mathbf{x}$ and $\mathbf{u}$ with sizes m and n respectively, denote $\mathbb{K}_{\mathbf{u},\mathbf{x}}$ the $n \times m$ matrix of which the (i,j) entry is $K(u_i, x_j)$. Then $\mathbb{K}_{\mathbf{u},\mathbf{x}} = \mathbb{K}_{\mathbf{x},\mathbf{u}}^T$, and $S_{\mathbf{u}} S_{\mathbf{x}}^T = \mathbb{K}_{\mathbf{u},\mathbf{x}}$. The Gram matrix $\mathbb{K}_{\mathbf{u},\mathbf{u}}$ of size $n \times n$ is similarly defined with the sample $\mathbf{u}$. The summary statistic $\mathbf{d}$ could be computed through

$$d_i = \left\langle \phi_i^{\mathbf{u}}, \frac{1}{m}\sum_{j=1}^m y_j K_{x_j} \right\rangle_K = \left\langle \frac{1}{\sqrt{n\lambda_i^{\mathbf{u}}}} S_{\mathbf{u}}^T V_i, \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} \right\rangle_K = \frac{1}{m\sqrt{n\lambda_i^{\mathbf{u}}}} \left\langle V_i, \mathbb{K}_{\mathbf{u},\mathbf{x}} \mathbf{y} \right\rangle_{\mathbb{R}^n},$$

where $V = [V_1, \ldots, V_n]$ is the orthogonal matrix defined by the eigen-decomposition $\frac{1}{n}\mathbb{K}_{\mathbf{u},\mathbf{u}} = V \mathrm{diag}\{\lambda_1^{\mathbf{u}}, \ldots, \lambda_n^{\mathbf{u}}\} V^T$.

3.2 Motivating applications

This paper is inspired by two recent works [38, 28] in statistics. Consider the linear regression model $y = \mathbb{X}\beta + \varepsilon$, and its least squares solution $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y$. Roughly speaking, the basic idea in [38, 28] is only to collect the summary statistic $\mathbb{X}^T y$ as a whole, and use a new estimator $\hat{\beta}' = (\tilde{\mathbb{X}}^T\tilde{\mathbb{X}})^{-1}\mathbb{X}^T y$ to replace $\hat{\beta}$. Here $\tilde{\mathbb{X}}$ is the coefficient matrix made by openly accessible and unlabeled data without privacy issues. Real applications with data of both $\mathbb{X}$ and $\tilde{\mathbb{X}}$ are studied in the works. The relation between the predicted function $f_\lambda^{\mathbf{z}}$ of regularized least squares, and the predicted function $f_\lambda^{\mathbf{u},\mathbf{z}}$ of LESS is similar to that between $\hat{\beta}$ and $\hat{\beta}'$. In fact, for any $f, g, h \in \mathcal{H}_K$, define $f \otimes g$ as an operator by $(f \otimes g)h = \langle g, h \rangle_K f$. Define $P_N : \mathcal{H}_K \to \mathcal{H}_K$ as the orthogonal projection onto the subspace spanned by $\{\phi_i^{\mathbf{u}}\}_{i=1}^N$. That is, $P_N = \sum_{i=1}^N \phi_i^{\mathbf{u}} \otimes \phi_i^{\mathbf{u}}$. It is well known [31] that $f_\lambda^{\mathbf{z}} = (L_K^{\mathbf{x}} + \lambda I)^{-1} \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y}$, and we can write $f_\lambda^{\mathbf{u},\mathbf{z}}$ by replacing $L_K^{\mathbf{x}}$ by $L_K^{\mathbf{u}}$, and inserting the projection P_N as a protocol,

$$f_\lambda^{\mathbf{u},\mathbf{z}} = (L_K^{\mathbf{u}} + \lambda I)^{-1} P_N \frac{1}{m} S_{\mathbf{x}}^T \mathbf{y}.$$

LESS can be used as a privacy-friendly substitute for regularized least squares (1). The solution $f_\lambda^{\mathbf{z}}$ in (2) of Problem (1) is a linear combination of kernel functions on the sample. To compute $f_\lambda^{\mathbf{z}}$, the sample $\mathbf{z}$ must be collected from the holder of data. To ship $f_\lambda^{\mathbf{z}}$ to the

users, at least the input part $\mathbf{x}$ should explicitly be shipped, and the labels y_i's could thus be estimated via $y_i \approx f_\lambda^{\mathbf{z}}(x_i)$. Although when the input space X is an Euclidean domain with low dimension, one may ship $f_\lambda^{\mathbf{z}}$ in terms of its local approximations with splines or wavelets, such approximation could be difficult when the dimension of X is high. LESS solves this problem by collecting only the summary statistic $\mathbf{d}$ and shipping the predicted function $f_\lambda^{\mathbf{u},\mathbf{z}}$ in terms of the linear combination of $\phi_i^{\mathbf{u}}$'s, which is eventually the linear combination of K_{u_i}'s, with $u_i \in \mathbf{u}$ free of privacy issues.

The dimension N of the summary statistic $\mathbf{d}$ balances the protection of privacy, and the least squares error of the predicted function $f_\lambda^{\mathbf{u},\mathbf{z}}$. As suggested by Assumption (A4) and Corollary 2.3, if N is large enough such that $\lambda_{N+1} \leq \kappa^2 \lambda$, $\mathbf{d}$ contains sufficient information that supports the optimal learning rate. In many applications the eigenvalues of L_K decay quickly and we do not need a large N to achieve (A4). For example, if X is an Euclidean domain and K is Sobolev smooth, then λ_i's decay polynomially [30]. If K is analytic, such as the widely used Gaussian kernel, then λ_i's decay exponentially [27]. From the proof of Theorem 2.1 and Corollary 2.3, we see that empirically, Assumption (A4) can be replaced by $\lambda_{N+1}^{\mathbf{u}} \leq \kappa^2 \lambda$ without affecting the error estimate. A better privacy protection can be achieved by adding noise to $\mathbf{d}$ (or to $\mathbf{d}^j$'s under the distributed setting). We leave the quantitative analysis of this approach as future work.

For the case the sample $\mathbf{z}$ is held separately by ℓ different sources $\mathbf{z} = \cup_{i=1}^{\ell} \mathbf{z}_i$, there are recent works [10, 26, 16] that study the method of inflating each sub-sample $\mathbf{z}_i$ with a separate unlabeled sample. The inflation is done as follows. Suppose $\mathbf{u}$ is an unlabeled sample divided into ℓ subsets $\mathbf{u} = \cup_{i=1}^{\ell} \mathbf{u}_i$. For each i, all the sample points in $\mathbf{u}_i$ are equipped with a fake label 0, and all the labels in $\mathbf{z}_i$ are scaled by the factor $(|\mathbf{z}_i|+|\mathbf{u}_i|)/|\mathbf{z}_i|$ to compensate for these fake labels. Then $\mathbf{z}_i$ and $\mathbf{u}_i$ are mixed as a sample to yield an output function $f_\lambda^{\mathbf{u}_i \cup \mathbf{z}_i}$ from regularized least squares. The overall output function $\bar{f}_\lambda^{\mathbf{z}}$ is the weighted average of $f_\lambda^{\mathbf{u}_i \cup \mathbf{z}_i}$'s. By this operation, [26] proved (with the assumptions $|\mathbf{z}_1| = \ldots = |\mathbf{z}_\ell|$ and $|\mathbf{u}_1| = \ldots = |\mathbf{u}_\ell|$) that when

$$\ell \leq \frac{1}{\log^5 m + 1} \min \left\{ (n+m)^{1/2} m^{-\frac{s+1}{4r+2s}}, (n+m)^{1/3} m^{\frac{2r+s-2}{6r+3s}} \right\}, \tag{12}$$

the output function $\bar{f}_\lambda^{\mathbf{z}}$ still achieves the minimax optimal learning rate.

Compared with the inflation method studied in [10, 26, 16], LESS provides a way better solution to the learning problems with multiple sources of data. First, although for the scenarios where it is not allowed to bring together the training data from different sources, the distributed-learning setting solves the training problem, one still has to ship out the new instances (to different sources of training data) for prediction. Usually, these instances also contain private information, and it is not appropriate to circulate them around. Second, in the worst case scenario, when the sample size of each subset $\mathbf{z}_i$ is $O(1)$, and without loss of generality we use $\ell = m$, then (12) implies (recall $0 < s \leq 1$)

$$n \gtrsim m^{2+\frac{2}{2r+s}}, \tag{13}$$

where $n \gtrsim f(m)$ means there exists some positive constant $0 < C < \infty$ such that $n = n(m) \geq Cf(m)$ for any positive integer m. Note that in Corollary 2.3, the functional relation $n(m)$ is implicitly given by the lower bound $n \geq \max\{m, m^{\frac{2}{2r+s}}\}$. The restriction (13) requires much more unlabeled sample points than LESS does

$$n \geq \max\{m, m^{\frac{2}{2r+s}}\}, \tag{14}$$

in Corollary 2.3. Third, when (13) is satisfied, in each single computing node (located at the corresponding data source), according to the analysis in [26], the regularized least squares algorithm would process an inflated sample of size

$$\frac{n}{m} \gtrsim m^{1+\frac{2}{2r+s}}. \tag{15}$$

While for LESS, since the computation is centralized, we do not need significant computation provided by the data sources, and the sample size to be processed by the central computing node for LESS could be reduced, as suggested by (14), to

$$O\left(\max\left\{m, m^{\frac{2}{2r+s}}\right\}\right),$$

which is even much smaller than (15).

Chaudhuri et al. [11] studied an algorithm that uses random features (instead of the empirical features we use) for learning. Noise is added to the coefficients of the random features to achieve differential privacy. Because of the adoption of random features, this algorithm in [11] works only with translation invariant kernels.

4 Proof of the Main Theorem

We cite the following lemma from [6, Lemma E.4] and [4, Theorem IX.2.1].

Lemma 4.1. *Let A and B be positive definite operators on a separable Hilbert space $\mathcal{H}$. Write $\|\cdot\|_{\mathsf{op}(\mathcal{H})}$ the operator norm of $\mathcal{H}$. Then for any $0 \leq s \leq 1$, we have*

$$\|A^s B^s\|_{\mathsf{op}(\mathcal{H})} \leq \|AB\|^s_{\mathsf{op}(\mathcal{H})}. \tag{16}$$

Write $f_\lambda = (L_K + \lambda I)^{-1} L_K f_\rho$. One has $\lambda f_\lambda = L_K(f_\rho - f_\lambda)$. Write $\|\cdot\|_{\mathsf{op}}$ the operator norm of all the bounded operators on $\mathcal{H}_K$.

Lemma 4.2. *We have the following error bound*

$$\left\|f_\lambda^{\mathbf{u},\mathbf{z}} - P_N f_\lambda\right\|_\rho \leq \Omega_{\mathbf{u},\lambda}\left(R_\lambda^{\mathbf{z}} + \|f_\rho\|_K W_\lambda^{\mathbf{x}} + \|g_\rho\|_\rho W_\lambda^{\mathbf{u}}\right), \tag{17}$$

where

$$\Omega_{\mathbf{u},\lambda} := \left\|(L_K^{\mathbf{u}} + \lambda I)^{-1}(L_K + \lambda I)\right\|_{\mathsf{op}}, \tag{18}$$

$$R_\lambda^{\mathbf{z}} := \left\|(L_K + \lambda I)^{-1/2}\left(\frac{1}{m} S_{\mathbf{x}}^T \mathbf{y} - L_K^{\mathbf{x}} f_\rho\right)\right\|_K, \tag{19}$$

$$W_\lambda^{\mathbf{u}} := \left\|(L_K + \lambda I)^{-1/2}(L_K - L_K^{\mathbf{u}})\right\|_{\mathsf{op}}, \tag{20}$$

and $W_\lambda^{\mathbf{x}}$ is defined in the same way as (20) by substituting $\mathbf{u}$ with $\mathbf{x}$.

Proof. Since $\operatorname{span}\{\phi_i^{\mathbf{u}}\}_{i=1}^N$ is an invariant subspace of $L_K^{\mathbf{u}}$, P_N and $L_K^{\mathbf{u}}$ commute. We have

$$\begin{aligned}
\left\|f_\lambda^{\mathbf{u},\mathbf{z}} - P_N f_\lambda\right\|_\rho &= \left\|L_K^{1/2}(L_K^{\mathbf{u}}+\lambda I)^{-1}P_N\frac{1}{m}S_{\mathbf{x}}^T\mathbf{y} - L_K^{1/2}P_N f_\lambda\right\|_K \\
&= \left\|L_K^{1/2}(L_K^{\mathbf{u}}+\lambda I)^{-1/2}P_N(L_K^{\mathbf{u}}+\lambda I)^{-1/2}\left(\frac{1}{m}S_{\mathbf{x}}^T\mathbf{y} - (L_K^{\mathbf{u}}+\lambda I)f_\lambda\right)\right\|_K \\
&= \left\|L_K^{1/2}(L_K^{\mathbf{u}}+\lambda I)^{-1/2}\right\|_{\mathsf{op}}\|P_N\|_{\mathsf{op}}\left\|(L_K^{\mathbf{u}}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}\right\|_{\mathsf{op}} \\
&\quad\times\left\|(L_K+\lambda I)^{-1/2}\left(\frac{1}{m}S_{\mathbf{x}}^T\mathbf{y} - (L_K^{\mathbf{u}}+\lambda I)f_\lambda\right)\right\|_K. \qquad (21)
\end{aligned}$$

The right-hand side of (21) is the product of four norms. Below we bound them one by one. First, obviously $\|P_N\|_{\mathsf{op}} \le 1$. Since L_K is positive semi-definite, for any $f \in \mathcal{H}_K$,

$$\langle f, L_K f\rangle_K \le \langle f, (L_K+\lambda I)f\rangle_K.$$

Therefore we apply Lemma 4.1 to bound the first and the third factor of the right-hand side of (21) by $\Omega_{\mathbf{u},\lambda}^{1/2}$.

$$\begin{aligned}
\left\|L_K^{1/2}(L_K^{\mathbf{u}}+\lambda I)^{-1/2}\right\|_{\mathsf{op}} &= \left\|(L_K^{\mathbf{u}}+\lambda I)^{-1/2}L_K(L_K^{\mathbf{u}}+\lambda I)^{-1/2}\right\|_{\mathsf{op}}^{1/2} \\
&\le \left\|(L_K^{\mathbf{u}}+\lambda I)^{-1/2}(L_K+\lambda I)(L_K^{\mathbf{u}}+\lambda I)^{-1/2}\right\|_{\mathsf{op}}^{1/2} \\
&= \left\|(L_K^{\mathbf{u}}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}\right\|_{\mathsf{op}} \le \Omega_{\mathbf{u},\lambda}^{1/2}. \qquad (22)
\end{aligned}$$

Since $r \ge 1/2$, we cite from [31] the bound that $\|f_\lambda\|_K \le \|g_\rho\|_\rho$. For the last factor of the right-hand side of (21), consider the following decomposition

$$\frac{1}{m}S_{\mathbf{x}}^T\mathbf{y} - (\lambda I + L_K^{\mathbf{u}})f_\lambda = \left(\frac{1}{m}S_{\mathbf{x}}^T\mathbf{y} - L_K^{\mathbf{x}}f_\rho\right) + (L_K^{\mathbf{x}} - L_K)\,f_\rho + (L_K - L_K^{\mathbf{u}})\,f_\lambda,$$

which leads to the bound $R_\lambda^{\mathbf{z}} + \|f_\rho\|_K W_\lambda^{\mathbf{x}} + \|g_\rho\|_\rho W_\lambda^{\mathbf{u}}$ of the fourth term of the right-hand side of (21), and thus completes the proof. □

Lemma 4.3. *Let $1/2 \le r \le 1$ and $\lambda > 0$. We have*

$$\|P_N f_\lambda - f_\lambda\|_\rho \le \Omega_{\mathbf{u},\lambda}^r(\lambda_{N+1}^{\mathbf{u}}+\lambda)^r\|g_\rho\|_\rho. \qquad (23)$$

Proof. Recall that P_N and $L_K^{\mathbf{u}}$ commute. In particular,

$$\begin{aligned}
(I-P_N)(L_K^{\mathbf{u}}+\lambda I)^r &= \left(\sum_{i\ge N+1}\phi_i^{\mathbf{u}}\otimes\phi_i^{\mathbf{u}}\right)\left(\sum_{j\ge 1}(\lambda_j^{\mathbf{u}}+\lambda)^r\phi_j^{\mathbf{u}}\otimes\phi_j^{\mathbf{u}}\right) \\
&= \sum_{j\ge N+1}(\lambda_j^{\mathbf{u}}+\lambda)^r\phi_j^{\mathbf{u}}\otimes\phi_j^{\mathbf{u}},
\end{aligned}$$

so $\|(I-P_N)(L_K^{\mathbf{u}}+\lambda)^r\|_{\mathsf{op}} = (\lambda_{N+1}^{\mathbf{u}}+\lambda)^r$. By Lemma 4.1 and Inequality (22), we have

$$\begin{aligned}
&\|P_N f_\lambda - f_\lambda\|_\rho \\
=&\left\| L_K^{1/2}(I-P_N)L_K^{\frac{1}{2}+r}(L_K+\lambda I)^{-1}L_K^{1/2}g_\rho \right\|_K \\
=&\left\| L_K^{1/2}(L_K^{\mathbf{u}}+\lambda I)^{-1/2}\right\|_{\mathsf{op}} \left\|(L_K^{\mathbf{u}}+\lambda I)^{1/2}(I-P_N)(L_K^{\mathbf{u}}+\lambda I)^{r-\frac{1}{2}}\right\|_{\mathsf{op}} \\
&\times \left\|(L_K^{\mathbf{u}}+\lambda I)^{-(r-\frac{1}{2})}(L_K+\lambda I)^{r-\frac{1}{2}}\right\|_{\mathsf{op}} \left\|L_K^{r+\frac{1}{2}}(L_K+\lambda I)^{-(r+\frac{1}{2})}\right\|_{\mathsf{op}} \|g_\rho\|_\rho \\
\le&\Omega_{\mathbf{u},\lambda}^r(\lambda_{N+1}^{\mathbf{u}}+\lambda)^r \|g_\rho\|_\rho .
\end{aligned}$$

The proof is complete. □

The following lemma is from [18, Proposition 1]. It is a powerful tool recently developed [25, 18] for the analysis of kernel-based regularized least squares and some related algorithms.

Lemma 4.4. *Let $\lambda > 0$ and $0 < \delta < 1$. One has with confidence at least $1-\delta$ that*

$$\Omega_{\mathbf{u},\lambda} \le \frac{2}{\lambda}\mathcal{B}_{n,\lambda}^2 \log^2\frac{2}{\delta} + 2. \tag{24}$$

Denote $\mathsf{HS}(\mathcal{H}_K)$ the Hilbert space of all the Hilbert-Schmidt operators on $\mathcal{H}_K$. Write $\|\cdot\|_{\mathsf{HS}}$ the norm of $\mathsf{HS}(\mathcal{H}_K)$. In the following lemma, Item 1 is the well-known Hoffman-Wielandt inequality [20, 23, 24, 5], and Item 2 is a standard corollary of Pinelis' vector-valued concentration inequality [29]. Detailed proof of Item 2 is available in [36, Proposition 5.3]. See also [22, 2, 8, 9, 15, 21, 25, 31, 36, 24, 34, 39].

Lemma 4.5. *1. We have*

$$\sum_{i=1}^{\infty}(\lambda_i - \lambda_i^{\mathbf{x}})^2 \le \|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}}^2. \tag{25}$$

2. For $0 < \delta < 1$, we have with confidence at least $1-\delta$ that

$$\|L_K - L_K^{\mathbf{x}}\|_{\mathsf{HS}} \le \frac{4\kappa^2}{\sqrt{m}}\log\frac{2}{\delta}. \tag{26}$$

For the following Lemma 4.6, the proof of (27) is available in [8]. The proof of (28) is available in [25, Lemma 17]. The bound (29) follows directly from Lemma 4.5 by substituting $\mathbf{x}$ with $\mathbf{u}$, and $m = |\mathbf{x}|$ with $n = |\mathbf{u}|$.

Lemma 4.6. *Let $0 < \delta < 1$. Each of the following bounds holds with confidence at least $1-\delta$.*

$$R_\lambda^{\mathbf{z}} \le \frac{M+\sigma}{\kappa}\mathcal{B}_{m,\lambda}\log\frac{2}{\delta}, \tag{27}$$

$$W_\lambda^{\mathbf{u}} \le \mathcal{B}_{n,\lambda}\log\frac{2}{\delta}, \quad \textit{and} \tag{28}$$

$$\lambda_i^{\mathbf{u}} \le \lambda_i + \frac{4\kappa^2}{\sqrt{n}}\log\frac{2}{\delta}, \quad \textit{for all } i = 1, 2, \cdots . \tag{29}$$

□

Proof of Theorem 2.1. Recall that $1/2 \le r \le 1$. By Lemma 4.2 and Lemma 4.3,

$$\begin{aligned}\left\|f_\lambda^{\mathbf{u},\mathbf{z}} - f_\rho\right\|_\rho &\le \left\|f_\lambda^{\mathbf{u},\mathbf{z}} - P_N f_\lambda\right\|_\rho + \left\|P_N f_\lambda - f_\lambda\right\|_\rho + \left\|f_\lambda - f_\rho\right\|_\rho \\ &\le \Omega_{\mathbf{u},\lambda}(R_\lambda^{\mathbf{z}} + \|f_\rho\|_K W_\lambda^{\mathbf{x}} + \|g_\rho\|_\rho W_\lambda^{\mathbf{u}}) \\ &\quad + \Omega_{\mathbf{u},\lambda}^r(\lambda_{N+1}^{\mathbf{u}} + \lambda)^r \|g_\rho\|_\rho + \lambda^r \|g_\rho\|_\rho, \end{aligned} \tag{30}$$

where we have used the estimate $\|f_\lambda - f_\rho\|_\rho \le \lambda^r \|g_\rho\|_\rho$ (see [31]). Let $0 < \delta < \frac{1}{5}$, then $\log \frac{2}{\delta} > \log 10 > 1$. From Lemma 4.4 and Lemma 4.6, we have with confidence at least $1 - \delta$ that (24), (27), (28) (for both $W_\lambda^{\mathbf{u}}$ and $W_\lambda^{\mathbf{x}}$ respectively), and (29) hold true simultaneously. Now we assume these five inequalities. Then

$$R_\lambda^{\mathbf{z}} + \|f_\rho\|_K W_\lambda^{\mathbf{x}} + \|g_\rho\|_\rho W_\lambda^{\mathbf{u}} \le \left(\frac{M+\sigma}{\kappa} + \|f_\rho\|_K + \|g_\rho\|_\rho\right) \mathcal{B}_{m,\lambda} \log \frac{2}{\delta}.$$

We combine the argument above and (29) to continue the bound (30).

$$\begin{aligned}\left\|f_\lambda^{\mathbf{u},\mathbf{z}} - f_\rho\right\|_\rho \le &\left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda} + 2\right)\left(\frac{M+\sigma}{\kappa} + \|f_\rho\|_K + \|g_\rho\|_\rho\right) \mathcal{B}_{m,\lambda} \log^3 \frac{2}{\delta} \\ &+ \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda} + 2\right)^r \left(\lambda + \frac{4\kappa^2}{\sqrt{n}} + \lambda_{N+1}\right)^r \|g_\rho\|_\rho \log^{3r} \frac{2}{\delta} + \|g_\rho\|_\rho \lambda^r,\end{aligned}$$

The proof is completed by scaling δ to $\delta/5$. □

Proof of Corollary 2.3. Recall that $1/2 \le r \le 1$, $n \ge m$, and $0 < s \le 1$. With the assumption $\mathcal{N}(\lambda) \le C_1 \lambda^{-s}$ and the setting $\lambda = m^{-\frac{1}{2r+s}}$, (8) implies

$$\mathcal{B}_{n,\lambda} \le \mathcal{B}_{m,\lambda} \le \frac{2\kappa^2}{m} m^{\frac{1/2}{2r+s}} + 2\kappa\sqrt{\frac{C_1}{m} m^{\frac{s}{2r+s}}} \le 2\kappa(\kappa + \sqrt{C_1}) m^{-\frac{r}{2r+s}}, \tag{31}$$

so

$$\frac{\mathcal{B}_{n,\lambda}^2}{\lambda} \le 4\kappa^2(\kappa + \sqrt{C_1})^2 m^{-\frac{2r-1}{2r+s}} \le 4\kappa^2(\kappa + \sqrt{C_1})^2. \tag{32}$$

Recall the assumptions $\lambda_{N+1} \le \kappa^2 \lambda$ and $n \ge m^{\frac{2}{2r+s}}$. Therefore $\frac{1}{\sqrt{n}} \le m^{-\frac{1}{2r+s}} = \lambda$ and

$$\frac{4\kappa^2}{\sqrt{n}} + \lambda_{N+1} \le 5\kappa^2 \lambda.$$

So, Theorem 2.1 implies that

$$\begin{aligned}
\left\|f_\lambda^{\mathbf{u},\mathbf{z}} - f_\rho\right\|_\rho \leq & \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda}+2\right)\left(\frac{M+\sigma}{\kappa}+\|f_\rho\|_K+\|g_\rho\|_\rho\right)\mathcal{B}_{m,\lambda}\log^3\frac{10}{\delta} \\
& + \left(\frac{2\mathcal{B}_{n,\lambda}^2}{\lambda}+2\right)^r\left(\lambda+\frac{4\kappa^2}{\sqrt{n}}+\lambda_{N+1}\right)^r\|g_\rho\|_\rho\log^{3r}\frac{10}{\delta}+\|g_\rho\|_\rho\lambda^r \\
\leq & (8\kappa^2(\kappa+\sqrt{C_1})^2+2)(2\kappa(\kappa+\sqrt{C_1})) \\
& \times\left(\frac{M+\sigma}{\kappa}+\|f_\rho\|_K+\|g_\rho\|_\rho\right)m^{-\frac{r}{2r+s}}\log^3\frac{10}{\delta} \\
& + (8\kappa^2(\kappa+\sqrt{C_1})^2+2)^r(1+5\kappa^2)^r\|g_\rho\|_\rho m^{-\frac{r}{2r+s}}\log^{3r}\frac{10}{\delta}+\|g_\rho\|_\rho m^{-\frac{r}{2r+s}} \\
\leq & C_2 m^{-\frac{r}{2r+s}}\log^3\frac{10}{\delta},
\end{aligned}$$

where $C_2 = (8\kappa^2(\kappa+\sqrt{C_1})^2+2)(2\kappa(\kappa+\sqrt{C_1}))\left(\frac{M+\sigma}{\kappa}+\|f_\rho\|_K+\|g_\rho\|_\rho\right)+(8\kappa^2(\kappa+\sqrt{C_1})^2+2)^r(1+5\kappa^2)^r\|g_\rho\|_\rho+\|g_\rho\|_\rho$. □

5 Acknowledgment

We would like to acknowledge Professor Jian Huang for the helpful discussions, in particular, the introduction of the works [38, 28] to us. The work described in this paper is supported partially by the Research Grants Council of Hong Kong [Project No. PolyU 15304917]. The corresponding author is Xin Guo.

References

[1] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68:337–404, 1950.

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

[3] Andrea L. Bertozzi, Xiyang Luo, Andrew M. Stuart, and Konstantinos C. Zygalakis. Uncertainty quantification in graph-based classification of high dimensional data. *SIAM/ASA J. Uncertain. Quantif.*, 6(2):568–595, 2018.

[4] Rajendra Bhatia. *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[5] Rajendra Bhatia and Ludwig Elsner. The Hoffman-Wielandt inequality in infinite dimensions. *Proc. Indian Acad. Sci. Math. Sci.*, 104(3):483–494, 1994.

[6] Gilles Blanchard and Nicole Krämer. Optimal learning rates for kernel conjugate gradient regression. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel,

and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 226–234. Curran Associates, Inc., 2010.

[7] Gilles Blanchard and Nicole Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Anal. Appl.*, 14(6):763–794, 2016.

[8] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.*, 7(3):331–368, 2007.

[9] Andrea Caponnetto and Yuan Yao. Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl. (Singap.)*, 8(2):161–183, 2010.

[10] Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.*, 18:Paper No. 46, 22, 2017.

[11] Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, 2011.

[12] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Appl. Comput. Harmon. Anal.*, 21(1):5–30, 2006.

[13] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49, 2002.

[14] Felipe Cucker and Ding-Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2007. With a foreword by Stephen Smale.

[15] Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, and Francesca Odone. Learning from examples as an inverse problem. *J. Mach. Learn. Res.*, 6:883–904, 2005.

[16] Xin Guo, Ting Hu, and Qiang Wu. Distributed minimum error entropy algorithms. Preprint, 2019.

[17] Xin Guo and Ding-Xuan Zhou. An empirical feature-based learning algorithm producing sparse approximations. *Appl. Comput. Harmon. Anal.*, 32(3):389–400, 2012.

[18] Zheng-Chu Guo, Shao-Bo Lin, and Ding-Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 33(7):074009, 29, 2017.

[19] Zheng-Chu Guo, Lei Shi, and Qiang Wu. Learning theory of distributed regression with bias corrected regularization kernel network. *J. Mach. Learn. Res.*, 18:Paper No. 118, 25, 2017.

[20] A. J. Hoffman and H. W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Math. J.*, 20:37–39, 1953.

[21] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Found. Comput. Math.*, 14(3):569–600, 2014.

[22] Ting Hu, Jun Fan, Qiang Wu, and Ding-Xuan Zhou. Regularization schemes for minimum error entropy principle. *Anal. Appl. (Singap.)*, 13(4):437–455, 2015.

[23] Tosio Kato. Variation of discrete spectra. *Comm. Math. Phys.*, 111(3):501–504, 1987.

[24] Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, 6(1):113–167, 2000.

[25] Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *J. Mach. Learn. Res.*, 18:Paper No. 92, 31, 2017.

[26] Shao-Bo Lin and Ding-Xuan Zhou. Distributed kernel-based gradient descent algorithms. *Constr. Approx.*, 47(2):249–276, 2018.

[27] G. Little and J. B. Reade. Eigenvalues of analytic kernels. *SIAM J. Math. Anal.*, 15(1):133–136, 1984.

[28] Jin Liu, Can Yang, Yuling Jiao, and Jian Huang. ssLasso: a summary-statistic-based regression using Lasso. Preprint, 2017.

[29] Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22(4):1679–1706, 1994.

[30] J. B. Reade. Eigenvalues of positive definite kernels. II. *SIAM J. Math. Anal.*, 15(1):137–142, 1984.

[31] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

[32] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

[33] Ingo Steinwart, Don R. Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[34] Ulrike von Luxburg, Mikhail Belkin, and Olivier Bousquet. Consistency of spectral clustering. *Ann. Statist.*, 36(2):555–586, 2008.

[35] Grace Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[36] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.

[37] Tong Zhang. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, pages 454–461, 2002.

[38] Xiang Zhu and Matthew Stephens. Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.*, 11(3):1561–1592, 2017.

[39] Laurent Zwald and Gilles Blanchard. On the convergence of eigenspaces in kernel principal component analysis. In *Advances in neural information processing systems*, pages 1649–1656, 2006.

[40] Laurent Zwald, Gilles Blanchard, Pascal Massart, and Régis Vert. Kernel projection machine: a new tool for pattern recognition. In *Advances in Neural Information Processing Systems*, pages 1649–1656, 2005.