Journal of Machine Learning Research 27 (2026) 1-28 Submitted 12/23; Revised 7/25; Published 3/26

Kernel-based Distributed Learning Beyond Least Squares

Heng Lian HENGLIAN@CITYU.EDU.HK
Department of Mathematics
City University of Hong Kong
Hong Kong, China

Xu Guo XUSTAT12@BNU.EDU.CN
School of Statistics
Beijing Normal University
Beijing, 100875, China

Editor: Sanmi Koyejo

Abstract

We consider one-shot distributed learning problems in a reproducing kernel Hilbert space framework. Current results are limited to the least-squares loss and extensions beyond this meet with some significant technical challenges. We establish the optimal rate of distributed learning for some general class of convex loss functions satisfying mild assumptions, using a novel empirical process on the Bregman divergence induced by the loss, which is essential for carrying out a quadratic approximation in the infinite-dimensional space. The empirical process is bounded by relating the Bregman divergence induced by the loss to the supremum norm and the L^2-norm of the functions. This framework incorporates many commonly used losses, including strongly smooth loss functions as well as Lipschitz continuous losses such as the quantile loss.

Keywords: Bregman divergence; Covering number; Divide-and-conquer; Learning rate; Rademacher complexity.

1. Introduction

In supervised learning problems, including regression and classification, the aim is to predict future outcomes based on given inputs using a learned function. A basic example is the prediction of the conditional mean of the response given a predictor. In the machine learning literature, a popular approach of nonparametric learning is based on the construction of a kernel function that induces a reproducing kernel Hilbert space (RKHS). For the least-squares loss, the early work of Caponnetto and De Vito (2007) obtained the minimax risk bound of the estimator based on Tikhonov regularization (also known as ridge regression), using both the source condition and the capacity condition. The results are significantly refined by Steinwart et al. (2009). Although general kernel-based learning has achieved significant advancements on an impressively wide range of problems over the past several decades (Mika et al., 1999; Lai and Fyfe, 2000; Shawe-Taylor et al., 2005; Shawe-Taylor, 2008), relatively complete theoretical results to characterize its performance are still only commonly found for the least-squares loss. Notable exceptions include for example Li et al. (2007) which dealt with quantile regression, and Steinwart and Scovel (2007) for the hinge loss used in the support vector machines. For smooth losses that are at least thrice

differentiable, Marteau-Ferey et al. (2019) used self-concordance assumption to derive the optimal estimation rate under *both source and capacity assumptions.*

Standard implementations of Tikhonov regularized learning algorithms all suffer from computational burdens for a very large sample size N. For example, to get the closed-form solution of the kernel ridge regression estimator with a least-squares loss, the inverse of the kernel matrix is computed, which usually has a time complexity of $O(N^3)$. Gradient-based methods can be used which scale better with the sample size, and works for general convex losses, but still has a time complexity of at least $O(N^2T)$ where T is the number of iterations. This motivates one to consider distributed learning algorithms (Zhang et al., 2013, 2015; Rosenblatt and Nadler, 2015; Balcan et al., 2015; Chang et al., 2017; Lee et al., 2017; Lin et al., 2017; Jordan et al., 2018; Volgushev et al., 2019; Lian and Fan, 2018). Here we focus on the one-shot data-parallel learning, which is an example of the divide-and-conquer strategy. The basic idea is very simple. We randomly divide a data set of size N into m subsets of equal sizes and compute an estimate using a certain algorithm on each subset/partition, and then take an average of the m 'local estimates' to get the final global estimate. In the distributed setting, different subsets can be dealt with on different machines that can communicate over a network, and the communication only needs to happen once at the end (thus the term 'one-shot learning'). Typically, when m does not diverge too fast compared to N, the averaged estimate can be shown to have the same risk bound as the central estimator (the one that inputs all data directly to the regularized algorithm). For kernel-based learning with the least-squares loss, this idea has been implemented and studied in several works (Lin et al., 2017; Guo et al., 2017; Chang et al., 2017; Lin and Cevher, 2020) based on Tikhonov regularization and also more general spectral regularization algorithms. These results critically depend on the availability of the closed-form expression of the local estimates, which makes it possible to study the properties of the averaged estimate which also has a closed-form expression. However, distributed estimation for kernel-based learning beyond the least-squares case has so far not been investigated, possibly due to the associated technical challenges caused by the fact that a closed-form solution is not available.

In this work, we consider Tikhonov regularized distributed learning in the RKHS setting for more general losses, with the statistical rate characterized by the source condition and the capacity condition as in the least-squares case. The main technical innovation is to use a new empirical process indexed by the Bregman divergence between a function and the truth (Lemma 3), which can be of a smaller order than the more commonly seen empirical process indexed by the risk difference (Lemma 4). See Remark 2 for some explanations why it is expected to be smaller. This bound then allows us to approximate the local estimates by a weighted least-squares estimate, for which the bound for the averaged estimate using the divide-and-conquer strategy can be more easily obtained. In particular, we can obtain the distributed learning rate which is the same as the rate for the central estimator, when the number of partitions is sufficiently small. Our results, in particular, include logistic regression and quantile regression, two primary examples used throughout the paper for more detailed illustrations.

Methodologically, our study is a straightforward extension of the pioneering work Zhang et al. (2015) for distributed learning in the RKHS, extending it to dealing with more general losses. However, this makes the proof very different from the least squares case. For example, Proof of Theorem 1 in Zhang et al. (2015) is based on a direct decomposition

into bias and variance, which is only possible for the least squares loss. Furthermore, their proofs proceed by explicitly expanding functions in terms of the eigenfunctions (such as in their Lemmas 6,8,9,10), which is part of the reason they require moment assumptions on eigenfunctions as in their Assumption A (or stronger boundedness assumption on eigenfunctions as in their Assumption A'), which we do not need (although we need other stronger assumptions for dealing with more general losses). For general losses, such an explicit decomposition seems not possible and our strategy is to show that the estimator is close to a carefully constructed weighted least squares problem. For this, we need an entropy bound for the class of functions containing the Bregman divergence as mentioned above, which requires a careful analytic construction as in our Lemmas 1 and 2. Equipped with this bound, our main result Theorem 2 uses some careful algebraic manipulations to show the distributed estimator is sufficiently close to the weighted least squares estimator. We also note that another related work Shang and Cheng (2013) used a similar general approach of using empirical processes techniques in the non-distributed setting. However, their results require thrice differentiability of the loss and thus cannot be directly applied to quantile loss or Huber's loss. Their proofs also rely on explicit expansion in terms of eigenfunctions and assume the eigenfunctions are uniformly bounded.

Our main contribution is a theoretical characterization of the averaged estimator in distributed learning for a wide range of loss functions, but we note that some results are new even in the classical non-distributed setting. For example, we provide learning rates that match the least-squares case for quantile loss which incorporates *both the source condition and the capacity condition*, which is not included in the framework of Marteau-Ferey et al. (2019) or Li et al. (2007). The current work is related to our conference paper (Lian, 2022), which focused on the quantile loss. Compared to the conference paper version, the new significant contributions include the following.

- We cover much more general losses, including logistic loss and Huber loss, using a unified treatment involving empirical processes bound on the Bregman divergence functions. As we see from our theoretical results, although logistic loss and Huber loss are also Lipschitz, it is better to treat them as smooth losses satisfying our assumption (A1)(i), due to that this can lead to stronger results. As we will see in the proof, smooth losses are technically more complicated to deal with due to that the error term is unbounded, while Lipschitz losses, such as the quantile loss, lead to a simpler bound.

- Using more careful arguments concerning the weighted least squares approximation, we are able to remove the unreasonable requirement that $\alpha \geq 3$ (see Remark 1 of Lian (2022)), making the theoretical result cleaner and more elegant.

The rest of the article is organized as follows. In Section 2, we present the Tikhonov regularized learning problem for convex losses, and explain the assumptions we use in some detail. The main results, together with their proofs, are presented in Section 3. Section 4 reports some numerical results on Huber's loss and logistic regression. We conclude with some discussions in Section 5.

Notation: the RKHS in this paper is always denoted by $\mathcal{H}$, with its inner product given by $\langle ., . \rangle_{\mathcal{H}}$ which induces the RKHS norm $\|.\|_{\mathcal{H}}$. ρ denotes the probability measure on the

input-output pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and ρ_x denotes the marginal distribution on $\mathcal{X}$. By convention, we also use P to denote the same probability measure as ρ, while the empirical counterpart is P_n or P_N for sample size n and N, respectively. For a function on $\mathcal{X} \times \mathcal{Y}$, $\|.\|$ denotes the $L^2(\rho)$ norm, which is just the square root of the second moment. For a random variable, $\|.\|_{\psi_1}$ denotes its Orlicz norm corresponding to the function $\psi_1(x) = e^x - 1$. For positive sequences a_n, b_n, both $a_n \lesssim b_n$ and $a_n = O(b_n)$ means a_n/b_n is bounded, and $a_n \asymp b_n$ means $a_n = O(b_n)$ and $b_n = O(a_n)$.

2. One-shot distributed learning in the RKHS

The function space $\mathcal{H} \subseteq L^2(\rho_x)$ we consider is a reproducing kernel Hilbert space (RKHS) containing functions with a compact domain $\mathcal{X}$, where ρ_x is a probability distribution on $\mathcal{X}$. The corresponding bivariate kernel function is denoted by $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The kernel function K defines an integral operator $\mathcal{L} : f \in \mathcal{H} \to \int K(x, .)f(x)d\rho_x$, which is a non-negative-definite operator whose eigenvalues $s_j, j = 1, 2, \ldots$, characterizes the complexity of $\mathcal{H}$ in learning theory. We also note that $K(x, .) \in \mathcal{H}, \forall x \in \mathcal{X}$ and $\langle K(x, .), f\rangle_{\mathcal{H}} = f(x), \forall f \in \mathcal{H}, x \in \mathcal{X}$ (the reproducing property).

Given a loss function $\ell(y, z)$ that is convex in z and an i.i.d. sample $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1, \ldots, N$, with $\mathcal{Y}$ denoting the support of y, the learned function in $\mathcal{H}$ is given by

$$\widehat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2, \tag{1}$$

where $\|.\|_{\mathcal{H}}$ denotes the RKHS norm. The unknown truth $f_0 = \arg\min E[\ell(y, f(x))]$ is assumed to be uniquely defined and $f_0 \in \mathcal{H}$. Two primary examples of the loss function include the logistic loss $\ell(y, f(x)) = h(yf(x)) := \log(1 + e^{-yf(x)}), y \in \{-1, 1\}$, and the quantile loss $\ell(y, f(x)) = \rho_\tau(y - f(x))$ with $\rho_\tau(u) = u(\tau - I\{u \le 0\})$. More generally, the condition on the loss function we consider in this work is specified by assumption (A1) below.

To compute $\widehat{f}$, one can resort to the Representer Theorem (Wahba, 1990), which states that $\widehat{f}$ will take the form

$$\widehat{f} = \sum_{i=1}^{N} a_i K(x_i, .), \tag{2}$$

for some unknown coefficients a_i that can be determined by solving (1) with (2) plugged into it. However, as mentioned in the introduction, when N is large, all algorithms will suffer from heavy computational burdens. In the distributed setting, the entire sample is partitioned into m subsets of equal sizes, distributed onto m machines and each machine computes a local estimate

$$\widehat{f_j} = \arg\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i \in \mathcal{S}_j} \ell(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2, \ j = 1, \ldots, m, \tag{3}$$

where $\mathcal{S}_j$ is the j-th subset with size $n = N/m$. The final averaged estimator is then

$$\bar{f} = \frac{1}{m} \sum_{j=1}^{m} \widehat{f_j}.$$

The main theoretical question concerning $\bar{f}$ is whether and when the statistical convergence rate of $\bar{f}$ will be the same as that of $\widehat{f}$, the latter of which uses all data directly and is called the central estimator.

In practice, it is important to choose the tuning parameter λ. In our simulation studies, we directly assume the central machine has a set of hold-out data that can be used for tuning. In practice, when such data are not present, we could do cross-validation (CV), which however requires more communication. More specifically, in 5-fold CV for example, each local machine j uses 80% of the local data to obtain $\widehat{f}_{j,\lambda}$, for $\lambda \in \Lambda$, where Λ is a finite set containing all potential tuning parameters used. All local estimates $\widehat{f}_{j,\lambda}, j = 1, \ldots, m, \lambda \in \Lambda$ are then sent to the central machine to obtain $\bar{f}_\lambda, \lambda \in \Lambda$. These estimates are then sent back to local machines which evaluate the cross-validation error based on 20% of the data not used in training, and the errors are sent back to the central machine. Finally, the server can determine which λ has the smallest cross-validation error. This is feasible but of course more time consuming in simulation. Note that assuming hold-out data is available roughly means we only partition the entire data into training the testing parts once, rather than five times, which speeds up the numerical procedure.

For smoothing splines (a special case of RKHS), Sun et al. (2021) provides a clever way to choose the tuning parameter when the sample size N is large. They proposed a two-step procedure. First, based on a random subsample of size n, some appropriate tuning parameter λ_n is obtained. Then one sets $\lambda = \lambda_n(N/n)^\nu$ for some ν. Here, ν is either a known constant, or needs to be estimated. In either case, this strategy seems to depend heavily on the specific smoothing splines method used. In our case, it seems challenging to estimate ν (in particular it depends on the smoothness of the true function as in our Assumption (A2) below). Thus at least the method needs to be further developed in order to be used in our setting.

Theoretical studies in an RKHS beyond the least-squares loss are scarce, even for the non-distributed setting. For a convex Lipschitz loss, Sridharan et al. (2009) derived a general non-asymptotic bound $O(1/N)$ for the excess risk, which is generally optimal without further assumptions. In order to obtain faster rates, for the least-squares loss, two conditions are often imposed. First, the faster decay rate of the eigenvalues of $\mathcal{L}$, or the smaller entropy number of the reproducing kernel Hilbert space (called the capacity condition), leads to an improved variance term, while a further smoothness on the true f_0 (called the source condition) leads to an improved bias term. Marteau-Ferey et al. (2019) incorporated these two conditions based on self-concordant analysis. Although this is an elegant and pioneering work, the definition of self-concordance requires the loss to be at least thrice differentiable and thus the framework excludes quantile loss and Huber's loss, for example. More importantly, these works did not consider the distributed setting. To achieve rates faster than $1/N$, we also make some assumptions on the source condition and the capacity condition.

Throughout the paper, C denotes a generic positive constant. C with subscripts such as $C_1, C_2, \ldots$ are used to denote specific constants to make the presentation clearer. The assumptions we used are listed below, followed by detailed discussions.

(A1) The loss $\ell(y, z)$ is convex in z and satisfies either one of the following two conditions.

(i) ℓ is strongly smooth with parameter C, uniformly in y, in the sense that ℓ is continuously differentiable in z and

$$|\partial_2\ell(y,z') - \partial_2\ell(y,z)| \le C|z'-z|, \forall y \in \mathcal{Y}, |z| \le 1, |z'| \le 1,$$

where $\mathcal{Y}$ is the support of y and $\partial_2\ell(y,z)$ denotes the partial (sub-)derivative with respect to z.

(ii) ℓ is Lipschitz continuous in z, uniformly in $y \in \mathcal{Y}$, in the sense that

$$|\partial_2\ell(y,z)| \le C, \forall y \in \mathcal{Y}, |z| \le 1.$$

(A2) $\sup_{x\in\mathcal{X}} K(x,x) < \infty$. We assume the following 'source condition' holds:

$$f_0 = \mathcal{L}^r g_0 \text{ for some } g_0 \in \mathcal{H}, r \in [0,1/2], \text{ with } \|g_0\|_{\mathcal{H}} \le 1. \tag{4}$$

(A3) $\epsilon := \partial_2\ell(y, f_0(x))$ is a mean-zero sub-exponential random variable.

(A4) (Capacity condition) We have the entropy condition $\log\mathcal{N}(\varepsilon, \mathcal{H}(1), L^2(P_n)) \le (C/\varepsilon)^{2/\alpha}$ for some $\alpha > 1$, where $\mathcal{H}(1) = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1\}$ and $\mathcal{N}(\varepsilon, \mathcal{H}(1), L^2(P_n))$ denotes the ε-covering number.

(A5) (Sup-norm assumption) For some $s \in (0,1]$, we have $\|f\|_\infty \le C_1\|f\|^{1-s}\|f\|_{\mathcal{H}}^s$, $\forall f \in \mathcal{H}$.

(A6) $E\ell(y,f(x)) - E\ell(y,f_0(x)) = \frac{1}{2}\langle f-f_0, \mathcal{L}_{\mathcal{H}}(f-f_0)\rangle_{\mathcal{H}} + O(E[(f(x)-f_0(x))^3]), \forall f \in \mathcal{H}$, for some operator $\mathcal{L}_{\mathcal{H}} : \mathcal{H} \to \mathcal{H}$ with $C_2\mathcal{L} \le \mathcal{L}_{\mathcal{H}} \le C_3\mathcal{L}$ for two positive constants C_2, C_3.

The source condition is about the smoothness of the true function f_0, while the capacity assumption and the sup-norm assumption are related to the RKHS used. These assumptions do not depend on the loss. Other assumptions including (A1),(A3) and (A6) are about the loss function. Below we provide more detailed explanations for these assumptions.

The assumption (A1) on the loss includes most popular convex losses in supervised learning. The least-squares loss obviously satisfies (i). Quantile loss and hinge loss satisfy (ii). The logistic loss and the Huber's loss satisfy both (i) and (ii). We will see that case (i) leads to less stringent bounds on the number of machines allowed, and thus we will regard these losses as case (i). In (A2), since we assume $\mathcal{X}$ is compact, $\sup_x K(x,x) < \infty$ as soon as K is continuous. This is a very mild assumption used in almost all existing studies. An important consequence is that the unit ball in the RKHS is bounded in supremum norm, since $\|f\|_\infty = \sup_x \langle K(x,.), f\rangle_{\mathcal{H}} \le \sup_x K(x,x)\|f\|_{\mathcal{H}}$. The source condition in (A2) is the same as that used in Lin et al. (2017). The constant 1 in $\|g_0\|_{\mathcal{H}} \le 1$ does not have any significance and can be replaced by any other fixed constant. Note that we merely assume such a constant bound for $\|g_0\|_{\mathcal{H}}$ exists and this constant does not have to be known to the user since it is not used in the optimization (1). In (A3), due to the first-order condition, $\partial_2\ell(y, f_0(x))$ automatically has mean zero. For the least-squares loss, assuming it has sub-exponential tails reduces to saying the additive error in the mean regression model is sub-exponential. This assumption is automatically true for the logistic loss, Huber loss, and any Lipschitz loss satisfying (ii), for which $\partial_2\ell(y, f_0(x))$ is actually a bounded random variable.

(A4) is slightly stronger than that often assumed in the literature concerning least-squares loss. It is related to, but stronger than, the common assumption that the eigenvalues of $\mathcal{L}$ decay with rate $s_j \sim j^{-\alpha}$. As stated in Steinwart et al. (2009), this eigenvalue assumption is equivalent to assuming (A4) but with $L^2(P_n)$ replaced by $L^2(P)$. However, it turns out to be harder to give general sufficient conditions for (A4) with the empirical measure P_n. Of course, given the $L^2(P_n)$ norm is bounded by the supremum norm, one could also replace $L^2(P_n)$ in (A4) with the $\|.\|_\infty$ metric. It is known to be satisfied for certain Sobolev spaces, but we do not find other examples in the literature. In fact, as mentioned in Steinwart et al. (2009), for m-times differentiable kernels on Euclidean balls of $\mathbb{R}^d$, it is known that (A4) holds for $\alpha = 2m/d$. In particular, it holds for the Sobolev space on the unit interval. Such an assumption is also used in Müller and van de Geer (2015), which used the supremum norm instead of the $L^2(P_n)$. We need (A4) to bound the empirical process for the Bregman divergence as in Lemma 3 below. For the least squares loss, we have $D_{f,f_0}(x,y) = \frac{1}{2}(f(x)-f_0(x))^2$. Thus $(P_n-P)D_{f,f_0} = \langle f-f_0, (\mathcal{L}_n-\mathcal{L})(f-f_0)$, where $\mathcal{L}_n$ is the empirical version of $\mathcal{L}$ (sample average). Thus we can directly study the properties of $\mathcal{L}_n - \mathcal{L}$, which is independent of f, making the use of empirical processes tools unnecessary. For losses other than the least squares loss, we apply bounds based on entropy number in $L^2(P_n)$ norm, which is slightly weaker than using L^∞ norm, and thus (A4) is required in our proof.

The sup-norm assumption (A5) is also used in Suzuki and Sugiyama (2013). It holds for $s = 1$ as soon as $\sup_x K(x,x)$ is bounded. Smaller s typically yields a better control for the sup-norm of a function in the RKHS. Steinwart et al. (2009) discussed the sup-norm assumption in detail. Again, this is closely related to assuming $s_j \sim j^{-\alpha}$ with $\alpha = 1/s$. For the Sobolev space of order m, the sup-norm assumption holds for $s = d/(2m)$ under mild regularity assumptions. Since the RKHSs of the Gaussian kernels are continuously embedded in all Sobolev spaces, it also satisfies the sup-norm assumption for all $s \in (0,1]$.

(A6) provides a quadratic approximation of the expected risk difference. On a high level, Assumption (A6) makes it possible to get a weighted least squares approximation for the estimator, by assuming an asymptotic expansion for the expected risk difference. In appearance, one might expect that (A6) would make the proof close to that for the least square problem, but this is not the case. On one hand, we still cannot have an explicit bias-variance decomposition as in the least squares case. On the other hand, it is stated for an expected loss which is not directly applicable to the empirical risk minimization problem. Instead, the empirical processes technique plays a key role to build a bridge between the empirical loss and the expected loss, with the difference being of higher order than the main stochastic error term.

We note that (A6) is not particularly restrictive. Intuitively, any smooth function would behave like a quadratic function near its minimum, which is a direct consequence of Taylor's expansion. Thus, assumption (A6) does not impose stringent restrictions on the loss function. To make it also applicable for nonsmooth losses, notably, we impose this assumption on the expected loss rather than the empirical loss. This makes it applicable for non-smooth losses including the quantile loss, for example. We verify in the following proposition that (A6) holds for logistic loss and quantile loss, although it is expected to hold also for many other losses. Unfortunately owing to fundamental analytical difficulties, we are unable to accommodate hinge loss in our theoretical framework.

Proposition 1 *(A6) holds for the logistic loss and the quantile loss. For the latter, we assume the conditional density of $\epsilon = y - f_0(x)$, $q(.|x)$, as well as its first derivative, is uniformly bounded, and $q(0|x)$ is uniformly bounded away from zero.*

3. Main results

We first state several useful lemmas in preparation for the proof of the main result. As mentioned in the introduction, our key innovation is to study the empirical process on the class of functions induced by the Bregman divergence, which is done in Lemma 3. Lemma 1 obtained some bounds for the Bregman divergence and Lemma 2 calculates the entropy bound for this class of functions, to be used in the proof of Lemma 3.

Define the Bregman divergence $D_{f_1,f_2}(x,y) = \ell(y, f_1(x)) - \ell(y, f_2(x)) - \partial_2\ell(y, f_2(x))(f_1(x) - f_2(x))$, where ∂_2 is the (sub-)derivative with respect to the second argument of ℓ. The following result relates the supremum norm and the L^2-norm of the Bregman divergence $D_{f,f_0}(x,y)$, to the L^2-norm and supremum norm of $f - f_0$.

Lemma 1 *Assume (A6) holds. For the loss satisfying (A1)(i),*

$$\begin{aligned} &\|D_{f,f_0}(x,y)\|_\infty \le C\|f - f_0\|^a \|f - f_0\|_\infty^b \\ &\|D_{f,f_0}(x,y)\| \le C\|f - f_0\|^{a'} \|f - f_0\|_\infty^{b'} \end{aligned} \tag{5}$$

for all $f \in \mathcal{H}$ with $\|f - f_0\|_\infty \le 1$, with $a = 0, b = 2, a' = b' = 1$. For the loss satisfying (A1)(ii) we have the above holds with $a = 0, b = 1, a' = 1, b' = 1/2$.

Remark 1 *Although the above Lemma gives specific values of a, b, a', b' in (5), for most parts of the proof we will just use letters a, b, a', b' in the proof for generality. Provided that some upper bounds for $\|D_{f,f_0}(x,y)\|_\infty$ and $\|D_{f,f_0}(x,y)\|$ in terms of $\|f - f_0\|_\infty$ and $\|f - f_0\|$ are available, we can establish an upper bound for the empirical process index by $D_{f,f_0}(x,y)$ as in Lemma 3 below. In particular, note that in our examples we always have $a = 0$.*

We also need a result on the covering number for the class $\{D_{f,f_0}(x,y)\}$ as in the lemma below.

Lemma 2 *Assume (A1), (A4) and (A5). Let $\mathcal{D}(u,v) = \{D_{f,f_0}(x,y) : \|f - f_0\| \le u, \|f - f_0\|_{\mathcal{H}} \le v\}$. We have $\mathcal{N}(\epsilon, \mathcal{D}(u,v), L^2(P_n)) \le e^{H(\epsilon,u,v)}$, with $H(\epsilon,u,v) = (Cvw/\epsilon)^{2/\alpha}$ where $w = u^{1-s}v^s$, if the loss satisfies (A1) (i), and $H(\epsilon,u,v) = (Cv/\epsilon)^{2/\alpha}$ if the loss satisfies (A1) (ii).*

Using Lemmas 1 and 2, we can obtain a bound for the empirical process.

Lemma 3 *Assume (A1)-(A6). For any $u, v > 0$, with probability $1 - \exp\{-H(c', u, v)\}$,*

$$\sup_{\|f-f_0\|\le u, \|f-f_0\|_{\mathcal{H}} \le v} (P_n - P)D_{f,f_0}(x,y) \le C\left(\frac{c'}{\sqrt{n}}\sqrt{H(c',u,v)} + \frac{c}{n}H(c',u,v)\right),$$

where $c' = Cu^{a'+(1-s)b'}v^{sb'}$, $c = Cu^{a+(1-s)b}v^{sb}$.

The following lemma concerns another (more commonly used) empirical process for the risk difference. It uses the Rademacher complexity $E[\sup_{\|f\|\le u,\|f\|_{\mathcal{H}}\le 1}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f(x_i)]$, where $\sigma_i \in \{-1,1\}$ are i.i.d. Rademacher variables. As shown in Mendelson (2002), we have a bound

$$E[\sup_{\|f\|\le u,\|f\|_{\mathcal{H}}\le 1}\frac{1}{N}\sum_{i=1}^{N}\sigma_i f(x_i)] \le C\mathcal{R}_N(u), \tag{6}$$

where $\mathcal{R}_N(u)=\sqrt{\frac{1}{N}\sum_{j=1}^{\infty}\min\{s_j,u^2\}}$. In the special case that $s_j \asymp Cj^{-\alpha}$, it is easy to see we can take $\mathcal{R}_N(u)=\frac{u^{1-\frac{1}{\alpha}}}{\sqrt{N}}$.

Lemma 4 *Assume (A1)-(A6). For any $u,v>0$, with probability at least $1-e^{-N\mathcal{R}_N^2(u/v)/(u/v)^2}-e^{-H(c',u,v)}$,*

$$\begin{aligned}&\sup_{\|f-f_0\|\le u,\|f-f_0\|_{\mathcal{H}}\le v}(P_N-P)\{\ell(y,f(x))-\ell(y,f_0(x))\}\\ \le\ & C\left(v\mathcal{R}_N(u/v)+\frac{c'}{\sqrt{N}}\sqrt{H(c',u,v)}+\frac{c}{N}H(c',u,v)+u^{1-s}v^s\frac{\mathcal{R}_N^2(u/v)}{(u/v)^2}\log N\right),\end{aligned} \tag{7}$$

where $c=u^{a+(1-s)b}v^{sb}$, $c'=u^{a'+(1-s)b'}v^{sb'}$.

In case of (A1)(ii), we have a simpler bound

$$\sup_{\|f-f_0\|\le u,\|f-f_0\|_{\mathcal{H}}\le v}(P_N-P)\{\ell(y,f(x))-\ell(y,f_0(x))\} \le C\left(v\mathcal{R}_N(u/v)+u^{1-s}v^s\frac{\mathcal{R}_N^2(u/v)}{(u/v)^2}\right), \tag{8}$$

with probability at least $1-e^{-N\mathcal{R}_N^2(u/v)/(u/v)^2}$.

Note that trivially the bounds also hold if the sample size N is replaced by n everywhere in the above (including replacing $\mathcal{R}_N$ by $\mathcal{R}_n$), which is useful in the distributed setting.

The final lemma is a strong local convexity property of the population loss, which is an immediate corollary of assumption (A6), and thus its proof is omitted.

Lemma 5 *Assume (A6) holds. Then, there exists constants C_4, C_5 such that $E[\ell(y,f(x))]-E[\ell(y,f_0(x))]\ge C_4\|f-f_0\|^2$, $\forall f\in\mathcal{H}$ with $\|f-f_0\|_\infty\le C_5$.*

Theorem 1 gives the bound for the central estimator with sample size N, as well as the local estimator based on one partition/machine with sample size n. Note that in order for the averaged estimator $\bar{f}$ to achieve the optimal rate of convergence, λ for the local estimators is chosen to be the same as that for the central estimator. Such a choice is the same as that already noted for the least-squares case (Lin et al., 2017). For the following, we note a unique positive u_N satisfying $\mathcal{R}_N(u_N)=u_N^{2+2r}$ exists as explained in Bartlett et al. (2005).

Theorem 1 *Assume (A1)-(A6) hold. Set $\lambda \asymp u_N^2$ with the positive value u_N satisfying $\mathcal{R}_N(u_N)=u_N^{2+2r}$ and. Assume the first term $v\mathcal{R}_N(u/v)$ in the bound of (7) (for (A1)(i))*

and (8) *(for (A1)(ii)) is the dominating term when* $u \asymp u_N^{1+2r}, v \asymp u_N^{2r}$*, and* $C_5/(C_1 u_N^{1+2r-s})$ *is sufficiently large, we have*

$$\|\widehat{f} - f_0\| + u_N \|\widehat{f} - f_0\|_{\mathcal{H}} \leq C u_N^{1+2r},$$

with probability at least $1 - e^{-N\mathcal{R}_N^2(u/v)/(u/v)^2} - e^{-H(c',u,v)}$ *where* $u \asymp u_N^{1+2r}, v \asymp u_N^{2r}$.

For the distributed setting where the local sample size is n*, we still set* $\lambda \asymp u_N^2$ *with* $\mathcal{R}_N(u_N) = u_N^{2+2r}$*. Assume the first term* $v\mathcal{R}_n(u/v)$ *in the bound of* (7) *(for (A1)(i)) and* (8) *(for (A1)(ii)) is the dominating term (all appearances of* N *in the bounds are replaced by* n*) when* $u \asymp \sqrt{N/n} u_N^{1+2r}, v \asymp \sqrt{N/n} u_N^{2r}$*, and that* $C_5/(C_1\sqrt{N/n} u_N^{1+2r-s})$ *is sufficiently large, then we have*

$$\|\widehat{f}_j - f_0\| + u_N \|\widehat{f}_j - f_0\|_{\mathcal{H}} \leq C\sqrt{\frac{N}{n}} u_N^{1+2r},$$

with probability at least $1 - e^{-N\mathcal{R}_N^2(u/v)/(u/v)^2} - e^{-H(c',u,v)}$ *where* $u \asymp \sqrt{N/n} u_N^{1+2r}, v \asymp \sqrt{N/n} u_N^{2r}$.

As in Lemma 4, the term $e^{-H(c',u,v)}$ *in the probability above can be omitted for case (A1)(ii).*

Remark 2 *The bound for the empirically process indexed by the Bregman divergence (Lemma 3) is expected to be smaller than that indexed by the risk difference (Lemma 4), as we indicate in the introduction. Intuitively, the reason we expect the bound for the empirically process indexed by the Bregman divergence to be smaller can be explained as follows. Assuming* $\ell(y, z)$ *is twice differentiable in* z*, Taylor's expansion implies* $\ell(y, f(x)) - \ell(y, f_0(x))$ *is proportional to* $f(x) - f_0(x)$ *and* $D_{f,f_0}(x, y)$ *is proportional to* $(f(x) - f_0(x))^2$*. We are mainly concerned with* f *in a shrinking neighborhood of* f_0*, and thus* $D_{f,f_0}(x, y)$ *is expected to be smaller than* $\ell(y, f(x)) - \ell(y, f_0(x))$*, making the associated empirical process smaller. Although for non-smooth losses such as the quantile loss we cannot directly use Taylor's expansion, Lemma 1 provides more general bounds that still makes the emprical process associated with the Bregman divergence smaller.*

Remark 3 *It is not easy to give general sufficient conditions when* $vR_n(u/v)$ *will dominate other terms. We expect* $vR_n(u/v)$ *is the key term since in kernel ridge regression, it has been shown in various proofs that this term is the main stochastic error term. Furthermore, the terms* $\frac{c'}{\sqrt{N}}\sqrt{H(c', u, v)} + \frac{c}{N}H(c', u, v)$ *come from the bound for empirical process indexed by the Bregman divergence. As we explained in the previous remark, it is expected (but not always) that this term would be smaller. The last term* $u^{1-s}v^s \frac{\mathcal{R}_N^2(u/v)}{(u/v)^2}\log N$ *comes from application of the concentration inequality. Since in many studies such as the current one, the concentration phenomenon (that the random quantity is tightly concentrated around its expectation) holds, we also expect this term is a high-order term.*

When u, v, u_N *are specifically chosen to balance some terms that represent bias and variance, we can verify that* $vR_n(u/v)$ *is the dominating term, or provide simple conditions that it is, as we did in Corollaries below (see the next remark), which shows that it is indeed the dominating term under reasonable assumptions.*

Remark 4 *When $s_j \asymp j^{-\alpha}$, we have $\mathcal{R}_N(u) = u^{1-1/\alpha}/\sqrt{N}$. Assume $1/\alpha = s$ (as mentioned in the discussions below the list of assumptions, in many cases we indeed have $s = 1/\alpha$). Then it can be calculated $u_N = N^{-\frac{1}{2(1+2r+s)}}$.*

In this specific setting, for the non-distributed setting (central estimator), with $u \asymp u_N^{1+2r}$ and $v \asymp u_N^{2r}$, the first term $v\mathcal{R}_N(u/v)$ in the bounds (7) and (8) always dominates. Here, the first term dominates means other terms are of smaller order or at most of the same order as $v\mathcal{R}_N(u/v)$. For the distributed setting, when $u \asymp \sqrt{N/n}u_N^{1+2r}$ and $v \asymp \sqrt{N/n}u_N^{2r}$, in (A2)(ii) the first term always dominates. For case (A1)(i) with bound (7), in the distributed setting the first term in the bound dominate if $m \lesssim N^{\frac{1+2r-s}{1+2r+s}}$. This and some other calculations immediately lead to the following corollary. The rate obtained in the corollary is the same as the least-squares case (Caponnetto and De Vito, 2007; Lin et al., 2017).

Corollary 1 *Assume $s_j \asymp j^{-\alpha}$ and (A1)-(A6) holds with $s = 1/\alpha$. For the central estimator, setting $\lambda \asymp u_N^2$ with $u_N = N^{-\frac{1}{2(1+2r+s)}}$, we have for N large enough,*

$$\|\widehat{f} - f_0\| + u_N\|\widehat{f} - f_0\|_{\mathcal{H}} \leq Cu_N^{1+2r},$$

with probability at least $1 - e^{-u_N^{-2s}}$.

For the distributed setting with local sample size n, setting $\lambda \asymp u_N^2$ with $u_N = N^{-\frac{1}{2(1+2r+s)}}$, we have for $m \leq cN^{\frac{1+2r-s}{1+2r+s}}$ with c sufficiently small,

$$\|\widehat{f}_j - f_0\| + u_N\|\widehat{f}_j - f_0\|_{\mathcal{H}} \leq C\sqrt{\frac{N}{n}}u_N^{1+2r},$$

with probability at least $1 - e^{-u_N^{-2s}}$.

We can now prove our main result for the distributed estimator. Furthermore, in Corollary 2, it is shown that under suitable conditions, the learning rate for the distributed estimator matches that of the central estimator.

Theorem 2 *Under assumptions (A1)-(A6), setting $\lambda \asymp u_N^2$, and assume that the dominating term in Lemma 4 is the first term, we have with probability $1 - e^{-N\mathcal{R}_N^2(u/v)/(u/v)^2} - e^{-H(c',u,v)}$ where $u \asymp \sqrt{N/n}u_N^{1+2r}, v \asymp \sqrt{N/n}u_N^{2r}$,*

$$\begin{aligned} & \|\bar{f} - f_0\|^2 + u_N^2\|\bar{f} - f_0\|_{\mathcal{H}}^2 \\ \leq\ & C\left(u_N^{2+4r} + u^{3-s}v^s + \frac{c'}{\sqrt{n}}\sqrt{H(c',u,v)} + \frac{c}{n}H(c',u,v)\right), \end{aligned}$$

where $c' = Cu^{a'+(1-s)b'}v^{sb'}$, $c = Cu^{a+(1-s)b}v^{sb}$ and the values of a, b, a', b' are as stated in Lemma 1 ($a = 0, b = 2, a' = b' = 1$ for case (A1)(i), and $a = 0, b = 1, a' = 1, b' = 1/2$ for case (A1)(ii)).

Remark 5 *When $s_j \asymp j^{-\alpha}$ and $s = 1/\alpha$, we can work out the conditions under which the first term u_N^{2+4r} in the rate dominates. This specific setting is stated as a corollary.*

Corollary 2 *Assume $s_j \asymp Cj^{-\alpha}$ and (A1)-(A6) hold with $s = 1/\alpha$. Set $\lambda \asymp u_N^2$ with $u_N = N^{-\frac{1}{2(1+2r+s)}}$. For (A1) (i), if $m^3 \lesssim N^{\frac{1+2r-s}{1+2r+s}}$, with probability $1-\exp\{-u_N^{-2s}\}$, or for (A2)(ii), if $m^{5-s} \lesssim N^{\frac{s^2-(2+2r)s+1+2r}{1+2r+s}}$, with probability $1-\exp\{-u_N^{-2s}\}-\exp\{-u_N^{-((3+2r)s-s^2)}m^{-s/2}\}$, we have*

$$\|\bar{f}-f_0\|^2 + u_N^2\|\bar{f}-f_0\|_{\mathcal{H}}^2 \quad \leq \quad Cu_N^{2+4r}.$$

Our rate $u_N^{2+4r} = N^{-\frac{1+2r}{1+2r+s}}$ matches the optimal rate for the least squares case as in Caponnetto and De Vito (2007); Guo et al. (2017). Note that Zhang et al. (2015) does not use the source condition (equivalent to saying $r = 0$ in our Assumption (A2)), and our rate in this special case $r = 0$ is also the same as the rate in their Corollary 4 (our parameter $\alpha = 1/s$ is equivalent to their parameter 2ν). In our proof, the extension to $r > 0$ for the general loss makes use of Young's inequality for operators. The lower bound for this minimax rate is derived by considering the special case of mean regression with Gaussian noise. Since quantile regression and Huber's regression model can also be applied to this generating model, the rate for quantile regression and Huber's regression is also minimax. The case of logistic regression requires re-examining the proof of the existing lower bounds, and we show in the Appendix that this rate is also minimax optimal for kernel logistic regression. Since all rates are the same, we *conjecture* these are all minimax rates under appropriate assumptions for other models. One might ask what the optimal number of machines/partitions m is. However, m should not be regarded as a tuning parameter. If computation is feasible, one should always use $m = 1$. The reason we want to choose $m > 1$ is simply because computation with $m = 1$ is infeasible (or too slow). This is a trade-off between statistical accuracy and computational speed. In the distributed setting, often m is fixed and pre-given and not something we can choose. The theory developed merely says that if m is not too large, we can still achieve the same rate as when $m = 1$.

Remark 6 *(On kernels with non-polynomial decaying eigenvalues) Zhang et al. (2015) consider kernels with finite rank and kernels with exponentially decaying eigenvalues (parametric rate or parameter rate with an additional logarithmic term can be achieved in these cases). In our case, we cannot directly establish general results for different types of eigenvalue decay rates. This is because for general losses, our assumption is not directly based on eigenvalue decay. Instead, we use the entropy condition and the sup-norm assumption, which do not have a one-one correspondence with the eigenvalue decay assumption. However, we can still say something about the case of finite-rank kernel, and the case about Gaussian kernel (the only well-known example with exponential eigenvalue decay).*

For a finite-rank kernel of dimension d, with d fixed, Lemma 2.6.15 and Theorem 2.6.7 of van der Vaart and Wellner (1996) show that (A4) is satisfied with the entropy bound $(2/\varepsilon)^{2/\alpha}$ replaced by $C\log(1/\varepsilon)$. Furthermore, (A5) holds with $s = 0$ since all norms are equivalent on a finite-dimensional space. Following the same proof strategy, we can derive that with probability approaching one,

$$\|\bar{f}-f_0\|^2 + u_N^2\|\bar{f}-f_0\|_{\mathcal{H}}^2 \quad \leq \quad \frac{C}{N},$$

if $m \lesssim N^{1/3}$ for case (A1)(i), or if $m \lesssim \frac{N^{1/5}}{(\log N)^{2/5}}$ for case (A1)(ii).

For kernels whose eigenvalues decay exponentially, that is $s_j \lesssim e^{-Cj^p}$ *for some* $p > 0$, *Section C.4. (Proof of Corollary 4.3) of Zhao et al. (2016) shows that* $\log\mathcal{N}(\varepsilon, \mathcal{H}(1), L_\infty) \lesssim (\log(\frac{1}{\varepsilon}))^{\frac{p+1}{p}}$. *Using this in place of Assumption (A4), and with (A5) holds trivially for* $s = 1$, *we can then calculate that* $\mathcal{R}_N(u_N) = u_N^{4r+2}$ *leads to* $u_N \asymp \left(\frac{(\log N)^{1/p}}{N}\right)^{\frac{1}{2(2r+1)}}$. *Then, using exactly the same proof steps, we can show that Corollary 1 holds with this* u_N. *That is, the convergence rate is* $u_N^{1+2r} \asymp \left(\frac{(\log N)^{1/p}}{N}\right)^{\frac{1}{2}}$, *an almost parametric rate. Furthermore, Corollary 2 holds for (A1)(i) if* $m^3 \lesssim \frac{N^{\frac{2r}{2r+1}}}{(\log N)^{1+\frac{1}{p}}}$, *and for (A1)(ii) if* $m^4 \lesssim \frac{N^{\frac{r}{2r+1}}}{(\log N)^{1+\frac{1}{p}}}$.

4. Numerical results

We illustrate the distributed learning for Huber's loss and the logistic loss. The results for the quantile loss were presented in Lian (2022) and thus we do not repeat them here. For nonparametric mean regression with Huber's loss, we generate the data from the model

$$y_i = f(x_i) + \epsilon_i,$$

where $f(x) = 2\sin(2\pi x)$, $x_i \overset{i.i.d}{\sim} Unif(0,1)$, $\epsilon_i \overset{i.i.d}{\sim} t_2$ (Students' t error with two degrees of freedom). We use the Sobolev kernel of order 2 given by $K(x_1, x_2) = \min\{x_1, x_2\}^2 \max\{x_1, x_2\}/2 - \min\{x_1, x_2\}^3/6 + x_1 x_2 + 1$. For the tuning parameter λ, it is determined by the prediction error on an independently generated data set for the averaged estimator $\bar{f}$. For simplicity we do not tune the robustness parameter and simply set it to be 2. Our main purpose to demonstrate the differences/similarities between the distributed estimator and the central estimator, rather than the statistical properties of Huber's estimator. The robustness parameter can also be chosen using cross-validation or hold-out data, like the choice of λ, which adds to the complexity of the algorithm. The estimation errors $\|\bar{f} - f_0\|$ for different pairs (n, m), averaged over 100 repetitions, are reported in Table 1 (top). We set $n \in \{32, 64, 128, 356, 512, 1024\}$ and $m \in \{1, 2, 4, 8, 16, 32\}$. For comparison, we also computed the nonparametric least squares estimator to demonstrate the robustness of the Huber loss (Table 1 bottom). We see that, unsurprisingly, for a fixed number of machines m, increasing n decreases the error, and for a fixed local sample size n, increasing m also decreases the error. More interestingly, for a fixed value of $N = nm$, increasing m (decreasing n at the same time) will lead to larger errors. In other words, more partitions imply less accuracy (the centralized estimator using all data is the best). For example, the cells in the table colored blue all represent the same total sample size $N = 1024$, and the error increases with m. The same phenomenon is observed for the cells colored yellow. This result is also as expected. One-shot distributed learning is computationally faster at the cost of increased error. Although the distributed estimator is worse than the centralized estimator, it is better than using local data only (data on one machine). Due to the heavy-tailed nature of the error term, we see that the least squares loss performs worse than the Huber's loss.

In the next experiment, we consider the logistic model with data generated from

$$y_i \sim Bernoulli(p_i), \log\frac{p_i}{1-p_i} = f(x_i),$$

Table 1: The estimation errors for different pairs of (n, m) with Huber's loss (top). The total sample size is $N = nm$. For comparison the results based on the least squares loss are also presented (bottom)

n \ m	1	2	4	8	16	32
32	4.220	2.058	1.099	0.917	0.527	0.393
64	1.523	0.709	0.449	0.260	0.135	0.095
128	0.623	0.434	0.237	0.122	0.085	0.057
256	0.389	0.216	0.119	0.078	0.044	0.036
512	0.140	0.075	0.044	0.023	0.013	0.007
1024	0.067	0.037	0.020	0.012	0.007	0.003
n \ m	1	2	4	8	16	32
32	6.996	3.305	2.279	1.806	0.947	0.733
64	2.629	1.509	0.820	0.514	0.392	0.165
128	1.477	0.805	0.495	0.308	0.149	0.088
256	0.796	0.472	0.256	0.133	0.081	0.052
512	0.319	0.173	0.106	0.053	0.029	0.019
1024	0.173	0.081	0.052	0.025	0.016	0.008

Table 2: The estimation errors for different pairs of (n, m) for the nonparametric logistic model. The total sample size is $N = nm$.

n \ m	1	2	4	8	16	32
32	13.016	9.280	8.673	7.638	7.112	6.827
64	9.205	8.128	5.753	4.056	3.876	3.356
128	4.479	3.492	1.666	1.053	0.695	0.526
256	2.571	1.525	0.943	0.458	0.387	0.180
512	1.452	0.888	0.461	0.302	0.168	0.111
1024	0.678	0.346	0.235	0.142	0.088	0.061

where the function f and the distribution of x_i is the same as before. We still use the Sobolev kernel of order 2 and select λ based on independently generated data. The estimation errors $\|\bar{f} - f_0\|$ are shown in Table 2. The general pattern for different (n, m) is the same as the regression case, for example larger m with the same $N = nm$ leads to larger errors.

Next we illustrate the distributed quantile regression. For a value $\tau \in (0, 1)$, the sample is generated from the model $y_i = f(x_i) + (1 + x_i)\sigma(\epsilon_i - \Phi^{-1}(\tau))$, where x_i are generated uniformly on $[0, 1]$, $\epsilon_i \sim N(0, 1)$, Φ^{-1} is the quantile function of the standard normal distribution so that $f(x_i)$ is indeed the conditional τ-quantile of y_i. We set $f(x) = \sin(2\pi x)$ and $\sigma = 0.5$. The simulation results are shown in Table 3 for $\tau = 0.5$.

Table 3: The estimation errors for different pairs of (n, m) when $\tau = 0.5$. The total sample size is $N = nm$.

n \ m	1	2	4	8	16	32
32	2.765	1.211	0.898	0.466	0.316	0.221
64	1.226	0.653	0.371	0.221	0.108	0.062
128	0.616	0.386	0.229	0.099	0.053	0.023
256	0.346	0.199	0.105	0.058	0.029	0.013
512	0.186	0.086	0.046	0.020	0.011	0.008
1024	0.089	0.047	0.020	0.010	0.008	0.004

Table 4: Computation time (in seconds) for different sample size n with different losses.

loss	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$
Huber	19.72	47.65	93.70	177.00	314.09	645.49
logistic	16.04	28.81	83.83	182.78	220.82	460.16
quantile	40.67	70.77	137.94	342.66	480.19	1166.63

Based on the representer theorem, solving the penalized problem with sample size n is an optimization problem in the n-dimensional Euclidean space. The computational complexity depends on the specific algorithm used. For Huber's loss and quantile loss, we used the package hqreg in R and for logistic loss we used the package glmnet. It is hard to analyze analytically the complexity since the algorithm is iterative and there seems no known theoretical results regarding the number of iterations required. As an indication of the savings in computation, we note that if the standard gradient descent is used, the complexity is $O(n^2)$ per iteration (quadratic in sample size). We report the real computation time for different losses in Table 4 below on our desktop computer with 12th Gen Intel(R) Core(TM) i5-12400F and 16GB of RAM. The table tells us the savings in computation time in a real distributed setting where different partitions are processed *parallelly*. For example, with Huber's loss, $N = 1024$ requires about 645 seconds, while if it is distributed as $(n = 128, m = 8)$, it only requires about 93 seconds.

Finally, we examine two real-world datasets using kernel logistic regression and quantile regression, respectively. The first dataset, HTRU2 (`https://archive.ics.uci.edu/dataset/372/htru2`), comprises 17,898 pulsar candidates from the High Time Resolution Universe Survey (South). Out of these, 1,639 are confirmed pulsars, and 16,259 are false positives, all annotated by humans. For computational feasibility and to make the two classes more balanced, we sample 2,457 obsevations from the second class and thus the total sample size is $N = 4,096$. Each candidate is characterized by 8 continuous features. We standardize these features and utilize a Gaussian kernel to train a logistic regression model. As the true function is unknown, we use the standard estimator without approximation as a benchmark to calculate the Relative errors for $n = \{32, 64, \ldots, 1024\}$, where the relative

Table 5: Relative errors for two real data sets both with $N = 4,096$, for logistic regression and quantile regression, respectively.

loss	$n = 32$	$n = 64$	$n = 128$	$n = 256$	$n = 512$	$n = 1024$
logistic	0.833	0.577	0.237	0.166	0.158	0.145
quant ($\tau = 0.5$)	0.562	0.356	0.240	0.113	0.066	0.047
quant ($\tau = 0.9$)	0.642	0.476	0.321	0.200	0.163	0.150

error is defined as $\|\widehat{f} - f_0\|/\|f_0\|$ and f_0 is the estimator using the entire sample with size N. The procedure is repeated 100 times and the average errors are reported in Table 5.

The second dataset is on air quality in an Italian city and is obtained from the UCI Machine Learning Repository (`https://archive.ics.uci.edu/dataset/360/air+quality`). This dataset includes 9,357 hourly averaged measurements from five metal oxide chemical sensors in an Air Quality Chemical Multisensor Device, collected from March 2004 to February 2005. It features fifteen variables, such as sensor responses for carbon monoxide (PT08.S1), non-methanic hydrocarbons (PT08.S2), total nitrogen oxides (PT08.S3), nitrogen dioxide (PT08.S4), and ozone (PT08.S5), along with temperature (T) and absolute humidity (AH) at various quantile levels. The objective is to model the true hourly averaged benzene concentration in $\mu g/m^3$ using these features. For this purpose, we fit quantile models at $\tau = 0.5$ and $\tau = 0.9$ using a Gaussian kernel on $N = 4,096$ observations. The relative errors for different (n, m) (again, treating the estimate based on sample size N as the truth) are also presented in Table 5.

5. Conclusions

In this paper, we considered one-shot distributed learning based on the simple divide-and-average strategy. The main contribution is to extend the existing results for least squares loss to a much larger class of convex losses. Such an extension beyond least squares is interesting and challenging. Due to the unavailability of the closed-form solution, the proof is rather different and more technical than the least squares case. We show that distributed learning can achieve the same convergence rate as centralized learning, which in turn is the same for all losses we consider (in particular all rates can be the same as the least squares case). An open question associated with our results is whether the required bound on m is tight. This appears to be a very challenging question, even for mean regression using the least squares loss.

The one-shot learning strategy has a very low communication cost and thus one may hope to improve accuracy using more communications. For this, one could adopt other learning strategies such as Jordan et al. (2018), which shows that by repeatedly sending gradients to the server to solve a surrogate optimization problem, using multiple rounds of communications can possibly improve accuracy (theoretically, the constraint in one-shot learning that the number of machines m cannot be too large can be relaxed or even removed). It remains to develop a nonparametric version in the framework of RKHS for the approach of Jordan et al. (2018), since the existing research is only on parametric models.

Furthermore, it would be a nontrivial problem to consider distributed learning over a general network using, for example, a gradient method. In the parametric (finite-dimensional) setting, learning over a general network has attracted much attention in the optimization literature (Nedić and Ozdaglar, 2009; Nedić et al., 2017). It is interesting to see how this works for the nonparametric (infinite-dimensional) setting.

Appendix A: Proofs for results in the main paper

Proof of Proposition 1. For logistic loss $h(yf(x)) = \log(1 + e^{-yf(x)})$, by Taylor's expansion, we have

$$\begin{aligned} & h(yf(x)) - h(yf_0(x)) \\ = & -\frac{ye^{-yf_0(x)}}{1+e^{-yf_0(x)}}(f(x) - f_0(x)) \\ & +\frac{1}{2}\frac{e^{-yf_0(x)}}{(1+e^{-yf_0(x)})^2}(f(x) - f_0(x))^2 \\ & -\frac{1}{6}\int_0^1 \frac{ye^u(1-e^u)}{(1+e^u)^4}du \cdot (f(x) - f_0(x))^3, \end{aligned}$$

where $u = -yf_0(x) - ty(f(x) - f_0(x))$. The first term on the right-hand side above has expectation zero since f_0 minimizes the expected loss. By defining $\mathcal{L}_{\mathcal{H}} = E[\frac{e^{-yf_0(x)}}{(1+e^{-yf_0(x)})^2}K_x \otimes K_x]$, where $K_x \otimes K_x$ is the operator that maps f to $f(x)K(x,.) \in \mathcal{H}$, the expectation of the second term can be written as $\frac{1}{2}\langle f - f_0, \mathcal{L}_{\mathcal{H}}(f - f_0)\rangle_{\mathcal{H}}$, Since $\frac{e^{-yf_0(x)}}{(1+e^{-yf_0(x)})^2}$ is bounded away from zero and infinity (f_0 is bounded by assumption), we indeed have $\mathcal{L}_{\mathcal{H}} \asymp \mathcal{L}$. The expectation of the third term is easily seen to be $O(E[(f(x) - f_0(x))^3])$.

For the quantile loss, using Knight's identity that $\rho_\tau(x - y) - \rho_\tau(x) = -y(\tau - I\{x \le 0\}) + \int_0^y I\{x \le t\} - I\{x \le 0\}dt$, we have

$$\begin{aligned} & \rho_\tau(y - f(x)) - \rho_\tau(y - f_0(x)) \\ = & -(f(x) - f_0(x))(\tau - I\{y \le f_0(x)\}) + \int_0^{f(x)-f_0(x)} I\{y - f_0(x) \le t\} - I\{y - f_0(x) \le 0\}dt. \end{aligned}$$

The expectation of the first term on the right-hand side is zero due to the first-order optimality condition for f_0. For the expectation of the second term, we have

$$\begin{aligned} & E\left[\int_0^{f(x)-f_0(x)} I\{y - f_0(x) \le t\} - I\{y - f_0(x) \le 0\}dt\right] \\ = & E\left[\int_0^{f(x)-f_0(x)} (q(0|x)t + q'(t^*|x)t^2)dt\right] \\ = & \frac{1}{2}E[q(0|x)(f(x) - f_0(x))^2] + O(E[(f(x) - f_0(x))^3]). \end{aligned}$$

By defining $\mathcal{L}_{\mathcal{H}} = E[q(0|x)K_x \otimes K_x]$, we have $E[q(0|x)(f(x) - f_0(x))^2] = \langle f - f_0, \mathcal{L}_{\mathcal{H}}(f - f_0)\rangle_{\mathcal{H}}$ and $\mathcal{L}_{\mathcal{H}} \asymp \mathcal{L}$ since $q(0|x)$ is bounded away from zero and infinity. □

Proof of Lemma 1. First, we show that (A6) together with the first part of (5) (bound on the supremum norm) implies the second part of (5) (bound on the L^2 norm) with $a' = a/2+1$, $b' = b/2$. Indeed, (A6) implies $E[D_{f,f_0}(x,y)] = O(\|f-f_0\|^2)$ when $\|f-f_0\|_\infty \le 1$. Thus $E[D^2_{f,f_0}(x,y)] \le \|D_{f,f_0}(x,y)\|_\infty \cdot E[D_{f,f_0}(x,y)] \le C\|f-f_0\|^{a+2}\|f-f_0\|^b_\infty$. Thus we can take $a' = a/2+1$, $b' = b/2$.

If the loss function is smooth in the sense that $|\partial_2\ell(y,z) - \partial_2\ell(y,z')| \le C|z-z'|$, we have

$$\begin{aligned} & D_{f,f_0}(x,y) = \ell(y,f(x)) - \ell(y,f_0(x)) - \partial_2\ell(y,f_0(x))(f(x)-f_0(x)) \\ = & \left(\int_0^1 \partial_2\ell\left(y, f_0(x) + t(f(x)-f_0(x))\right) dt - \partial_2\ell(y,f_0(x))\right)(f(x)-f_0(x)) \\ \le & C(f(x)-f_0(x))^2, \end{aligned}$$

and thus the first part of (5) holds with $a = 0, b = 2$, which then means we can set $a' = 1, b' = 1$.

For a loss function that is Lipschitz continuous in the sense that $|\ell(y,z) - \ell(y,z')| \le C|z-z'|$, we easily see that the first part of (5) holds with $a = 0, b = 1$. By the arguments at the beginning of the current proof, we can set $a' = 1, b' = 1/2$. □

Proof of Lemma 2. First we note that for $f \in \mathcal{D}(u,v)$, $\|f-f_0\|_\infty \le Cw$. In case of (A1)(i), we have for $f_1, f_2 \in \mathcal{D}(u,v)$,

$$\begin{aligned} & |D_{f_1,f_0}(x,y) - D_{f_2,f_0}(x,y)| \\ = & |\ell(y,f_1(x)) - \ell(y,f_2(x)) - \partial_2\ell(y,f_0(x))(f_1(x)-f_2(x))| \\ = & |\partial_2\ell(y,z)(f_1(x)-f_2(x)) - \partial_2\ell(y,f_0(x))(f_1(x)-f_2(x))| \\ \le & C|z-f_0(x)| \cdot |f_1(x)-f_2(x)|, \end{aligned}$$

where z lies between $f_1(x)$ and $f_2(x)$. Since $|z-f_0(x)| \le \max\{|f_1(x)-f_0(x)|, |f_2(x)-f_0(x)|\} \le Cw$. we get

$$|D_{f_1,f_0}(x,y) - D_{f_2,f_0}(x,y)| \le Cw|f_1(x)-f_2(x)|.$$

The above means

$$\begin{aligned} \mathcal{N}(\epsilon \cdot Cvw, \mathcal{D}(u,v), L^2(P_n)) & \le \mathcal{N}(\epsilon v, \{\|f-f_0\|_{\mathcal{H}} \le v\}, L^2(P_n)) \\ & = \mathcal{N}(\epsilon, \mathcal{H}(1), L^2(P_n)) \le \exp\{(C/\epsilon)^{2/\alpha}\}, \end{aligned}$$

which implies

$$\mathcal{N}(\epsilon, \mathcal{D}(u,v), L^2(P_n)) \le \exp\{(Cvw/\epsilon)^{2/\alpha}\}.$$

For the case of (A1)(ii), we have

$$\begin{aligned} & |D_{f_1,f_0}(x,y) - D_{f_2,f_0}(x,y)| \\ = & |\ell(y,f_1(x)) - \ell(y,f_2(x)) - \partial_2\ell(y,f_0(x))(f_1(x)-f_2(x))| \\ \le & C|f_1(x)-f_2(x)|. \end{aligned}$$

Thus

$$\mathcal{N}(C\epsilon v, \mathcal{D}(u,v), L^2(P_n)) \le \mathcal{N}(\epsilon v, \{\|f-f_0\|_{\mathcal{H}} \le v\}, L^2(P_n)) \le \exp\{(C/\epsilon)^{2/\alpha}\},$$

leading to

$$\mathcal{N}(\epsilon, \mathcal{D}(u,v), L^2(P_n)) \leq \exp\{(Cv/\epsilon)^{2/\alpha}\}.$$

□

Proof of Lemma 3. Using Lemma 1 and the sup-norm assumption (A5), we have $\|D_{f,f_0}\|_\infty \leq c := Cu^{a+(1-s)b}v^{sb}$, $\|D_{f,f_0}\| \leq c' := Cu^{a'+(1-s)b'}v^{sb'}$. Thus, the entropy bound in Lemma 2, Theorem 3.12 of Koltchinskii (2011) implies

$$E[\sup_{g\in\mathcal{D}(u,v)}(P_n - P)g] \leq C\left(\frac{c'}{\sqrt{n}}\sqrt{H(c',u,v)} + \frac{c}{n}H(c',u,v)\right).$$

Using Talagrand's concentration inequality, we have with probability $1-e^{-t}$,

$$\begin{aligned}\sup_{g\in\mathcal{G}}(P_n - P)g &\leq CE[\sup_{g\in\mathcal{G}}(P_n - P)g] + C\sqrt{\frac{t}{n}}c' + C\frac{tc}{n} \\ &\leq C\left(\frac{c'}{\sqrt{n}}\sqrt{H(c',u,v)} + \frac{c}{n}H(c',u,v) + \sqrt{\frac{t}{n}}c' + \frac{tc}{n}\right),\end{aligned}$$

and setting $t = H(c',u,v)$ proves the lemma. □

Proof of Lemma 4. We have $\ell(y,f(x)) - \ell(y,f_0(x)) = D_{f,f_0}(x,y) + \partial_2\ell(y,f_0(x))(f(x) - f_0(x))$ and thus

$$(P_N - P)(\ell(y,f(x)) - \ell(y,f_0(x))) = (P_N - P)D_{f,f_0}(x,y) + P_N\partial_2\ell(y,f_0(x))(f(x) - f_0(x)).$$

The first term on the right-hand side above is bounded by Lemma 3. For the second term, we write

$$P_N\partial_2\ell(y,f_0(x))(f(x) - f_0(x)) = \frac{1}{N}\sum_i \epsilon_i(f(x_i) - f_0(x_i)),$$

where $\epsilon_i = \partial_2\ell(y_i, f_0(x_i))$. We can show

$$E\left[\sup_{\|f\|\leq u, \|f\|_{\mathcal{H}}\leq v}\frac{1}{N}\sum_i \epsilon_i f(x_i)\right] \leq Cv\mathcal{R}(\frac{u}{v}).$$

In fact, we can bound the left side of the above by $\left(E\left[\sup_{\|f\|\leq u,\|f\|_{\mathcal{H}}\leq v}\left|\frac{1}{N}\sum_i \epsilon_i f(x_i)\right|^2\right]\right)^{1/2}$, and then use exactly the same arguments as in the proof of Theorem 42 in Mendelson (2002).

Note that since $\|f\|_\infty \lesssim u^{1-s}v^s$, an envelope function for the class $\{\epsilon f(x) : \|f\| \leq u, \|f\|_{\mathcal{H}} \leq v\}$ is $F(\epsilon) = C|\epsilon|u^{1-s}v^s$ and by that ϵ is sub-exponential, $\|\max_{1\leq i\leq N} F(\epsilon_i)\|_{\psi_1} \lesssim (\log N)u^{1-s}v^s$. Then, using this envelope function, by Adamczak bound (Koltchinskii, 2011), which is a concentration-type inequality for unbounded variables, we have with probability $1-e^{-t}$,

$$\begin{aligned}\sup_{\|f\|\leq u,\|f\|_{\mathcal{H}}\leq v}\frac{1}{N}\sum_i \epsilon_i f(x_i) &\leq CE\left[\sup_{\|f\|\leq u,\|f\|_{\mathcal{H}}\leq v}\frac{1}{N}\sum_i \epsilon_i f(x_i)\right] \\ &\quad + C\left(u\sqrt{\frac{t}{n}} + (\log N)u^{1-s}v^s\frac{t}{n}\right). \qquad (9)\end{aligned}$$

Setting $t = N\mathcal{R}_N^2(u/v)/(u/v)^2$, we get with probability $1 - e^{-N\mathcal{R}_N^2(u/v)/(u/v)^2}$,

$$\sup_{\|f\|\le u, \|f\|_{\mathcal{H}}\le v} \frac{1}{N}\sum_i \epsilon_i f(x_i) \quad \le \quad Cv\mathcal{R}_N(u/v) + Cu^{1-s}v^s \frac{\mathcal{R}_N^2(u/v)}{(u/v)^2}\log N.$$

This proves (7).

In the case of (A1)(ii), we can directly use

$$\begin{aligned} & E\left[\sup_{\|f-f_0\|\le u, \|f-f_0\|_{\mathcal{H}}\le v} (P_N - P)\{\ell(y, f(x)) - \ell(y, f_0(x))\}\right] \\ \le \quad & 2E\left[\sup_{\|f-f_0\|\le u, \|f-f_0\|_{\mathcal{H}}\le v} \frac{1}{N}\sum_i \sigma_i\{\ell(y_i, f(x_i)) - \ell(y_i, f_0(x_i))\}\right] \\ \le \quad & CE\left[\sup_{\|f-f_0\|\le u, \|f-f_0\|_{\mathcal{H}}\le v} \frac{1}{N}\sum_i \sigma_i(f(x_i) - f_0(x_i))\right] \\ \le \quad & Cv\mathcal{R}_N(u/v), \end{aligned}$$

where the first step used symmetrization and the second step used the contraction inequality for the Rademacher process (Theorem 2.3 in Koltchinskii (2011)), and the concentration inequality gives with probability $1 - e^{-t}$,

$$\begin{aligned} \sup_{\|f\|\le u, \|f\|_{\mathcal{H}}\le v} \frac{1}{N}\sum_i \sigma_i f(x_i) \quad \le \quad & CE\left[\sup_{\|f\|\le u, \|f\|_{\mathcal{H}}\le v} \frac{1}{N}\sum_i \sigma_i f(x_i)\right] \\ & + C\left(u\sqrt{\frac{t}{n}} + u^{1-s}v^s\frac{t}{n}\right) \end{aligned}$$

(note unlike (9), the $log(N)$ term is not necessary here due to the Rademacher variables σ_i are bounded). □

Proof of Theorem 1. Suppose $\|\widehat{f} - f_0\| + u_N\|\widehat{f} - f_0\|_{\mathcal{H}} \ge \gamma := Lu_N^{1+2r}$ for a sufficiently large constant L. This means

$$\inf_{\|f-f_0\|+u_N\|f-f_0\|_{\mathcal{H}}\ge\gamma} \frac{1}{N}\sum_i \ell(y_i, f(x_i)) + \lambda\|f\|_{\mathcal{H}}^2 - \frac{1}{N}\sum_i \ell(y_i, f_0(x_i)) - \lambda\|f_0\|_{\mathcal{H}}^2 < 0. \quad (10)$$

By convexity of the functional in f in the above, we have

$$\inf_{\|f-f_0\|+u_N\|f-f_0\|_{\mathcal{H}}=\gamma} \frac{1}{N}\sum_i \ell(y_i, f(x_i)) + \lambda\|f\|_{\mathcal{H}}^2 - \frac{1}{N}\sum_i \ell(y_i, f_0(x_i)) - \lambda\|f_0\|_{\mathcal{H}}^2 < 0. \quad (11)$$

In fact, to see that (10) implies (11), we define the left-hand side for which we take the infimum of by $G(f)$, considered as a function of f, which is convex in f. We trivially have $G(f_0) = 0$. (10) means there exists some f with $\|f - f_0\| + u_N\|f - f_0\|_{\mathcal{H}} \ge \gamma$ and $G(f) < 0$. Let $f' = (1-t)f + tf_0$ for some $t \in [0, 1)$ chosen such that $\|f' - f_0\| + u_N\|f' - f_0\|_{\mathcal{H}} = \gamma$. Then by convexity we have $G(f') = G((1-t)f + tf_0) \le (1-t)G(f) < 0$. This implies (11).

Note that when $\|f-f_0\|+u_N\|f-f_0\|_{\mathcal{H}}=\gamma$, $\|f-f_0\|_\infty \le C_1\gamma u_N^{-s} = C_1 L u_N^{1+2r-s} \le C_5$, and thus by Lemma 4 and Lemma 5,

$$\begin{aligned}
& C\|f-f_0\|^2 \\
\le\ & E[\ell(y,f(x))] - E[\ell(y,f_0(x))] \\
\le\ & \lambda\|f_0\|_{\mathcal{H}}^2 - \lambda\|f\|_{\mathcal{H}}^2 + Cu_N^{-1}\mathcal{R}_N(u_N)\gamma \\
=\ & -2\lambda\langle f_0, f-f_0\rangle_{\mathcal{H}} - \lambda\|f-f_0\|_{\mathcal{H}}^2 + Cu_N^{-1}\mathcal{R}_N(u_N)\gamma \\
\le\ & \sqrt{2r}\lambda^{\frac{1+2r}{2}}\|f-f_0\| + \lambda^{1+r}\|f-f_0\|_{\mathcal{H}} - \lambda\|f-f_0\|_{\mathcal{H}}^2 \\
& + Cu_N^{-1}\mathcal{R}_N(u_N)\gamma \\
\le\ & \sqrt{2r}\lambda^{\frac{1+2r}{2}}\|f-f_0\| + \frac{1}{2}\lambda^{1+2r} + \frac{\lambda}{2}\|f-f_0\|_{\mathcal{H}}^2 - \lambda\|f-f_0\|_{\mathcal{H}}^2 \\
& + Cu_N^{-1}\mathcal{R}_N(u_N)\gamma,
\end{aligned} \tag{12}$$

where in the 3rd inequality above, we used that

$$\begin{aligned}
& |\lambda\langle f_0, f-f_0\rangle_{\mathcal{H}}| \\
=\ & \lambda|\langle \mathcal{L}^r g_0, f-f_0\rangle_{\mathcal{H}}| \\
=\ & \lambda|\langle g_0, \mathcal{L}^r(f-f_0)\rangle_{\mathcal{H}}| \\
\le\ & \lambda^{\frac{1}{2}+r}\|\lambda^{\frac{1}{2}-r}\mathcal{L}^r(f-f_0)\|_{\mathcal{H}} \\
\le\ & \lambda^{\frac{1}{2}+r}\sqrt{\langle f-f_0, ((1-2r)\lambda+2r\mathcal{L})(f-f_0)\rangle_{\mathcal{H}}} \\
\le\ & \sqrt{2r}\lambda^{\frac{1}{2}+r}\|f-f_0\| + \lambda^{1+r}\|f-f_0\|_{\mathcal{H}},
\end{aligned}$$

where the first line used the source condition (A2), and the second to last line used Young's inequality for positive operators $\lambda^{1-2r}\mathcal{L}^{2r} \le (1-2r)\lambda + 2r\mathcal{L}$.

Re-arranging the terms, we have obtained that

$$\begin{aligned}
& C\|f-f_0\|^2 + \frac{\lambda}{2}\|f-f_0\|_{\mathcal{H}}^2 \\
\le\ & \left(\sqrt{2r}\lambda^{\frac{1+2r}{2}} + C\frac{\mathcal{R}_N(u_N)}{u_N}\right)\gamma + \frac{1}{2}\lambda^{1+2r}.
\end{aligned}$$

Then, by the specified value of γ and λ, the above leads to

$$L^2 u_N^{2+4r} \le C\lambda^{1/2+r} L u_N^{1+2r},$$

which is a contradiction if L is sufficiently large. This means $\|\widehat{f}-f_0\| + u_N\|\widehat{f}-f_0\|_{\mathcal{H}} \le L u_N^{1+2r}$.

When we consider the distributed setting, we again take $\lambda \asymp u_N^2$. We still have, following the same steps,

$$\begin{aligned}
& C\|f-f_0\|^2 + \frac{\lambda}{2}\|f-f_0\|_{\mathcal{H}}^2 \\
\le\ & \left(\sqrt{2r}\lambda^{\frac{1+2r}{2}} + C\frac{\mathcal{R}_n(u_N)}{u_N}\right)\gamma + \frac{1}{2}\lambda^{1+2r}.
\end{aligned}$$

Note that we have used $\mathcal{R}_n$ instead of $\mathcal{R}_N$ above, since the estimator is based on a sample of size n now. In this case, setting $\gamma = L\mathcal{R}_n(u_N)/u_N = L\sqrt{\frac{N}{n}}u_N^{1+2r}$ (then we have $\|f-f_0\|_\infty \leq C_1 L\sqrt{\frac{N}{n}}u_N^{1+2r-s} \leq C_5$), the above leads to

$$L^2\frac{N}{n}u_N^{2+4r} \leq CL\frac{N}{n}u_N^{2+4r},$$

again a contraction for L large enough. This means $\|\widehat{f}_j - f_0\| + u_N\|\widehat{f}_j - f_0\|_{\mathcal{H}} \leq L\sqrt{\frac{N}{n}}u_N^{1+2r}$. □

Proof of Theorem 2. By assumption (A6), we have, when $\|f-f_0\| \leq u, \|f-f_0\|_{\mathcal{H}} \leq v$,

$$\begin{aligned} ED_{f,f_0}(x,y) &= \frac{1}{2}\langle f-f_0, \mathcal{L}_H(f-f_0)\rangle_{\mathcal{H}} + O(E[(f(x)-f_0(x))^3]) \\ &= \frac{1}{2}\langle f-f_0, \mathcal{L}_H(f-f_0)\rangle_{\mathcal{H}} + Cu^{3-s}v^s, \end{aligned} \tag{13}$$

using $E[(f(x)-f_0(x))^3] \leq \|f-f_0\|_\infty \|f-f_0\|^2 \leq C\|f-f_0\|^{3-s}\|f-f_0\|_{\mathcal{H}}^s$ based on the sup-norm assumption.

Let $\widetilde{f}_j$ be the minimizer of $\frac{1}{2}\langle f-f_0, \mathcal{L}_H(f-f_0)\rangle_{\mathcal{H}} + P_n((f(x)-f_0(x))\epsilon) + \lambda\|f\|_{\mathcal{H}}^2$, where $\epsilon = \partial_2\ell(y, f_0(x))$. It is easy to see that $\widetilde{f}_j$ has a closed-form expression

$$\widetilde{f}_j = (\mathcal{L}_H + \lambda)^{-1}\mathcal{L}_H f_0 + (\mathcal{L}_{\mathcal{H}} + \lambda)^{-1}\frac{\sum_{i\in\mathcal{S}_j}\epsilon_i K(x_i, .)}{n}. \tag{14}$$

We will show that

$$\|\widetilde{f}_j - f_0\| + u_N\|\widetilde{f}_j - f_0\|_{\mathcal{H}} \leq C\sqrt{\frac{N}{n}}u_N^{1+2r}. \tag{15}$$

In fact, since $\widetilde{f}_j$ is the minimizer of $\frac{1}{2}\langle f-f_0, \mathcal{L}_H(f-f_0)\rangle_{\mathcal{H}} + P_n((f(x)-f_0(x))\epsilon) + \lambda\|f\|_{\mathcal{H}}^2$, by comparing the objective function value at $\widetilde{f}_j$ with that at f_0, we have

$$\frac{1}{2}\langle \widetilde{f}_j - f_0, \mathcal{L}_H(\widetilde{f}_j - f_0)\rangle_{\mathcal{H}} + \frac{1}{n}\sum_i \epsilon_i(\widetilde{f}_j(x_i) - f_0(x_i)) + \lambda\|\widetilde{f}_j\|_{\mathcal{H}}^2 \leq \lambda\|f_0\|_{\mathcal{H}}^2.$$

By the same arguments as at the beginning of the proof of Theorem 1, if $\|\widetilde{f}_j - f_0\| + u_N\|\widetilde{f}_j - f_0\|_{\mathcal{H}} \geq \gamma = L\sqrt{\frac{N}{n}}u_N^{1+2r}$, there exists some f with $\|f-f_0\| + u_N\|f-f_0\|_{\mathcal{H}} = \gamma$ such that

$$\frac{1}{2}\langle f-f_0, \mathcal{L}_H(f-f_0)\rangle_{\mathcal{H}} + \frac{1}{n}\sum_i \epsilon_i(f(x_i) - f_0(x_i)) + \lambda\|f\|_{\mathcal{H}}^2 \leq \lambda\|f_0\|_{\mathcal{H}}^2.$$

Since $\langle f-f_0, \mathcal{L}_H(f-f_0)\rangle_{\mathcal{H}} \geq C\langle f-f_0, \mathcal{L}(f-f_0)\rangle_{\mathcal{H}} = C\|f-f_0\|^2$, we have

$$C\|f-f_0\|^2 \leq \lambda\|f_0\|_{\mathcal{H}}^2 - \lambda\|f\|_{\mathcal{H}}^2 + Cu^{-1}\mathcal{R}_n(u)\gamma,$$

which is the same as (12) and by the same arguments as in the proof of Theorem 1, the above displayed equation would lead to the same rate $\|\widetilde{f}_j - f_0\| + u_N\|\widetilde{f}_j - f_0\|_{\mathcal{H}} \leq C\sqrt{\frac{N}{n}}u_N^{1+2r}$.

Furthermore, using (14), since $\sum_j \widetilde{f}_j/m = (\mathcal{L}_H+\lambda)^{-1}\mathcal{L}_H f_0 + (\mathcal{L}_{\mathcal{H}}+\lambda)^{-1}\frac{\sum_{i=1}^N \epsilon_i K(x_i,.)}{N}$. is almost the same as $\widetilde{f}_j$ except that the sample size is N, we immediately have

$$\|\sum_j \widetilde{f}_j/m - f_0\| + u_N\|\sum_j \widetilde{f}_j/m - f_0\|_{\mathcal{H}} \leq C u_N^{1+2r}.$$

Now we bound $\|\widehat{f}_j - \widetilde{f}_j\| + u_N\|\widehat{f}_j - \widetilde{f}_j\|_{\mathcal{H}}$. For simplicity of notation, define the function $h(u,v) = u^{3-s}v^s + \frac{c'}{\sqrt{n}}\sqrt{H(c',u,v)} + \frac{c}{n}H(c',u,v)$.

By Lemma 3, (A6), and (13), for any $f \in \mathcal{H}$ with $\|f - f_0\| \leq u$ and $\|f - f_0\|_{\mathcal{H}} \leq v$,

$$\begin{aligned} & \frac{1}{n}\sum_i \ell(y_i, f(x_i)) - \frac{1}{n}\sum_i \ell(y_i, f_0(x_i)) - \frac{1}{n}\sum_i \epsilon_i(f(x_i) - f_0(x_i)) \\ & -\frac{1}{2}\langle f - f_0, \mathcal{L}_H(f - f_0)\rangle_{\mathcal{H}} \\ \leq\ & Ch(u,v). \end{aligned} \tag{16}$$

Since $\|\widehat{f}_j - f_0\| \leq u$, $\|\widehat{f}_j - f_0\|_{\mathcal{H}} \leq v$, $\|\widetilde{f}_j - f_0\| \leq u$, $\|\widetilde{f}_j - f_0\|_{\mathcal{H}} \leq v$ with $u = C\sqrt{\frac{N}{n}}u_N^{1+2r}$, $v = C\sqrt{\frac{N}{n}}u_N^{2r}$ by Theorem 1 and equation (15), taking the difference when plugging $f = \widetilde{f}_j$ and $f = \widehat{f}_j$ into (16), we get

$$\begin{aligned} & -\frac{1}{n}\sum_i \ell(y_i, \widehat{f}_j(x_i)) + \frac{1}{n}\sum_i \ell(y_i, \widetilde{f}_j(x_i)) + \frac{1}{n}\sum_i \epsilon_i(\widehat{f}_j(x_i) - \widetilde{f}_j(x_i)) \\ & +(1/2)\langle \widehat{f}_j - f_0, \mathcal{L}_H(\widehat{f}_j - f_0)\rangle_{\mathcal{H}} - (1/2)\langle \widetilde{f}_j - f_0, \mathcal{L}_H(\widetilde{f}_j - f_0)\rangle_{\mathcal{H}} \\ \leq\ & Ch(u,v). \end{aligned} \tag{17}$$

Since $\widetilde{f}_j$ minimizes $\frac{1}{2}\langle f - f_0, \mathcal{L}_H(f - f_0)\rangle_{\mathcal{H}} + \frac{1}{n}\sum_i \epsilon_i(f(x_i) - f_0(x_i)) + \lambda\|f\|_{\mathcal{H}}^2$, the first order condition yields

$$\frac{1}{n}\sum_{i=1}^n \epsilon_i K(x_i,.) + \mathcal{L}_H(\widetilde{f}_j - f_0) + 2\lambda\widetilde{f}_j = 0. \tag{18}$$

Thus we have

$$\begin{aligned} & (1/2)\langle \widehat{f}_j - f_0, \mathcal{L}_H(\widehat{f}_j - f_0)\rangle_{\mathcal{H}} - (1/2)\langle \widetilde{f}_j - f_0, \mathcal{L}_H(\widetilde{f}_j - f_0)\rangle_{\mathcal{H}} \\ & +\frac{1}{n}\sum_i \epsilon_i(\widehat{f}_j(x_i) - \widetilde{f}_j(x_i)) \\ =\ & \frac{1}{2}\langle \widehat{f}_j - \widetilde{f}_j, \mathcal{L}_H(\widehat{f}_j - \widetilde{f}_j)\rangle_{\mathcal{H}} + \langle \widehat{f}_j - \widetilde{f}_j, \mathcal{L}_H(\widetilde{f}_j - f_0)\rangle_{\mathcal{H}} \\ & +\langle \widehat{f}_j - \widetilde{f}_j, \frac{1}{n}\sum_i \epsilon_i K(x_i,.)\rangle_{\mathcal{H}} \\ =\ & \frac{1}{2}\langle \widehat{f}_j - \widetilde{f}_j, \mathcal{L}_H(\widehat{f}_j - \widetilde{f}_j)\rangle_{\mathcal{H}} - 2\lambda\langle \widetilde{f}_j, \widehat{f}_j - \widetilde{f}_j\rangle_{\mathcal{H}}, \end{aligned}$$

where the last step used (18). Thus, using the above in (17), we get

$$\begin{aligned}
& -\frac{1}{n}\sum_i \ell(y_i, \widehat{f}_j(x_i)) + \frac{1}{n}\sum_i \ell(y_i, \widetilde{f}_j(x_i)) \\
& +\frac{1}{2}\langle \widehat{f}_j - \widetilde{f}_j, \mathcal{L}_H(\widehat{f}_j - \widetilde{f}_j)\rangle_{\mathcal{H}} - 2\lambda\langle \widetilde{f}_j, \widehat{f}_j - \widetilde{f}_j\rangle_{\mathcal{H}} \\
= & -\frac{1}{n}\sum_i \ell(y_i, \widehat{f}_j(x_i)) - \lambda\|\widehat{f}_j\|_{\mathcal{H}}^2 + \frac{1}{n}\sum_i \ell(y_i, \widetilde{f}_j(x_i)) + \lambda\|\widetilde{f}_j\|_{\mathcal{H}}^2 \\
& +\frac{1}{2}\langle \widehat{f}_j - \widetilde{f}_j, \mathcal{L}_H(\widehat{f}_j - \widetilde{f}_j)\rangle_{\mathcal{H}} + \lambda\|\widehat{f}_j - \widetilde{f}_j\|_{\mathcal{H}}^2 \\
\leq & Ch(u,v). \qquad (19)
\end{aligned}$$

Noting

$$\frac{1}{n}\sum_i \ell(y_i, \widehat{f}_j(x_i)) + \lambda\|\widehat{f}_j\|_{\mathcal{H}}^2 \leq \frac{1}{n}\sum_i \ell(y_i, \widetilde{f}_j(x_i)) + \lambda\|\widetilde{f}_j\|_{\mathcal{H}}^2,$$

since $\widehat{f}_j$ is the minimizer of $\frac{1}{n}\sum_{i\in\mathcal{S}_j} \ell(y_i, f(x_i)) + \lambda\|f\|_{\mathcal{H}}^2$, the above immediately implies

$$\|\widehat{f}_j - \widetilde{f}_j\|^2 + u_N^2\|\widehat{f}_j - \widetilde{f}_j\|_{\mathcal{H}}^2 \leq Ch(u,v).$$

Finally, the distributed estimator satisfies

$$\begin{aligned}
& \|\bar{f} - f_0\|^2 + u_N^2\|\bar{f} - f_0\|_{\mathcal{H}}^2 \\
\leq & 2\|\frac{\sum_j \widehat{f}_j}{m} - \frac{\sum_j \widetilde{f}_j}{m}\|^2 + 2\|\frac{\sum_j \widetilde{f}_j}{m} - f_0\|^2 + 2u_N^2\|\frac{\sum_j \widehat{f}_j}{m} - \frac{\sum_j \widetilde{f}_j}{m}\|_{\mathcal{H}}^2 + 2u_N^2\|\frac{\sum_j \widetilde{f}_j}{m} - f_0\|_{\mathcal{H}}^2 \\
\leq & 2\max_j(\|\widehat{f}_j - \widetilde{f}_j\|^2 + u_N^2\|\widehat{f}_j - \widetilde{f}_j\|_{\mathcal{H}}^2) + 2\|\frac{\sum_j \widetilde{f}_j}{m} - f_0\|^2 + 2u_N^2\|\frac{\sum_j \widetilde{f}_j}{m} - f_0\|_{\mathcal{H}}^2 \\
\leq & C\left(u^{3-s}v^s + \frac{c'}{\sqrt{n}}\sqrt{H(c',u,v)} + \frac{c}{n}H(c',u,v) + u_N^{2+4r}\right),
\end{aligned}$$

where $u = \sqrt{\frac{N}{n}}u_N^{1+2r}$, $v = \sqrt{\frac{N}{n}}u_N^{2r}$, $c = u^{a+(1-s)b}v^{sb}$, $c' = u^{a'+(1-s)b'}v^{sb'}$. □

Appendix B: On optimal rates for kernel logistic regression

In this section, we prove the following lower bound for kernel logistic regression. We use the following assumption.

(B1) Assume the kernel for the RKHS $\mathcal{H}$ is bounded. The operator $\mathcal{L}$ has eigenvalues $s_j \asymp j^{-\alpha}$ for some $\alpha > 1$. Define the class of functions $\mathcal{H}_r := \{f \in \mathcal{H} : f = \mathcal{L}^r g$ for some $g \in \mathcal{H}$ with $\|g\| \leq 1\}$. The true model is $y \sim Ber(\frac{e^{f(x)}}{1+e^{f(x)}})$ for some $f \in \mathcal{H}_r$ and $Ber(.)$ indicates the Bernoulli distribution.

Proposition 2 *Under Assumption (B1) stated above, there is a constant $C > 0$ such that*

$$\inf_{\widehat{f}} \sup_{f_0\in\mathcal{H}_r} P(\|\widehat{f} - f_0\|_2 \geq Cn^{-\frac{\alpha(2r+1)}{2\alpha(2r+1)+2}}) \geq 1/2,$$

where the infimum is over all possible estimators.

Proof of Proposition 2. The proof follows basically the same lines as for Theorem 2 of Raskutti et al. (2012) or Theorem 4 of Suzuki and Sugiyama (2013) (although both studied a more general case of sparse additive models under least squares loss), both of which used arguments of Yang and Barron (1999). We outline the proof here and point out the modifications required. Let δ_n, ϵ_n be positive sequences (specific choices to be stated later) and $M := \mathcal{M}(\delta_n, \mathcal{H}_r, L^2)$ be the δ_n-packing number of $\mathcal{H}_r$ and $N := \mathcal{N}(\epsilon_n, \mathcal{H}_r, L^2)$ be the ϵ_n-covering number of $\mathcal{H}_r$, with the δ_n-packing set $\{f^1, \ldots, f^M\}$ and the ϵ_n-net $\{g^1, \ldots, g^N\}$. Let Θ be a discrete random variable taking value in $\{1, \ldots, M\}$ which indicates the true function is f^m if $\Theta = m$.

As in Raskutti et al. (2012); Suzuki and Sugiyama (2013), we have

$$\inf_{\widehat{f}} \sup_{F \in \mathcal{F}} P(\|\widehat{f} - f_0\|_2^2 \geq \delta_n^2/2) \geq \left(1 - \frac{E[I_{x_1^n}(\Theta, y_1^n)] + \log(2)}{\log(M)}\right),$$

where $I_{x_1^n}(\Theta, y_1^n)$ is the mutual information between Θ and $y_1^n = (y_1, \ldots, y_n)^{\mathrm{T}}$ given some fixed $x_1^n = (x_1, \ldots, x_n)^{\mathrm{T}}$.

Furthermore, they showed that

$$I_{x_1^n}(\Theta, y_1^n) \leq \log(N) + \frac{1}{M} \sum_{m=1}^{M} \min_{\ell} D(P_{y_1^n|x_1^n, f^m} \| P_{y_1^n|x_1^n, g^\ell}),$$

where $\|f\|_n^2 = \frac{1}{n}\sum_i f^2(x_i)$, $P_{y_1^n|x_1^n, f}$ denotes the conditional distribution of y_1^n given x_1^n when f is the true function, and $D(.\|.)$ is the Kullback-Leibler (KL) divergence.

For least squares regression with Gaussian noise $N(0, \sigma^2)$, it can be easily calculated that $D(P_{y_1^n|x_1^n, f^m} \| P_{y_1^n|x_1^n, g^\ell}) = \frac{n}{2\sigma^2}\|f^m - g^\ell\|_n^2$. Using this closed-form expression of divergence, and that $\log(N) \sim \epsilon_n^{-\frac{2}{\alpha(2r+1)}}, \log(M) \sim \delta_n^{-\frac{2}{\alpha(2r+1)}}$, choosing $\epsilon_n \sim n^{-\frac{\alpha(2r+1)}{2\alpha(2r+1)+2}}$ and $\delta_n = C\epsilon_n$ for a sufficiently small constant $C > 0$, we can get

$$1 - \frac{E[I_{x_1^n}(\Theta, y_1^n)] + \log(2)}{\log(M)} \geq 1/2,$$

for the Gaussian mean regression model.

By the above arguments, exactly the same lower bound can be obtained if $D(P_{y_1^n|x_1^n, f^m} \| P_{y_1^n|x_1^n, g^\ell}) = \frac{n}{2\sigma^2}\|f^m - g^\ell\|_n^2$ is replaced by the inequality

$$D(P_{y_1^n|x_1^n, f^m} \| P_{y_1^n|x_1^n, g^\ell}) \leq Cn\|f^m - g^\ell\|_n^2, \tag{20}$$

and thus we only need to show the logistic regression model satisfies (20).

In fact, when using negative log-likelihood as the loss function, we have $E_{f^m}[\ell(y, g^\ell(x)|x] - E_{f^m}[\ell(y, f^m(x)|x] = D(P_{y_1^n|x_1^n, f^m} \| P_{y_1^n|x_1^n, g^\ell})$ where $E_{f^m}[.|x]$ is the conditional expectation when f^m is the true function (using the subscript to emphasize the truth here), and thus (20) can be verified similar to the proof of Proposition 1 using Taylor's expansion. This completes the proof. □

Acknowledgments

We sincerely thank Professor Pradeep Ravikumar and Professor Sanmi Koyejo for handling our paper and providing insightful comments. Thanks also go to three anonymous reviewers for their great comments and suggestions that substantially improved the manuscript. The research of Heng Lian is partially supported by NSFC 12371297 at CityUHK Shenzhen Research Institute, NSFC-Shenzhen Projedt No. JCYJ20250604191213017 at CityUHK Shenzhen Research Institute, and by Hong Kong RGC general research fund 11300721 and 11311822, 11300424.

References

Maria-Florina Balcan, Yingyu Liang, Le Song, David Woodruff, and Bo Xie. Communication efficient distributed kernel principal component analysis. *arXiv:1503.06858*, mar 2015.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *Annals of Statistics*, 33:1497–1537, 2005.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational mathematics*, 7(3):331–368, 2007. ISSN 1615-3375.

Xiangyu Chang, Shao-Bo Lin, and Ding-Xuan Zhou. Distributed semi-supervised learning with kernel ridge regression. *Journal of Machine Learning Research*, 18:1–46, 2017.

Zheng Chu Guo, Shao Bo Lin, and Ding Xuan Zhou. Learning theory of distributed spectral algorithms. *Inverse Problems*, 77:074009, 2017. ISSN 13616420. doi: 10.1088/1361-6420/aa72b2.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114:668–681, 2018. ISSN 1537274X. doi: 10.1080/01621459.2018.1429274.

V Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*. Springer, New York, 2011.

P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10:365–277, 2000. ISSN 0129-0657. doi: 10.1142/S012906570000034X.

Jason D Lee, Qiang Liu, Yuekai Sun, and Jonathan E Taylor. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18:1–30, 2017.

Youjuan Li, Yufeng Liu, and Ji Zhu. Quantile regression in reproducing kernel Hilbert spaces. *Journal of the American Statistical Association*, 102:255–268, 2007. ISSN 01621459. doi: 10.1198/016214506000000979.

Heng Lian. Distributed learning of conditional quantiles in the reproducing kernel hilbert space. *Advances in Neural Information Processing Systems*, 35:11686–11696, 2022.

Heng Lian and Zengyan Fan. Divide-and-conquer for debiased l1-norm support vector machine in ultra-high dimensions. *Journal of Machine Learning Research*, 1:1–26, 2018.

Junhong Lin and Volkan Cevher. Optimal convergence for distributed learning with stochastic gradient methods and spectral algorithms. *Journal of Machine Learning Research*, 21: 1–63, 2020.

Shao-Bo Lin, Xin Guo, and Ding-Xuan Zhou. Distributed learning with regularized least squares. *Journal of Machine Learning Research*, 18:1–31, 2017.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: fast rates for regularized empirical risk minimization through self-concordance. In *32nd Annual Conference on Learning Theory*, 2019.

Shahar Mendelson. Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43, 2002.

S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.R. Mullers. Fisher discriminant analysis with kernels. In *Neural Networks for Signal Processing*, pages 41–48, 1999. doi: 10.1109/nnsp.1999.788121.

Patric Müller and Sara van de Geer. The partial linear model in high dimensions. *Scandinavian Journal of Statistics*, 42:580–608, 2015. ISSN 14679469. doi: 10.1111/sjos.12124.

Angelia Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4): 2597–2633, 2017. ISSN 10526234. doi: 10.1137/16M1084316.

Angelina Nedić and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009. ISSN 00189286. doi: 10.1109/TAC.2008.2009515.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research*, 13:389–427, 2012. ISSN 1532-4435.

Jonathan Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *arXiv:1407.2724*, 2015.

Zuofeng Shang and Guang Cheng. Local and global asymptotic inferences for the smoothing spline estimate. *Annals of Statistics*, 41:2608–2638, 2013.

J Shawe-Taylor. Kernel learning for novelty detection. *NIPS 2008 Workshop on Kernel Learning: Automatic Selection of Optimal Kernel*, 2008.

John Shawe-Taylor, Christopher K.I. Williams, Nello Cristianini, and Jaz Kandola. On the eigenspectrum of the gram matrix and the generalization error of Kernel-PCA. *IEEE Transactions on Information Theory*, 51:2510—-2522, 2005. ISSN 00189448. doi: 10.1109/TIT.2005.850052.

Karthik Sridharan, Nathan Srebro, and Shai Shalev-Shwartz. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21 - Proceedings of the 2008 Conference*, 2009.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines using Gaussian kernels. *The Annals of Statistics*, 35:575–607, 2007.

Ingo Steinwart, Don Hush, and Clint Scovel. Optimal rates for regularized least squares regression. In *COLT 2009 - The 22nd Conference on Learning Theory*, 2009.

Xiaoxiao Sun, Wenxuan Zhong, and Ping Ma. An asymptotic and empirical smoothing parameters selection method for smoothing spline ANOVA models in large samples. *Biometrika*, 108(1):149–166, 2021. ISSN 14643510. doi: 10.1093/biomet/asaa047.

Taiji Suzuki and Masashi Sugiyama. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics*, 41:1381–1405, 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1095.

A W van der Vaart and J A Wellner. *Weak convergence and empirical processes*. Springer Verlag, New York, 1996.

S. Volgushev, S. Chao, and G. Cheng. Distributed inference for quantile regression processes. *Annals of Statistics*, 47:1634–1662, 2019.

Grace Wahba. *Spline models for observational data*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1990.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999. ISSN 00905364. doi: 10.1214/aos/1017939142.

Yuchen Zhang, John C. Duchi, and Martin J. Wainwright. Communication-efficient algorithms for statistical optimization. *Journal of Machine Learning Research*, 14:3321–3363, 2013.

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

Tianqi Zhao, Guang Cheng, and Han Liu. A Partially Linear Framework for Massive Heterogeneous Data. *Annals of Statistics*, 44, 2016.