Distributed Nonparametric Regression under Communication Constraints

Yuancheng Zhu [1] **John Lafferty** [2]

Abstract

This paper studies the problem of nonparametric estimation of a smooth function with data distributed across multiple machines. We assume an independent sample from a white noise model is collected at each machine, and an estimator of the underlying true function needs to be constructed at a central machine. We place limits on the number of bits that each machine can use to transmit information to the central machine. Our results give both asymptotic lower bounds and matching upper bounds on the statistical risk under various settings. We identify three regimes, depending on the relationship among the number of machines, the size of data available at each machine, and the communication budget. When the communication budget is small, the statistical risk depends solely on this communication bottleneck, regardless of the sample size. In the regime where the communication budget is large, the classic minimax risk in the non-distributed estimation setting is recovered. In an intermediate regime, the statistical risk depends on both the sample size and the communication budget.

1. Introduction

Classic statistical theory studies the difficulty of estimation under various models, and attempts to find the optimal estimation procedures. Such studies usually assume that all of the collected data are available to construct the estimators. In this paper, we study the problem of statistical estimation with data residing at multiple machines. Estimation in distributed settings is becoming common in modern data analysis tasks, as the data can be collected or stored at different locations. In order to obtain an estimate of some statistical functional, information needs to be gathered and aggregated from the multiple locations to form the final estimate. However, the communication between machines may be limited. For instance, there may be a communication budget that limits how much information can be transmitted. In this setting, it is important to understand how the statistical risk of estimation degrades as the communication budget becomes more limited.

A similar problem, called the CEO problem, was first studied in the electrical engineering community from a rate-distortion-theory perspective (Berger et al., 1996; Viswanathan & Berger, 1997). More recently, several studies have focused on more specific statistical tasks and models; see, for example, Zhang et al. (2013a); Shamir (2014); Battey et al. (2015); Braverman et al. (2016); Diakonikolas et al. (2017); Fan et al. (2017); Lee et al. (2017) treating mean estimation, regression, principal eigenspace estimation, discrete density estimation and other problems. Most of this existing research focuses on parametric and discrete models, where the parameter of interest has a finite dimension. While there are also studies of nonparametric problems and models (Zhang et al., 2013b; Blanchard & Mücke, 2016; Chang et al., 2017; Shang & Cheng, 2017), the fundamental limits of distributed nonparametric estimation are still under-explored.

In this paper, we consider a fundamental nonparametric estimation task—estimating a smooth function in the white noise model. We assume observation of the random process

$$\mathrm{d}Y(t) = f(t)\mathrm{d}t + \frac{1}{\sqrt{n}}\mathrm{d}W(t), \quad 0 \le t \le 1, \tag{1.1}$$

where $\frac{1}{\sqrt{n}}$ is the noise level, $W(t)$ is a standard Wiener process, and f is the underlying function to be estimated. The white noise model is a centerpiece of nonparametric estimation, being asymptotically equivalent to nonparametric regression and density estimation (Brown & Low, 1996; Nussbaum, 1996). We intentionally express the noise level as $\frac{1}{\sqrt{n}}$ to reflect the connection between the white noise model and a nonparametric regression problem with n evenly spaced observations. We focus on the important case where the regression function lies in the Sobolev space $\mathcal{F}(\alpha, c)$ of order α and radius c; the exact definition of this function space is given in the following section.

In a distributed setting, instead of observing a single sample

[1]Department of Statistics, Wharton School, University of Pennsylvania [2]Department of Statistics and Data Science, Yale University. Correspondence to: Yuancheng Zhu <yuancheng.zhu@gmail.com>.

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

path $Y(t)$, we assume there are m machines, each of which observes an independent copy of the stochastic process. That is, the jth machine gets

$$\mathrm{d}Y_j(t) = f(t)\mathrm{d}t + \frac{1}{\sqrt{n}}\mathrm{d}W_j(t), \quad 0 \le t \le 1,$$

for $j = 1, \ldots, m$ where $W_j(t)$'s are mutually independent standard Wiener processes. Furthermore, each machine has a budget of b bits to communicate with a central machine, where a final estimate $\widehat{f}$ is formed based on the messages received from the m machines. Specifically, we denote by Π_j the message that the jth machine sends to the central estimating machine; each Π_j can be viewed as a (possibly random) functional of the stochastic process $Y_j(t)$. In this way, the tuple (n, m, b) defines a problem instance for the function class $\mathcal{F}(\alpha, c)$. We use the minimax risk

$$\begin{aligned} &R(n, m, b; \mathcal{F}(\alpha, c)) \\ &= \inf_{\widehat{f}, \Pi_{1:m}} \sup_{f \in \mathcal{F}(\alpha, c)} \mathbb{E}\, \|f - \widehat{f}(\Pi_1, \ldots, \Pi_m)\|^2 \end{aligned}$$

to quantify the hardness of distributed estimation of f in the Sobolev space $\mathcal{F}(\alpha, c)$.

The main contribution of the paper is to identify the following three asymptotic regimes.

- An *insufficient regime* where $mb \ll n^{\frac{1}{2\alpha+1}}$. Under this scaling, the total number of bits, mb, is insufficient to preserve the classical, non-distributed, minimax rate of convergence for the sample size n on a single machine. Therefore, the communication budget becomes the main bottleneck, and we have

  $$R(n, m, b; \mathcal{F}(\alpha, c)) \asymp (mb)^{-2\alpha}.$$

- A *sufficient regime* where $b \gg (mn)^{\frac{1}{2\alpha+1}}$. In this case, the number of bits allowed per machine is relatively large, and we have the minimax risk

  $$R(n, m, b; \mathcal{F}(\alpha, c)) \asymp (mn)^{-\frac{2\alpha}{2\alpha+1}}.$$

  Note that this is also the optimal convergence rate if all the data were available at the central machine.

- An *intermediate regime* where $b \lesssim (mn)^{\frac{1}{2\alpha+1}}$ and $mb \gtrsim n^{\frac{1}{2\alpha+1}}$. In this regime, the minimax risk depends on all three parameters, and scales according to

  $$R(n, m, b; \mathcal{F}(\alpha, c)) \asymp (mnb)^{-\frac{\alpha}{\alpha+1}}.$$

Together, these three regimes give a sharp characterization of the statistical behavior of distributed nonparametric estimation for the Sobolev space $\mathcal{F}(\alpha, c)$ under communication constraints, covering the full range of parameters and problem settings. The Bayesian framework adopted in this paper to establish the lower bounds is different from the techniques used in previous work, which typically rely on Fano's lemma and the strong data processing inequality. Finally, we note that an essentially equivalent set of minimax convergence rates is obtained in a simultaneously and independently written paper by Szabo & van Zanten (2018).

The paper is organized as follows. In the next section, we explain our notation and give a brief introduction of nonparametric estimation over a Sobolev space for the usual non-distributed setting and a distributed setting. In Section 3, we state our main results on the risk of distributed nonparametric estimation with communication constraints. We outline the proof strategy for the lower bounds in Section 3.1, deferring some of the technical details and proofs to the supplementary material. In Section 4, we show achievability of the lower bounds by a particular distributed protocol and estimator. We conclude the paper with a discussion of possible directions for future work.

2. Problem formulation

The Sobolev space of order α and radius c is defined by

$$\begin{aligned} \mathcal{F}(\alpha, c) = \Bigg\{ &f : f^{(\alpha-1)} \text{ is absolutely continuous}, \\ &\int_0^1 (f^{(\alpha)}(t))^2 \mathrm{d}t \le c^2, \text{ and } f \in [0,1] \to \mathbb{R} \Bigg\}. \end{aligned}$$

Intuitively, it is a space of functions having a certain degree of smoothness, increasing with α. The periodic Sobolev space is defined by

$$\begin{aligned} \widetilde{\mathcal{F}}(\alpha, c) = &F(\alpha, c) \bigcap \\ &\Big\{ f^{(j)}(0) = f^{(j)}(1),\ j = 0, 1, \ldots, \alpha - 1 \Big\}. \end{aligned}$$

The white noise model (1.1) can be reformulated in terms of an infinite Gaussian sequence model. Let $(\varphi_i)_{i=1}^{\infty}$ be the trigonometric basis, and let

$$\theta_i = \int_0^1 \varphi_i(t) f(t) \mathrm{d}t, \quad i = 1, 2, \ldots$$

be the Fourier coefficients. It is known that f belongs to $\widetilde{\mathcal{F}}(\alpha, c)$ if and only if the sequence θ belongs to the Sobolev ellipsoid $\Theta(\alpha, c)$, defined as

$$\Theta(\alpha, c) = \left\{ \theta : \sum_{i=1}^{\infty} a_i^2 \theta_i^2 \le \frac{c^2}{\pi^{2\alpha}} \right\}$$

where

$$a_i = \begin{cases} i^\alpha & \text{if } i \text{ is even} \\ (i-1)^\alpha & \text{if } i \text{ is odd}. \end{cases}$$

To ease the analysis, we will assume $a_i = i^\alpha$ and use $\widetilde{c}^2$ in the place of $\frac{c^2}{\pi^{2\alpha}}$. Expanding the observed process $Y(t)$ in terms of the same basis we obtain the Gaussian sequence

$$X_i = \int_0^1 \varphi_i(t)\,\mathrm{d}Y(t) \sim N\left(\theta_i, 1/n\right).$$

Given an estimator $\widehat{\theta}$ for θ, we can formulate a corresponding estimator for f by

$$\widehat{f}(t) = \sum_{i=1}^{\infty} \widehat{\theta}_i \varphi_i(t),$$

and the squared errors satisfy $\|\widehat{\theta} - \theta\|^2 = \|\widehat{f} - f\|^2$. In this way, estimating the function f in the white noise model is equivalent to estimating the means θ in the Gaussian sequence model.

The minimax risk of estimating f over the periodic Sobolev space is defined as

$$R(n; \widetilde{\mathcal{F}}(\alpha, c)) = \inf_{\widehat{f}} \sup_{f \in \mathcal{F}(\alpha,c)} \mathbb{E}\,\|\widehat{f} - f\|^2,$$

which, as just shown, is equal to the minimax risk of estimating θ over the Sobolev ellipsoid in the corresponding Gaussian sequence model,

$$R(n; \Theta(\alpha, \widetilde{c})) = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta(\alpha, \widetilde{c})} \mathbb{E}\,\|\widehat{\theta} - \theta\|^2.$$

It is well known that the asymptotic minimax risk scales according to

$$R(n; \widetilde{\mathcal{F}}(\alpha, c)) = R(n; \Theta(\alpha, c)) \asymp n^{-\frac{2\alpha}{2\alpha+1}}$$

as $n \to \infty$ (Tsybakov, 2008).

In a distributed setting, we suppose there are m machines, and the jth machine independently observes $Y_j(t)$ such that

$$\mathrm{d}Y_j(t) = f(t)\mathrm{d}t + \frac{1}{\sqrt{n}}\mathrm{d}W_j(t) \quad 0 \le t \le 1$$

for $j = 1, \ldots, m$. Equivalently, if we express this in terms of the Gaussian sequence model, the jth machine observes data

$$X_{ij} \sim N(\theta_i, 1/n), \quad i = 1, 2, \ldots.$$

We further assume there is a central machine where a final estimator needs to be calculated based on messages received from the m local machines. Local machine j sends a message of length b_j bits to the central machine; we denote this message by Π_j. Then $\Pi_j = \Pi_j(X_{1j}, X_{2j}, \ldots)$ can be viewed as a (possibly random) mapping from $\mathbb{R}^\infty$ to $\{1, 2, \ldots, 2^{b_j}\}$. The final estimator $\widehat{\theta}$ is then a functional of the collection of messages. The mechanism can be summarized by the following diagram:

$$f \longrightarrow \left\{ \begin{array}{c} Y_1(t) \longrightarrow X_{11}, \ldots, X_{n1} \xrightarrow{b_1} \Pi_1 \\ Y_2(t) \longrightarrow X_{12}, \ldots, X_{n2} \xrightarrow{b_2} \Pi_2 \\ \vdots \\ Y_m(t) \longrightarrow X_{1m}, \ldots, X_{nm} \xrightarrow{b_m} \Pi_m \end{array} \right\} \longrightarrow \widehat{\theta} \longrightarrow \widehat{f}.$$

Suppose that the communication is restricted by one of two types of constraints: An *individual constraint*, where $b_j \le b$, for each $j = 1, \ldots, m$ and a given budget b, and a *sum constraint*, where $\sum_{j=1}^m b_j \le mb$. We call the set of mappings $\Pi_1, \ldots, \Pi_m$ and $\widehat{\theta}$ a *distributed protocol*, and denote by $\Gamma_{\text{ind}}(m, b)$ and $\Gamma_{\text{sum}}(m, b)$ the collection of all such protocols, operating under the individual constraint and the sum constraint, respectively.

We note here that for simplicity we consider only one round of communication. A variant is to allow multiple rounds of communication, for which the local machines can get access to a "blackboard" where the central machine broadcasts information back to the distributed nodes.

The minimax risk of the distributed estimation problem under the communication constraint is defined by

$$\begin{aligned} &R(n, m, b; \Theta(\alpha, c)) \\ &= \inf_{(\Pi_1,\ldots,\Pi_m,\widehat{\theta}) \in \Gamma(m,b)} \sup_{\theta \in \Theta(\alpha,c)} \mathbb{E}\,\|\widehat{\theta}(\Pi_1, \ldots, \Pi_m) - \theta\|^2. \end{aligned} \tag{2.1}$$

Here Γ represents either Γ_{ind} or Γ_{sum}. In fact, it will be clear that the minimax risks under the two types of constraints are asymptotically equivalent.

3. Lower bounds for distributed estimation

In what follows, we will work in an asymptotic regime where the tuple (n, m, b) goes to infinity while satisfying some relationships, and show how the minimax risk for the distributed estimation problem scales accordingly. The main result can be summarized in the following theorem.

Theorem 3.1. *Let $R(n, m, b; \Theta(\alpha, c))$ be defined as in (2.1) with $\Gamma = \Gamma_{sum}$*

1. *If $b(mn)^{-\frac{1}{2\alpha+1}} \to \infty$, then*
$$\liminf_{mn\to\infty} (mn)^{\frac{2\alpha}{2\alpha+1}} R(n, m, b; \Theta(\alpha, c)) \ge C.$$
2. *If $b(mn)^{-\frac{1}{2\alpha+1}} = O(1)$ and $mbn^{-\frac{1}{2\alpha+1}} \to \infty$, then*
$$\liminf_{mn\to\infty} (mnb)^{\frac{\alpha}{\alpha+1}} R(n, m, b; \Theta(\alpha, c)) \ge C.$$
3. *If $mbn^{-\frac{1}{2\alpha+1}} = O(1)$, then*
$$\liminf_{mn\to\infty} (mb)^{2\alpha} R(n, m, b; \Theta(\alpha, c)) \ge C.$$

Remark 3.1. The lower bounds are valid for both the sum constraint and the individual constraint. In fact, the individual constraint is more stringent than the sum constraint, so in terms of lower bounds, it suffices to prove it for the sum constraint.

Remark 3.2. To put the result more concisely, we can write

$$R(n, m, b; \Theta(\alpha, c)) \gtrsim \begin{cases} (mn)^{-\frac{2\alpha}{2\alpha+1}} & \text{if } b(mn)^{-\frac{1}{2\alpha+1}} \to \infty \\ (mbn)^{-\frac{\alpha}{\alpha+1}} & \text{if } b(mn)^{-\frac{1}{2\alpha+1}} = O(1) \\ & \quad \text{and } mbn^{-\frac{1}{2\alpha+1}} \to \infty \\ (mb)^{-2\alpha} & \text{if } mbn^{-\frac{1}{2\alpha+1}} = O(1) \end{cases}.$$

There are multiple ways to interpret this main result and here we illustrate one of the many possibilities. Fixing m and b, and viewing the minimax risk as a function of n, the sample size on each machine, we have

$$R(n) \gtrsim \begin{cases} n^{-\frac{2\alpha}{2\alpha+1}} m^{\frac{2\alpha}{2\alpha+1}} & \text{if } n \lesssim \frac{b^{2\alpha+1}}{m} \\ n^{-\frac{\alpha}{\alpha+1}} (mb)^{-\frac{\alpha}{\alpha+1}} & \text{if } \frac{b^{2\alpha+1}}{m} \ll n \ll (mb)^{2\alpha+1} \\ (mb)^{-2\alpha} & \text{if } n \gtrsim (mb)^{2\alpha+1} \end{cases}.$$

This indicates that when the configuration of machines and communication budget stay the same, as we increase the sample size at each machine, the risk starts to decay at the optimal rate with exponent $-\frac{2\alpha}{2\alpha+1}$. Once the sample size is large enough, the convergence rate slows down to an exponent $-\frac{\alpha}{\alpha+1}$. Eventually, the sample size exceeds a threshold, beyond which any further increase won't decrease the risk due to the communication constraint.

Remark 3.3. This work can be viewed as a natural generalization of Zhu & Lafferty (2017), where the authors consider estimation over a Sobolev space with a single remote machine and communication constraints. Specifically, by setting $m = 1$ we recover the main results in Zhu & Lafferty (2017) up to some constant factor. However, with more than one machine, it is non-trivial to uncover the minimax convergence rate, especially in the intermediate regime.

3.1. Proof of the lower bounds

We now proceed to outline the proof of the lower bounds in Theorem 3.1. Most existing results rely on Fano's lemma and the strong data processing inequality (Zhang et al., 2013a; Braverman et al., 2016). An extension of this information-theoretic approach is used by Szabo & van Zanten (2018) in the nonparametric setting to obtain essentially the same lower bounds as we establish here. However, we develop the Bayesian framework for deriving minimax lower bounds (Johnstone, 2017), circumventing the need for both Fano's lemma and the strong data processing inequality, and associating the lower bounds with the solution of an optimization problem.

We consider a prior distribution $\pi(\theta)$ asymptotically supported on the parameter space Θ. For any estimator $\widehat{\theta}$ that follows the distributed protocol, we have

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \, \|\widehat{\theta} - \theta\|^2 \gtrsim \int_\Theta \mathbb{E}_\theta \, \|\widehat{\theta} - \theta\|^2 \mathrm{d}\pi(\theta). \tag{3.1}$$

That is, the worst-case risk associated with $\widehat{\theta}$ is bounded from below by the integrated risk. We specifically consider the Gaussian prior distribution $\theta_i \sim N(0, \sigma_i^2)$ for $i = 1, \ldots, \ell$, and $\mathbb{P}(\theta_i = 0) = 1$ for $i = \ell + 1, \ldots,$ where the sequence σ_i satisfies $\sum_{i=1}^\ell i^{2\alpha}\sigma_i^2 \le c^2$. We make (3.1) clear in the following lemma, whose proof can be found in the supplementary material.

Lemma 3.1. *Suppose that a sequence of Gaussian prior distributions for θ and estimator $\widehat{\theta}$ satisfy*

$$\frac{\sum_{i=1}^\ell i^{2\alpha}\sigma_i^2}{\max_{1\le i\le \ell} i^{2\alpha}\sigma_i^2} = O(\ell) \text{ and} \qquad \int_\Theta \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|^2]\mathrm{d}\pi(\theta) = O(\ell^\delta) \tag{3.2}$$

for some $\delta > 0$ as $\ell \to \infty$. Then

$$\sup_{\theta\in\Theta} \mathbb{E}_\theta \, \|\widehat{\theta} - \theta\|^2 \ge \int_\Theta \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|^2]\mathrm{d}\pi(\theta) \cdot (1 + o(1)).$$

The next step is to lower bound the integrated risk $\int_\Theta \mathbb{E}_\theta[\|\widehat{\theta} - \theta\|^2]\mathrm{d}\pi(\theta)$. Lemma 3.2 is derived from a result that appears in Wang et al. (2010); for completeness we include the proof in the supplementary material.

Lemma 3.2. *Suppose $\theta_i \sim N(0, \sigma_i^2)$ and $X_{ij} \sim N(\theta_i, \varepsilon^2)$ for $i = 1, \ldots, \ell$ and $j = 1, \ldots, m$. Let $\Pi_j : \mathbb{R}^\ell \to \{1, \ldots, M_j\}$ be a (random) mapping, which takes up to M_j different values. Let $\widehat{\theta} : \{1, \ldots, M_1\} \times \cdots \times \{1, \ldots, M_m\} \to \mathbb{R}^\ell$ be an estimator based on the messages created by $\Pi_1, \ldots, \Pi_m$. Under the constraint that $\frac{1}{m}\sum_{j=1}^m \log M_j \le b$, $\mathbb{E}\,\|\widehat{\theta} - \theta\|^2$ can be lower bounded by the value of the following optimization problem*

$$\begin{aligned} &L(m, b, \varepsilon; \sigma) \triangleq \\ &\min_{d_i,\ i=1,\ldots,\ell} \ \sum_{i=1}^\ell d_i \\ &\quad \text{s.t.} \ \sum_{i=1}^\ell \left(\frac{1}{2}\log\frac{\sigma_i^2}{d_i} + \frac{m}{2}\log\frac{\frac{m}{\varepsilon^2}}{\frac{1}{\sigma_i^2} + \frac{m}{\varepsilon^2} - \frac{1}{d_i}} \right) \le mb \\ &\quad \frac{\sigma_i^2 \frac{\varepsilon^2}{m}}{\sigma_i^2 + \frac{\varepsilon^2}{m}} \le d_i \le \sigma_i^2 \text{ for } i = 1, \ldots, \ell. \end{aligned} \tag{3.3}$$

Combining the Lemma 3.1 and 3.2, we have the following asymptotic lower bound

$$R(m, b, n; \Theta(\alpha, c)) \gtrsim L(m, b, n^{-\frac{1}{2}}; \sigma)$$

for sequences σ_i satisfying $\sum_{i=1}^\ell i^{2\alpha}\sigma_i^2 \le \tilde{c}^2$ and $\frac{\sum_{i=1}^\ell i^{2\alpha}\sigma_i^2}{\max_{1\le i\le \ell} i^{2\alpha}\sigma_i^2} = O(\ell)$ as $\ell \to \infty$.

Next, based on the optimization problem formulated above, we work under three different regimes, and derive three forms of lower bounds of the minimax risk. The key is to choose appropriate sequences of prior variances σ_i^2 for different regimes, as we shall illustrate.

1. Suppose that $d_1, \dots, d_\ell$ is a feasible solution to the problem (3.3). Using the first constraint, we have

$$\begin{aligned} mb &\geq \sum_{i=1}^{\ell} \left(\frac{1}{2} \log \frac{\sigma_i^2}{d_i} + \frac{m}{2} \log \frac{\frac{m}{\varepsilon^2}}{\frac{1}{\sigma_i^2} + \frac{m}{\varepsilon^2} - \frac{1}{d_i}} \right) \\ &\geq \sum_{i=1}^{\ell} \frac{1}{2} \log \frac{\sigma_i^2}{d_i} \\ &\geq \sum_{i=1}^{\ell} \frac{1}{2} \log \sigma_i^2 - \frac{\ell}{2} \log \left(\frac{1}{\ell} \sum_{i=1}^{\ell} d_i \right) \end{aligned}$$

where we have used Jensen's inequality. Therefore,

$$\begin{aligned} \sum_{i=1}^{\ell} d_i &\geq \ell \exp \left(\frac{1}{\ell} \sum_{i=1}^{\ell} \log \sigma_i^2 - \frac{2mb}{\ell} \right) \\ &= \ell \left(\prod_{i=1}^{\ell} \sigma_i^2 \right)^{\frac{1}{\ell}} e^{-\frac{2mb}{\ell}}. \end{aligned}$$

Consider an asymptotic regime where $mbn^{-\frac{1}{2\alpha+1}} = O(1)$, and pick a sequence of corresponding prior distributions with $\ell = \gamma mb$ for some constant γ and $\sigma_i^2 = \frac{\widetilde{c}^2}{i^{2\alpha} \ell}$ for $i = 1, \dots, \ell$. Note that this choice satisfies condition (3.2). With such a choice of the prior distribution, we have

$$\begin{aligned} \sum_{i=1}^{\ell} d_i &\geq \ell \left(\prod_{i=1}^{\ell} \frac{\widetilde{c}^2}{i^{2\alpha} \ell} \right)^{\frac{1}{\ell}} e^{-\frac{2mb}{\ell}} \\ &= \widetilde{c}^2 e^{-\frac{2mb}{\ell}} (\ell!)^{-\frac{2\alpha}{\ell}} \\ &\geq \widetilde{c}^2 e^{-\frac{2mb}{\ell}} (e\ell^{\ell+\frac{1}{2}} e^{\ell})^{-\frac{2\alpha}{\ell}} \\ &\geq \widetilde{c}^2 e^{-4\alpha} \cdot e^{-\frac{2mb}{\ell}} \ell^{-2\alpha - \frac{\alpha}{\ell}} \\ &\sim (mb)^{-2\alpha}. \end{aligned}$$

2. Again suppose that $d_1, \dots, d_\ell$ is a feasible solution to the problem (3.3). This time we take another viewpoint of the first constraint

$$\begin{aligned} mb &\geq \sum_{i=1}^{\ell} \left(\frac{1}{2} \log \frac{\sigma_i^2}{d_i} + \frac{m}{2} \log \frac{\frac{m}{\varepsilon^2}}{\frac{1}{\sigma_i^2} + \frac{m}{\varepsilon^2} - \frac{1}{d_i}} \right) \\ &\geq \sum_{i=1}^{\ell} \frac{m}{2} \log \frac{\frac{m}{\varepsilon^2}}{\frac{1}{\sigma_i^2} + \frac{m}{\varepsilon^2} - \frac{1}{d_i}}. \end{aligned}$$

To minimize $\sum_{i=1}^{\ell} d_i$ under the constraint that $\sum_{i=1}^{\ell} \frac{1}{2} \log \frac{\frac{m}{\varepsilon^2}}{\frac{1}{\sigma_i^2} + \frac{m}{\varepsilon^2} - \frac{1}{d_i}} \leq b$, we write the Lagrangian

$$L = \sum_{i=1}^{\ell} d_i + \lambda \left(\sum_{i=1}^{\ell} \frac{1}{2} \log \frac{\frac{m}{\varepsilon^2}}{\frac{1}{\sigma_i^2} + \frac{m}{\varepsilon^2} - \frac{1}{d_i}} - b \right)$$

and set

$$0 = \frac{\partial L}{\partial d_i} = 1 - \frac{\lambda}{2} \frac{1}{\frac{1}{\sigma_i^2} + \frac{m}{\varepsilon^2} - \frac{1}{d_i}}.$$

Solving this gives us that

$$\begin{aligned} \sum_{i=1}^{\ell} d_i &\geq \sum_{i=1}^{\ell} \frac{1}{\frac{1}{\sigma_i^2} + \frac{m}{\varepsilon^2} \left(1 - e^{-\frac{2b}{\ell}} \right)} \\ &\geq \sum_{i=1}^{\ell} \frac{1}{\frac{1}{\sigma_i^2} + \frac{2mb}{\varepsilon^2 \ell}}. \end{aligned}$$

This time, consider a regime where $b(mn)^{-\frac{1}{2\alpha+1}} = O(1)$ and $mbn^{-\frac{1}{2\alpha+1}} \to \infty$. Pick a sequence of corresponding prior distributions with $\ell = (\gamma mbn)^{\frac{1}{2\alpha+2}}$ for some constant γ and $\sigma_i^2 = \frac{\widetilde{c}^2}{\sum_{i=1}^{\ell} i^{2\alpha}}$ for $i = 1, \dots, \ell$, which satisfies condition (3.2). With this choice and replacing ε^2 by $\frac{1}{n}$, we have

$$\begin{aligned} \sum_{i=1}^{\ell} d_i &\geq \sum_{i=1}^{\ell} \frac{1}{\frac{\sum_{i=1}^{\ell} i^{2\alpha}}{\widetilde{c}^2} + \frac{2mbn}{\ell}} \\ &\geq \frac{\ell}{\frac{(\ell+1)^{2\alpha+1}}{\widetilde{c}^2(2\alpha+1)} + \frac{2mbn}{\ell}} \\ &= \frac{\widetilde{c}^2(2\alpha+1)}{2\widetilde{c}^2(2\alpha+1)+1} \ell^{-2\alpha}(1+o(1)) \\ &\sim (mbn)^{-\frac{\alpha}{\alpha+1}}. \end{aligned}$$

3. For the last regime where $b(mn)^{-\frac{1}{2\alpha+1}} \to \infty$, we use the constraint that $d_i \geq \frac{\sigma_i^2 \frac{\varepsilon^2}{m}}{\sigma_i^2 + \frac{\varepsilon^2}{m}}$ and write

$$\sum_{i=1}^{\ell} d_i \geq \sum_{i=1}^{\ell} \frac{\sigma_i^2 \frac{\varepsilon^2}{m}}{\sigma_i^2 + \frac{\varepsilon^2}{m}} = \sum_{i=1}^{\ell} \frac{\sigma_i^2 \frac{1}{mn}}{\sigma_i^2 + \frac{1}{mn}}.$$

Let $\ell = (\gamma mn)^{\frac{1}{2\alpha+1}}$ and $\sigma_i^2 = \frac{\widetilde{c}^2}{\sum_{i=1}^{\ell} i^{2\alpha}}$ satisfying (3.2), and we have

$$\begin{aligned} \sum_{i=1}^{\ell} d_i &\geq \sum_{i=1}^{\ell} \frac{\frac{\widetilde{c}^2}{\sum_{i=1}^{\ell} i^{2\alpha}} \frac{1}{mn}}{\frac{\widetilde{c}^2}{\sum_{i=1}^{\ell} i^{2\alpha}} + \frac{1}{mn}} \\ &\geq \sum_{i=1}^{\ell} \frac{\frac{\ell^{2\alpha+1}}{2\alpha+1} \frac{1}{mn}}{\frac{\ell^{2\alpha+1}}{2\alpha+1} + \frac{1}{mn}} \\ &\sim (mn)^{-\frac{2\alpha}{2\alpha+1}}. \end{aligned}$$

Thus, combining the previous three scenarios, we conclude the lower bound in 3.1.

4. Achievability

In this section, we describe how the lower bound can be achieved through the use of a certain distributed protocol. Unlike for the lower bound, we shall work under the individual constraint on the communication budget, instead of the sum constraint. However, a protocol satisfying the individual constraint automatically satisfies the sum constraint.

4.1. High-level idea

In nonparametric estimation theory, it is known that for the Gaussian sequence model $X_i \sim N\left(\theta, \frac{1}{n}\right)$ for $i = 1, \ldots, \infty$ with $\theta \in \Theta(\alpha, c)$, the optimal scaling of the ℓ_2 risk is $n^{-\frac{2\alpha}{2\alpha+1}}$, and this can be achieved by truncating the sequence at $i = O(n^{\frac{1}{2\alpha+1}})$. That is, the estimator

$$\widehat{\theta} = \begin{cases} X_i & \text{if } i \le n^{\frac{1}{2\alpha+1}} \\ 0 & \text{if } i > n^{\frac{1}{2\alpha+1}} \end{cases}$$

has worst-case risk $\sup_{\theta\in\Theta(\alpha,c)} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}}$. We are going to build on this simple but rate-optimal estimator in our distributed protocol. But before carefully defining and analyzing the protocol, we first give a high-level idea of how it is designed.

In our distributed setting, we have a total budget of mb bits to communicate from the local machines to the central machine, which means that we can transmit $O(mb)$ random variables to a certain degree of precision.

In the first regime where we have $mb \lesssim n^{\frac{1}{2\alpha+1}}$, the communication budget is so small that the total number of bits is smaller than the effective dimension for the noise level $1/n$. In this case, we let each machine transmit information regarding a unique set of $O(b)$ components of θ. Thus, at the central machine, we can decode and obtain information about the first $O(mb)$ components of θ_i. This is equivalent to truncating a centralized Gaussian sequence at $i = O(mb)$, and gives us a convergence rate of $(mb)^{-2\alpha}$.

In the second regime ($b \ll (mn)^{\frac{1}{2\alpha+1}}$ and $mb \gg n^{\frac{1}{2\alpha+1}}$), we have a larger budget at our disposal, and can thus afford to transmit more than one random variable containing information about θ_i. Suppose that for a specific i we quantize and transmit X_{ij} for k different values of j, namely at k different machines. The budget of $O(mb)$ random variables will allow us to acquire information about the first $O(\frac{mb}{k})$ components of θ. When aggregating at the central machine, we have $Z_i \sim N\left(\theta_i, \frac{1}{nk}\right)$ for $i = 1, \ldots, O\left(\frac{mb}{k}\right)$, and no information about θ_i for $i \ge O\left(\frac{mb}{k}\right)$. Now consider the effect of choosing different values of k. In choosing a smaller k, we will be able to estimate more components of θ, but each at a lower accuracy. On the other hand, a larger k leads to fewer components being estimated, but with smaller error. We know from nonparametric estimation theory that the tradeoff is optimized when $(kn)^{\frac{1}{2\alpha+1}} \asymp \frac{mb}{k}$. This gives us the optimal choice $k \asymp (mb)^{\frac{2\alpha+1}{2\alpha+2}} n^{-\frac{1}{2\alpha+2}}$, with risk scaling as $(mbn)^{-\frac{\alpha}{\alpha+1}}$.

In the last regime, we have $b \gtrsim (mn)^{\frac{1}{2\alpha+1}}$. In this case, the number of bits available at each machine is larger than the effective dimension associated with the global noise level $\frac{1}{mn}$. We simply quantize and transmit the first $O(mn)^{\frac{1}{2\alpha+1}}$ of X_{ij} from each machine to the central machine, where we decode and simply average the received random variables.

4.2. Algorithm

First we state a lemma describing and analyzing a simple scalar quantization method.

Lemma 4.1. *Suppose that X is a random variable supported on $[-c, c]$, and that $U \sim \mathrm{Unif}(0, \delta)$ independently, for some constant $\delta > 0$. Let $G(u, \delta) = \{u + i\delta : i = 0, \pm1, \pm2, \ldots\}$ be a grid of points with base point u and skip δ. Define*

$$q(x; u, \delta) = \operatorname*{arg\,min}_{g\in G(u,\delta)} |x - g|.$$

Let $E = q(X; U, \delta) - X$. Then X and E are independent, and $E \sim \mathrm{Unif}(-\frac{\delta}{2}, \frac{\delta}{2})$.

Proof. Let us condition on the event that $X = x$. We have for $\epsilon \in (-\frac{\delta}{2}, \frac{\delta}{2})$

$$\begin{aligned} &\mathbb{P}(E \in (\epsilon, \epsilon + \mathrm{d}\epsilon) \,|\, X = x) \\ &= \mathbb{P}(q(X) \in (x + \epsilon, x + \epsilon + \mathrm{d}\epsilon) \,|\, X = x) \\ &= \mathbb{P}(U \in (x + \epsilon - \delta\lfloor (x + \epsilon)/\delta \rfloor, \\ &\qquad x + \epsilon - \delta\lfloor (x + \epsilon)/\delta \rfloor + \mathrm{d}\epsilon) \,|\, X = x) \\ &= \mathbb{P}(U \in (x + \epsilon - \delta\lfloor (x + \epsilon)/\delta \rfloor, \\ &\qquad x + \epsilon - \delta\lfloor (x + \epsilon)/\delta \rfloor + \mathrm{d}\epsilon)) \\ &= \frac{d\epsilon}{\delta}. \end{aligned}$$

We thus conclude that $E|X \sim \mathrm{Unif}(-\frac{\delta}{2}, \frac{\delta}{2})$, and therefore E and X are independent. □

By this lemma, we know that with a public key for randomness, we can transmit a random variable X supported on $(-c, c)$ using $\log_2 \frac{2c}{\delta}$ bits, so that the central machine receives $X + E$ with $E \sim \mathrm{Unif}(-\frac{\delta}{2}, \frac{\delta}{2})$ and independent to X. We are now ready to describe the algorithm of estimating θ.

Algorithm

1. Input
   - α: order of the Sobolev space.
   - c: radius of the Sobolev space.

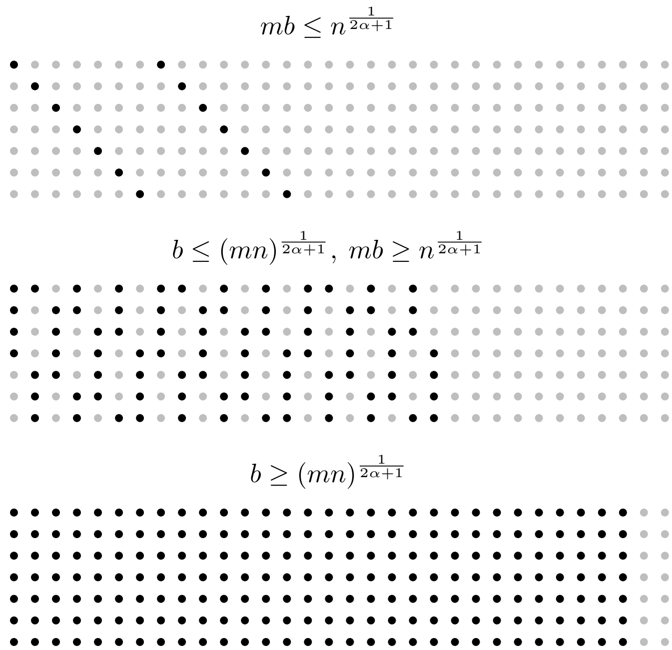

Figure 1. Allocation of communication budget for the three regimes. Each dot represents a random variable X_{ij}. The jth row represents the random variables on the jth machine, and the ith dot in that row is for the random variable $X_{ij} \sim N(\theta_i, \frac{1}{n})$. If a dot is colored black, it means that the random variable is quantized and transmitted to the central machine; otherwise, we don't spend any communication budget on it. In the first regime, each θ_i is only estimated on at most one machine, while in the second regime, it is estimated on multiple but not all machines. In the last regime, we quantize and transmit all random variables associated with θ_i on the m machines, before truncating at some position.

- X_{ij}: independent $N\left(\theta_i, \frac{1}{n}\right)$ r.v. for $i = 1, \dots, \infty$ at machine j for $j = 1, \dots, m$.
- b: number of bits for communication at each machine.

Calculate

- $\delta = \max\left\{(mb)^{-\frac{2\alpha+1}{2}}, n^{-\frac{1}{2}}\right\}$.
- $b_0 = \log_2 \delta$, $\widetilde{b} = \lfloor b/b_0 \rfloor$.
- $k = \left(\lfloor (mb)^{\frac{2\alpha+1}{2\alpha+2}} n^{-\frac{1}{2\alpha+2}} \rfloor \vee 1\right) \wedge m$.

2. At the jth machine (for $j = 1, \dots, m$), let $I_j = \left\{\lceil (ms+j)/k \rceil : s = 0, \dots, \widetilde{b}-1\right\}$.
   (a) Generate a random seed shared with the central machine.
   (b) For $i \in I_j$, generate $U_{ij} \sim \mathrm{Unif}(0, \delta)$ independently based on the seed.
   (c) For $i \in I_j$, Winsorize X_{ij} at $[-c, c]$ and quantize
   $$\widetilde{X}_{ij} = q\left((X_{ij} \wedge c) \vee (-c); U_{ij}, \delta\right).$$
   (d) Transmit the quantized random variables $\left\{\widetilde{X}_{ij} : i \in I_j\right\}$ to the central machine using $\lfloor b/b_0 \rfloor b_0 \leq b$ bits.

3. At the central machine, decode the messages and construct the estimator
$$\widehat{\theta}_i = \begin{cases} \frac{1}{k} \sum_{j:\, i \in I_j} \widetilde{X}_{ij} & \text{if } i \leq \lfloor m\widetilde{b}/k \rfloor \wedge (mn)^{\frac{1}{2\alpha+1}} \\ 0 & \text{otherwise} \end{cases}.$$

A graphical illustration of the algorithm is shown in Figure 1. We must also note that while the algorithm is rate optimal, it is not adaptive, in the sense that it requires knowledge of the parameter α.

4.3. Analysis

We now analyze the statistical risk associated with the algorithm described in the previous section. Suppose that $\theta \in \Theta(\alpha, c)$. First notice that the Winsorization in Step 2(c) makes X_{ij} bounded prior to quantization and it only decreases the risk. Write $i^* = \lfloor m\widetilde{b}/k \rfloor \wedge (mn)^{\frac{1}{2\alpha+1}}$. The risk of the final estimator satisfies

$$\mathbb{E}_\theta\left[\|\widehat{\theta} - \theta\|^2\right] \sum_{i=1}^{i^*} \mathbb{E}_\theta\left[\left(\frac{1}{k} \sum_{j:\, i \in I_j} \widetilde{X}_{ij} - \theta_i\right)^2\right] + \sum_{i=i^*+1}^{\infty} \theta_i^2$$

where

$$
\begin{aligned}
&\sum_{i=1}^{i^*} \mathbb{E}_\theta \left[\left(\frac{1}{k} \sum_{j:\, i \in I_j} \widetilde{X}_{ij} - \theta_i \right)^2 \right] \\
&= \sum_{i=1}^{i^*} \mathbb{E}_\theta \left[\left(\frac{1}{k} \sum_{j:\, i \in I_j} \left(X_{ij} + E_{ij} \right) - \theta_i \right)^2 \right] \\
&= \sum_{i=1}^{i^*} \mathbb{E}_\theta \left[\left(\frac{1}{k} \sum_{j:\, i \in I_j} X_{ij} - \theta_i \right)^2 \right] + \mathbb{E} \left[\left(\frac{1}{k} \sum_{j:\, i \in I_j} E_{ij} \right)^2 \right] \\
&\le \frac{\lfloor m\widetilde{b}/k \rfloor \wedge (mn)^{\frac{1}{2\alpha+1}}}{nk} + \frac{\left(\lfloor m\widetilde{b}/k \rfloor \wedge (mn)^{\frac{1}{2\alpha+1}} \right) \delta^2}{3k}
\end{aligned}
$$

where E_{ij} denotes the uniform error introduced by quantizing X_{ij} and we have used the fact that they are mutually independent and independent to X_{ij}. Also recall the definitions of $\delta, k, \widetilde{b}$ as appearing in the algorithm. Therefore, we have

$$
\begin{aligned}
&\mathbb{E}_\theta \left[\|\widehat{\theta} - \theta\|^2 \right] \\
&\le \frac{\lfloor m\widetilde{b}/k \rfloor \wedge (mn)^{\frac{1}{2\alpha+1}}}{nk} + \frac{\left(\lfloor m\widetilde{b}/k \rfloor \wedge (mn)^{\frac{1}{2\alpha+1}} \right) \delta^2}{3k} \\
&\quad + \frac{c^2}{\left(\lfloor m\widetilde{b}/k \rfloor \wedge (mn)^{\frac{1}{2\alpha+1}} \right)^{2\alpha}}
\end{aligned}
$$

Now we analyze the risk for the three regimes respectively.

In the first regime where $mb \le n^{\frac{1}{2\alpha+1}}$, we have

$$
\begin{aligned}
k &= \left(\lfloor (mb)^{\frac{2\alpha+1}{2\alpha+2}} n^{-\frac{1}{2\alpha+2}} \rfloor \vee 1 \right) \wedge m = 1 \\
\delta &= \max\left\{ (mb)^{-\frac{2\alpha+1}{2}}, n^{-\frac{1}{2}} \right\} = (mb)^{-\frac{2\alpha+1}{2}},
\end{aligned}
$$

and consequently

$$
\begin{aligned}
\mathbb{E}_\theta \left[\|\widehat{\theta} - \theta\|^2 \right] &\le \frac{m\widetilde{b}}{n} + \frac{m\widetilde{b}}{3(mb)^{2\alpha+1}} + \frac{c^2}{(m\widetilde{b})^{2\alpha}} \\
&= O\left((mb)^{-2\alpha} \log(mb) \right).
\end{aligned}
$$

In the second regime where $b \le (mn)^{\frac{1}{2\alpha+1}}$ and $mb \ge n^{\frac{1}{2\alpha+1}}$, we have

$$
\begin{aligned}
k &= \left(\lfloor (mb)^{\frac{2\alpha+1}{2\alpha+2}} n^{-\frac{1}{2\alpha+2}} \rfloor \vee 1 \right) \wedge m = \lfloor (mb)^{\frac{2\alpha+1}{2\alpha+2}} n^{-\frac{1}{2\alpha+2}} \rfloor, \\
\delta &= \max\left\{ (mb)^{-\frac{2\alpha+1}{2}}, n^{-\frac{1}{2}} \right\} = n^{-\frac{1}{2}},
\end{aligned}
$$

and it follows that

$$
\begin{aligned}
\mathbb{E}_\theta \left[\|\widehat{\theta} - \theta\|^2 \right] &\le \frac{4\lfloor m\widetilde{b}/k \rfloor}{3nk} + \frac{c^2}{\lfloor m\widetilde{b}/k \rfloor^{2\alpha}} \\
&= O\left((mbn)^{-\frac{\alpha}{\alpha+1}} (\log n)^{2\alpha} \right).
\end{aligned}
$$

In the last regime where $b \ge (mn)^{\frac{1}{2\alpha+1}}$, we have

$$
\begin{aligned}
k &= \left(\lfloor (mb)^{\frac{2\alpha+1}{2\alpha+2}} n^{-\frac{1}{2\alpha+2}} \rfloor \vee 1 \right) \wedge m = m, \\
\delta &= \max\left\{ (mb)^{-\frac{2\alpha+1}{2}}, n^{-\frac{1}{2}} \right\} = n^{-\frac{1}{2}},
\end{aligned}
$$

and then

$$
\begin{aligned}
&\mathbb{E}_\theta \left[\|\widehat{\theta} - \theta\|^2 \right] \\
&\le \frac{(mn)^{\frac{1}{2\alpha+1}}}{mn} + \frac{(mn)^{\frac{1}{2\alpha+1}}}{3mn} + \frac{c^2}{\widetilde{b}^{2\alpha}} \\
&= \begin{cases} O\left((mn)^{-\frac{2\alpha}{2\alpha+1}} \right) & \text{if } b \ge (mn)^{\frac{1}{2\alpha+1}} \log n \\ O\left((mn)^{-\frac{2\alpha}{2\alpha+1}} (\log n)^{2\alpha} \right) & \text{otherwise} \end{cases}.
\end{aligned}
$$

5. Future directions

One interesting direction for future work is to study adaptivity in distributed estimation. An adaptive protocol $(\Pi, \widehat{\theta})$ satisfies

$$
\liminf_{n\to\infty} \frac{\sup_{\theta \in \Theta(\alpha,c)} \mathbb{E}\left[\|\widehat{\theta}(\Pi) - \theta\|^2 \right]}{\inf_{(\Pi,\check{\theta}) \in \Gamma(m,b)} \sup_{\theta \in \Theta(\alpha,c)} \mathbb{E}\left[\|\check{\theta}(\Pi) - \theta\|^2 \right]} < \infty,
$$

for (almost) all α and c. That is, the protocol should be minimax optimal for all θ without prior knowledge of the parameter space in which θ resides. While we had conjectured that this may not be possible with only one round of communication, Szabo & van Zanten (2018) recently developed an adaptive estimator using a modification of Lepski's method.

A second interesting direction for future work is distributed estimation of other functionals. For instance, one might study the sum of squares (or ℓ_2 norm) of the mean of a normal random vector. It would be of interest to understand the minimax risk of the norm of the mean in a distributed setting, and to develop optimal distributed protocols for this functional.

Finally, other nonparametric problems should be considered in a distributed estimation setting. For example, it will be interesting to study nonparametric estimation of functions with varying smoothness (e.g., over Besov bodies), and with shape constraints such as monotonicity and convexity.

Acknowledgment

Research supported in part by ONR grant N00014-12-1-0762 and NSF grant DMS-1513594.

References

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*, 2015.

Berger, T., Zhang, Z., and Viswanathan, H. The CEO problem. *IEEE Trans. Inform. Theory*, 42(3):887–902, 1996.

Blanchard, G. and Mücke, N. Parallelizing spectral algorithms for kernel learning. *arXiv preprint arXiv:1610.07487*, 2016.

Braverman, M., Garg, A., Ma, T., Nguyen, H. L., and Woodruff, D. P. Communication lower bounds for statistical estimation problems via a distributed data processing inequality. In *Proceedings of the Forty-eighth Annual ACM Symposium on Theory of Computing*, STOC '16, pp. 1011–1020, 2016.

Brown, L. D. and Low, M. G. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996.

Chang, X., Lin, S.-B., and Zhou, D.-X. Distributed semi-supervised learning with kernel ridge regression. *Journal of Machine Learning Research*, 18(46):1–22, 2017.

Diakonikolas, I., Grigorescu, E., Li, J., Natarajan, A., Onak, K., and Schmidt, L. Communication-efficient distributed learning of discrete distributions. In *Advances in Neural Information Processing Systems*, pp. 6394–6404, 2017.

Fan, J., Wang, D., Wang, K., and Zhu, Z. Distributed estimation of principal eigenspaces. *arXiv preprint arXiv:1702.06488*, 2017.

Johnstone, I. M. Gaussian estimation: Sequence and wavelet models. Unpublished manuscript, 2017.

Lee, J. D., Liu, Q., Sun, Y., and Taylor, J. E. Communication-efficient sparse regression. *Journal of Machine Learning Research*, 18(5):1–30, 2017.

Nussbaum, M. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. of Statist.*, pp. 2399–2430, 1996.

Shamir, O. Fundamental limits of online and distributed algorithms for statistical learning and estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pp. 163–171, 2014.

Shang, Z. and Cheng, G. Computational limits of a distributed algorithm for smoothing spline. *The Journal of Machine Learning Research*, 18(1):3809–3845, 2017.

Szabo, B. and van Zanten, H. Adaptive distributed methods under communication constraints. *arXiv preprint arXiv:1804.00864*, 2018.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Series in Statistics, 1st edition, 2008.

Viswanathan, H. and Berger, T. The quadratic Gaussian ceo problem. *IEEE Transactions on Information Theory*, 43 (5):1549–1559, Sep 1997.

Wang, J., Chen, J., and Wu, X. On the sum rate of Gaussian multiterminal source coding: New proofs and results. *IEEE Transactions on Information Theory*, 56(8):3946–3960, 2010.

Zhang, Y., Duchi, J., Jordan, M. I., and Wainwright, M. J. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems*, pp. 2328–2336, 2013a.

Zhang, Y., Duchi, J., and Wainwright, M. Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pp. 592–617, 2013b.

Zhu, Y. and Lafferty, J. Quantized minimax estimation over Sobolev ellipsoids. *Information and Inference*, 2017.