Submitted to the Annals of Statistics

OPTIMAL FEDERATED LEARNING FOR NONPARAMETRIC REGRESSION WITH HETEROGENEOUS DISTRIBUTED DIFFERENTIAL PRIVACY CONSTRAINTS

BY T. TONY CAI[1,a], ABHINAV CHAKRABORTY[1,b] AND LASSE VUURSTEEN[1,c]

[1]*Department of Statistics and Data Science, University of Pennsylvania,* [a]*tcai@wharton.upenn.edu;* [b]*abch@wharton.upenn.edu;* [c]*lassev@wharton.upenn.edu*

This paper studies federated learning for nonparametric regression in the context of distributed samples across different servers, each adhering to distinct differential privacy constraints. The setting we consider is heterogeneous, encompassing both varying sample sizes and differential privacy constraints across servers. Within this framework, both global and pointwise estimation are considered, and optimal rates of convergence over the Besov spaces are established.

Distributed privacy-preserving estimators are proposed and their risk properties are investigated. Matching minimax lower bounds, up to a logarithmic factor, are established for both global and pointwise estimation. Together, these findings shed light on the tradeoff between statistical accuracy and privacy preservation. In particular, we characterize the compromise not only in terms of the privacy budget but also concerning the loss incurred by distributing data within the privacy framework as a whole. This insight captures the folklore wisdom that it is easier to retain privacy in larger samples, and explores the differences between pointwise and global estimation under distributed privacy constraints.

1. Introduction. In today's data-driven world, the proliferation of personal data and technological advancements has made the protection of privacy a matter of paramount importance. Developing statistical methods with privacy guarantees is becoming increasingly important. Differential privacy (DP), one of the most widely adopted privacy frameworks, ensures that statistical analysis results do not divulge any sensitive information about the input data. DP was introduced in the seminal work by Dwork et al. [30]. Since its inception, DP has garnered significant academic attention [5, 32, 33] and notable applications within industry leaders, including Google [36], Microsoft [25], and Apple [67]. It has also been embraced by governmental entities like the US Census Bureau [60].

A common setting in many real-life applications is the distributed nature of data collection and analysis. For example, medical data is spread across various hospitals in healthcare, customer data is stored in different branches or databases in financial institutions and various modern technologies rely on federated learning from networks of users, see, for example, [9, 42, 50, 54, 58]. DP has found applications in many of these domains relating to, for example, healthcare, finance, tech and social sciences, where preserving individuals' data privacy is of utmost concern. In such scenarios, it is vital to develop efficient estimation techniques that respect privacy constraints while harnessing the collective potential of distributed data.

Federated learning is a machine learning paradigm designed to address the challenges of data governance and privacy. It enables organizations or groups, whether from diverse geographic regions or within the same organization, to collaboratively train and improve a shared global statistical model without external sharing of raw data. The learning process

MSC2020 subject classifications: Primary 62G08; secondary 62G20.

Keywords and phrases: Besov Spaces, Distributed Computation, Differential Privacy, Minimax Risk, Nonparametric Regression, Function Estimation.

occurs locally at each participating entity, which we shall refer to as *servers*. The servers exchange only characteristics of their data, such as parameter estimates or gradients, in a way that preserves privacy of the individuals comprising their data. Federated learning facilitates secure collaboration across industries like retail, manufacturing, healthcare, and financial services, allowing them to harness the power of data analysis while upholding data privacy and security.

Rigorous study of theoretical performance in federated learning settings with communication constraints has been conducted in, for example, bandwidth constraint parametric problems [2, 7, 10, 20, 29, 41, 65] and bandwidth constraint nonparametric estimation and testing [19, 63, 64, 66, 73]. Under DP constraints, theoretical performance in federated learning settings have been studied for various parametric estimation and testing problems [3, 53, 55, 57]. Federated learning settings where each server's sample consists of one individual observation (referred to as *local* differential privacy settings) have been studied in many-normal-means model, discrete distributions and parametric models [1, 6, 27, 28, 71] and nonparametric density estimation [12, 51, 61].

This paper investigates the statistical optimality of federated learning under DP constraints in the context of nonparametric regression. We consider a setting where data is distributed among different entities, such as hospitals, that are concerned about sharing their data with other entities due to privacy concerns for their patients. Each entity communicates a transcript that fulfills a distinct DP requirement, and we assume a setting with m servers, each with n_j observations where $j = 1, \ldots, m$.

Our goals are two-fold: firstly, to establish optimal rates of convergence, measured in terms of minimax risk, for estimating the nonparametric regression function while adhering to DP constraints; secondly, to construct a rate-optimal estimator under these DP constraints. We explore both global and pointwise estimation, aiming to provide quantifiable measures of the trade-off between accuracy and privacy preservation. These convergence rates offer insights into the best achievable estimation performance in distributed settings while ensuring privacy. Recognizing that global estimation exhibits different characteristics compared to its pointwise counterpart in the classical setting [15], we investigate how DP constraints impact global and pointwise estimation risks differently.

1.1. *Problem formulation.* We will begin by formally introducing the general framework of distributed estimation under privacy constraints. Consider a family of probability measures $\{P_f\}_{f\in\mathcal{F}}$ on the measurable space $(\mathcal{Z}, \mathscr{Z})$, parameterized by $f \in \mathcal{F}$. We consider a setting where $N = \sum_{j=1}^{m} n_j$ i.i.d. observations are drawn from a distribution P_f and distributed across m servers. Each server $j = 1, \ldots, m$ holds n_j observations.

Let us denote by $Z^{(j)} = \{Z_i^{(j)}\}_{i=1}^{n_j}$ the n_j realizations from P_f on the j-th server. For each server, we output a (randomized) transcript $T^{(j)}$ based on $Z^{(j)}$, where the law of the transcript is given by a distribution conditional on $Z^{(j)}$, $\mathbb{P}(\cdot|Z^{(j)})$ on a measurable space $(\mathcal{T}, \mathscr{T})$. The transcript $T^{(j)}$ has to satisfy a $(\varepsilon_j, \delta_j)$-differential privacy constraint, which is defined as follows.

DEFINITION 1.1. The transcript $T^{(j)}$ is $(\varepsilon_j, \delta_j)$-differentially private if for all $A \in \mathscr{T}$ and $z, z' \in \mathcal{Z}^{n_j}$ differing in one individual datum, it holds that

$$\mathbb{P}\left(T^{(j)} \in A | Z^{(j)} = z\right) \leq e^{\varepsilon_j} \mathbb{P}\left(T^{(j)} \in A | Z^{(j)} = z'\right) + \delta_j.$$

In the above definition, "differing in one datum" should be understood in terms of being Hamming distance "neighbors." To clarify, the local datasets $Z^{(j)}$ and $\tilde{Z}^{(j)}$ are deemed *neighboring* if their Hamming distance is at most 1. The Hamming distance is calculated

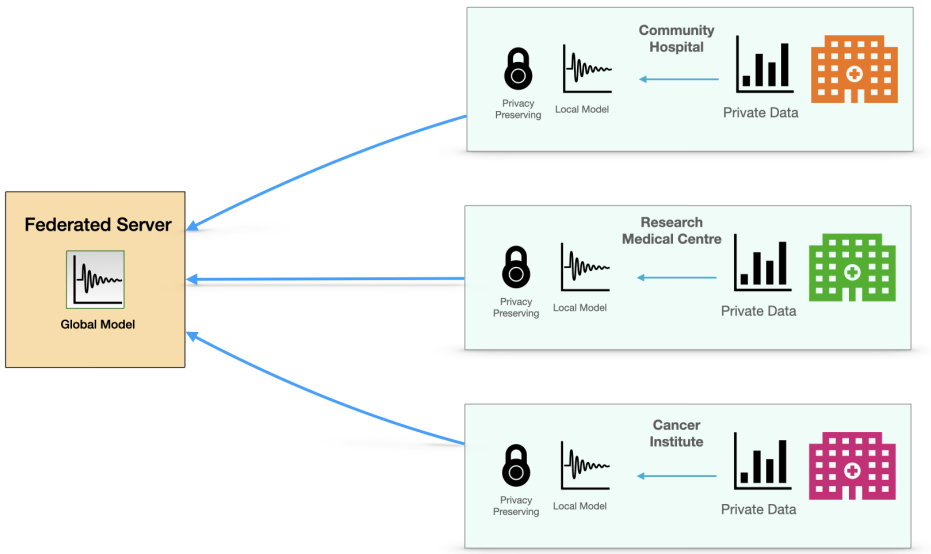

FIG 1. *An illustration of the federated learning framework.*

over $\mathcal{Z}^{n_j} \times \mathcal{Z}^{n_j}$. In other words, $\tilde{Z}^{(j)}$ can be derived from $Z^{(j)}$ by modifying at most one of the observations $Z_1^{(j)}, \ldots, Z_{n_j}^{(j)}$. The smaller the value of ϵ_j and δ_j, the more stringent the privacy constraint. We shall consider $\varepsilon_j \leq C_\varepsilon$ for $j = 1, \ldots, m$ for a fixed but arbitrarily constant $C_\varepsilon > 0$, where the choice of the constant does not affect the rates in the results derived.

We focus on distributed protocols that apply to situations in which sensitive data is held by multiple parties, each generating an output while ensuring differential privacy. Within such a distributed protocol, the transcripts from each server only depend on its local data, and no information is exchanged between the servers. This occurs, for example, when multiple trials concerning the same population are conducted, but each location (e.g. hospital) does not wish to pool their original data because of privacy concerns.

Each server transmits its transcript to the central server. The central server, utilizing all transcripts $T := (T^{(1)}, \ldots, T^{(m)})$, computes an estimator $\hat{f} : \mathcal{T}^m \to \mathcal{F}$. We refer to the pair $(\hat{f}, \{(\mathbb{P}(\cdot|z))_{z\in\mathcal{Z}}\}_{j=1}^m)$ as a *distributed estimation protocol*, which we shall sometimes just denote as $\hat{f}$. We denote the vector of the differing DP levels by $(\boldsymbol{\varepsilon}, \boldsymbol{\delta}) = \{(\varepsilon_j, \delta_j)\}_{j=1}^m$ and denote the class of *distributed estimation protocols*, i.e. $\left(\hat{f}, \{(\mathbb{P}(\cdot|z))_{z\in\mathcal{Z}}\}_{j=1}^m\right)$ satisfying Definition 1.1, with $\mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$. We let $\mathbb{P}_f$ denote the joint law of transcripts and the $N = \sum_{j=1}^m n_j$ i.i.d. observations generated from P_f. We let $\mathbb{E}_f$ denote the expectation corresponding to $\mathbb{P}_f$.

In the context of nonparametric regression, the distributed estimation problem arises when data is distributed among multiple servers. Specifically, for each server j, the data $Z^{(j)} = \{(Y_i^{(j)}, X_i^{(j)})\}_{i=1}^{n_j}$ consists of n_j pairs of observations $(Y_i^{(j)}, X_i^{(j)})$. Here, $X_i^{(j)}$ represents the input variable, and $Y_i^{(j)}$ represents the corresponding response variable.

We assume that under P_f, $X_i^{(j)}$ and $Y_i^{(j)}$ are generated by the relationship

$$Y_i^{(j)} = f\left(X_i^{(j)}\right) + \xi_i^{(j)}, \quad X_i^{(j)} \sim U[0,1]. \tag{1}$$

Here, f is an unknown function representing the underlying relationship between the input and response variables. The term $\xi_i^{(j)}$ represents random noise, assumed to be independent

of $X_i^{(j)}$, and follows a Gaussian distribution with mean 0 and known variance σ^2. Without loss of generality, we shall assume $\sigma = 1$ throughout the paper.

The aim is to estimate the function f based on the distributed data. The difficulty of this estimation task arises from both the distributed nature of the data and privacy constraints that limit the sharing of information between servers. As in the conventional decision-theoretical framework, for global estimation, the estimation accuracy of a distributed estimator $\hat{f} \equiv \hat{f}(T)$ is measured by the integrated mean squared error (IMSE), $\mathbb{E}_f \|\hat{f} - f\|_2^2$, where the expectation is taken over the randomness in both the data (under P_f) and construction of the transcripts. As in the conventional framework, a quantity of particular interest in federated learning is the *global minimax risk* for the distributed private protocols over function class $\mathcal{F}$,

$$\inf_{\hat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{f \in \mathcal{F}} \mathbb{E}_f \|\hat{f} - f\|_2^2. \tag{2}$$

The global risk characterizes the difficulty of the distributed learning problem over the function class $\mathcal{F}$ when trying to infer the entire function underlying the data whilst adhering to the heterogeneous privacy constraints.

Besides global estimation, it is also of interest to estimate f at a fixed point $x_0 \in (0,1)$ under the mean squared error (MSE). The *pointwise minimax risk* in that case is given by

$$\inf_{\hat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{f \in \mathcal{F}} \mathbb{E}_f (\hat{f}(x_0) - f(x_0))^2, \text{ for } x_0 \in (0,1), \tag{3}$$

where $\hat{f}(x_0)$ denotes the estimated function value at $x_0 \in (0,1)$. The pointwise risk is particularly useful in understanding the behavior of estimators at specific points within the domain, which can be crucial in applications where certain regions are of particular interest or have higher consequences associated with estimation errors. It is known that in the classical setting, without privacy constraints, there are important differences between the global risk and pointwise risk in terms of performance. See, for example, [14].

We consider estimating f over the Besov ball of radius $R > 0$, denoted as $\mathcal{B}_{p,q}^{\alpha,R}[0,1]$ (defined in (11)), where $p \geq 2$, $q \geq 1$ and $\alpha - 1/p > 1/2$. This Besov space offers a suitable framework for analyzing functions with specific smoothness characteristics. Operating within this space allows us to encompass diverse function classes, accommodating varying levels of smoothness and complexity.

1.2. *Main contribution.* We quantify the cost of differential privacy for both the minimax global risk given by (2) and the pointwise risk as in (3). To achieve this, we introduce two differentially private estimators – one for global and one for pointwise estimation. We obtain matching minimax lower bounds, up to logarithmic factors, thereby establishing their optimality.

Our analysis reveals interesting phenomena, that go unobserved in settings where servers are assumed to have homogeneous privacy budgets. Further discussion on these broader findings is deferred to Section 2. The results for the homogeneous case, where privacy budgets are equal among servers ($\varepsilon_j = \varepsilon$, $\delta_j = \delta$, and $n_j = n$ for $j = 1, \ldots, m$), yield novel insights. In this case, our results yield the following minimax rate for global estimation,

$$\inf_{\hat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{f \in \mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \|\hat{f} - f\|_2^2 \asymp \min\left\{ M_{m,n} \cdot \left((mn^2\varepsilon^2)^{-\frac{2\alpha}{2\alpha+2}} + (mn)^{-\frac{2\alpha}{2\alpha+1}} \right), 1 \right\}, \tag{4}$$

where $M_{m,n} \geq 1$ is a sequence at most of the order $\log(mn) \cdot \log(1/\delta)$. The rate $(mn)^{-\frac{2\alpha}{2\alpha+1}}$ is the minimax rate for the global risk in the unconstrained problem, and is attained whenever $n\varepsilon^2 \gtrsim (mn)^{\frac{1}{2\alpha+1}}$. The unconstrained optimal rate is attainable (up to a possibly poly-logarithmic factor) under DP constraints in the homogeneous setting as long as

$n\varepsilon^2 \gtrsim (mn)^{\frac{1}{2\alpha+1}}$. Whenever $n\varepsilon^2 \ll (mn)^{\frac{1}{2\alpha+1}}$, the first term dominates and the minimax rate becomes $(mn^2\varepsilon^2)^{-\frac{2\alpha}{2\alpha+2}}$. As expected in this regime, a smaller ε, which indicates a stronger privacy guarantee, results in an larger minimax estimation error. Whenever $\varepsilon \ll (\sqrt{mn})^{-1}$, consistent estimation ceases to be possible altogether.

This result recover the known minimax rates under local DP constraints (i.e. $n = 1$) that were derived for the problem of nonparametric density estimation for the L_2-risk in [12] and the squared Hellinger in [61], up to logarithmic factor differences. When $n > 1$, the different powers with which n and m appear in the minimax rate reveal an important difference between the general distributed setting and local DP; if one distributes $N = mn$ observations across m machines, the task becomes more challenging as the N observations are spread over a greater number of machines, rather than having a large number of observations on a smaller number of machines. This phenomenon has an intuitive explanation; it is easier to retain privacy in larger samples, as each individual's data will have only a small influence on the aggregate statistics of interest.

For pointwise estimation, we establish the minimax rate in the homogeneous setting;

(5)
$$\inf_{\hat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{f\in\mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f|\hat{f}(x_0) - f(x_0)|^2 \asymp \min\left\{M_{m,n}\cdot\left(\left(mn^2\varepsilon^2\right)^{-\frac{2\nu}{2\nu+2}} + (mn)^{-\frac{2\nu}{2\nu+1}}\right), 1\right\},$$

where $M_{m,n} \geq 1$ is a sequence at most of the order $\log(mn)$. The rate reveals similar phenomena as the one for the global risk above, where for $n = 1$ we recover the known minimax rate for the problem of nonparametric density estimation for the pointwise risk under local DP constraints as studied in [51]. An important difference is the quantity $\nu = \alpha - 1/p$ appearing in the exponent instead of α. This implies that privacy constraints impact pointwise estimation differently than global estimation, with the Besov parameter p influencing both the relative privacy cost and the distribution of the $N = mn$ observations, as discussed further in Section 2.1.

Our findings have substantial implications for the development of federated learning algorithms that balance distributed privacy with accuracy. A clear understanding of the optimal convergence rate under distributed privacy constraints allows the design of algorithms that strike the right balance between accuracy and privacy trade-offs. This study contributes significantly to the growing knowledge on distributed settings for privacy-preserving machine learning, offering valuable insights for future research in this domain.

1.3. *Related Work.* The nonparametric regression setting considered in this work bears relationships with that of nonparametric density estimation as studied in the privacy setting for global risk [12, 28, 61] and pointwise risk [51]. The aforementioned papers consider the setting of local DP, in which the privacy protection is applied at the level of individual data entries or observations. This corresponds to the case wherein $n_j = 1$ for $j = 1, \ldots, m$ in our setting.

Distributed DP as considered in this paper, where DP applies at the level of the local sample consisting of multiple observations, has been studied for the homogeneous estimation setting for discrete distributions [3, 55] and parametric mean estimation [53, 57]. In the paper [21], the authors consider discrete distribution testing in a two server setting ($m = 2$) with differing DP constraints.

Settings in which the full data is assumed to be on a single server (i.e. $m = 1$), where a single privacy constraint applies to all the observations, have also been studied for various parametric high-dimensional problems [8, 17, 35, 47, 48, 56, 62]. The problem of mean estimation with a single server having heterogeneous privacy constraints for each individual observation have been studied in [22, 37].

1.4. *Organization of the paper.* The rest of the paper is organized as follows. We conclude this section with notation, definitions, and assumptions. Section 2 summarizes and discusses the minimax optimal convergence rates for global and pointwise risks under privacy constraints. In Section 3, we present distributed estimation procedures that achieve optimal global and pointwise risk while adhering to distributed privacy constraints, along with an upper bound on their statistical performance. The matching minimax lower bounds for the global and pointwise risks are derived in Section 4. The proofs of the main results are given in the Supplementary Material [16].

1.5. *Notation, definitions and assumptions.* Throughout the article, we shall write $N := \sum_{j=1}^m n_j$ and consider asymptotics in m, the n_j's and the privacy budget $(\boldsymbol{\epsilon}, \boldsymbol{\delta}) := \{\varepsilon_j, \delta_j\}_{j=1}^m$, where we assume that $N \to \infty$. For two positive sequences a_k, b_k we write $a_k \lesssim b_k$ if the inequality $a_k \leq C b_k$ holds for some universal positive constant C. Similarly, we write $a_k \asymp b_k$ if $a_k \lesssim b_k$ and $b_k \lesssim a_k$ hold simultaneously and let $a_k \ll b_k$ denote that $a_k/b_k = o(1)$.

We use the notations $a \vee b$ and $a \wedge b$ for the maximum and minimum, respectively, between a and b. For $k \in \mathbb{N}$, $[k]$ shall denote the set $\{1, \ldots, k\}$. Throughout the paper c and C denote universal constants whose value can differ from line to line. The Euclidean norm of a vector $v \in \mathbb{R}^d$ is denoted by $\|v\|_2$. For a matrix $M \in \mathbb{R}^{d \times d}$, the norm $M \mapsto \|M\|$ is the spectral norm and $\mathrm{Tr}(M)$ is its trace. Furthermore, we let I_d denote the $d \times d$ identity matrix.

Throughout this paper, we shall let $\nu := \alpha - 1/p > 1/2$, which is a required assumption for estimation in Besov spaces (see e.g. [44]). We let $\mathcal{B}_{p,q}^{\alpha,R}$ denote the closed Besov ball of radius R, i.e. $\{f \in \mathcal{B}_{p,q}^{\alpha}[0,1] : \|f\|_{\mathcal{B}_{p,q}^{\alpha}} \leq R\}$, where $R > 0$ is taken to be a constant.

For random variables U and V with probability measures P and Q defined on the same measurable space, we let $D_{\mathrm{TV}}(U, V)$ denote the total variation norm between P and Q, i.e. $\|P - Q\|_{\mathrm{TV}}$. Whenever $P \ll Q$, we write $D_{\mathrm{KL}}(U, V)$ for the Kullback-Leibler divergence between P and Q: $D_{\mathrm{KL}}(P; Q) = \int \log \frac{dP}{dQ} dP$. Our lower bound results hold for transcripts taking values in standard Borel measure spaces. Different measure spaces or larger sigma-algebras can be considered (which only make the privacy constraint more stringent, see e.g. [70]) as long as the quantities in the proofs are appropriately measurable.

2. Minimax optimal rates of convergence.

In this section, we present our primary findings regarding the minimax rate of convergence under DP constraints. Our results address both the global and pointwise risks.

For the global risk, the minimax rates are encapsulated in the upper bound of Theorem 3.2 and the lower bound of Theorem 4.1, derived in Sections 3.2 and 4.1. Similarly, for the pointwise risk, our findings are summarized in Theorems 3.4 and 4.4, in the form of an upper bound and lower bound respectively, in Sections 3.3 and 4.2. Together, these theorems are summarized by the following result.

THEOREM 2.1. *For $\gamma > 0$, let $D > 0$ be the number solving the equation*

$$D^{2\gamma+2} = \sum_{j=1}^{m} \left(n_j^2 \varepsilon_j^2\right) \wedge \left(n_j D\right). \tag{6}$$

Taking $\gamma = \alpha$, the minimax rate for the global risk is given by

$$\inf_{\hat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{f \in \mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \|\hat{f} - f\|_2^2 \asymp \left(M_N D^{-2\alpha} \wedge 1\right),$$

whenever for all $j = 1, \ldots, m$ we have $\delta_j \lesssim (n_j^{1/2} \varepsilon_j^2 (D \vee 1)^{-1})^{1+\kappa}$ for some $\kappa > 0$ and where $M_N \geq 1$ is a sequence of the order at most $\log(N) \log(1/\min_{j \in [m]} \delta_j)$.

For $\gamma = \nu$, the minimax rate for the pointwise risk is given by

$$\inf_{\hat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{f\in\mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \left|\hat{f}(x_0) - f(x_0)\right|^2 \asymp \left(M_N D^{-2\nu} \wedge 1\right),$$

whenever $\sum_j n_j\delta_j \to 0$, for a sequence $M_N \geq 1$ of the order at most $\log(N)$.

We briefly comment on the derived result. First, we note that a unique positive solution to (14) always exists. To see this, note that the exponent $2\gamma + 2 > 2$ implies that the left-hand side is smaller than the right-hand side for $D > 0$ small enough, whilst the right-hand side grows linearly for small enough $D > 0$. Furthermore, the right-hand side increases sublinearly in D, whilst the left-hand side increases superlinearly (strictly so).

When the privacy budget is large enough (e.g. $\varepsilon_j = \infty$ for $j = 1, \ldots, m$), D can be seen to correspond with the 'effective resolution level' of the estimation problem. That is, D would be proportional to the number of wavelet coefficients needed to obtain a wavelet estimator that attains the optimal estimation rate, see for example [26]. For $\alpha > 0$ smooth functions in a Besov space, the optimal resolution level of a wavelet estimator would correspond to $(1 + 2\alpha)^{-1}\lceil \log_2 N \rceil$ for the global risk. However, under privacy constraints, the effective resolution level changes to $(2+2\alpha)^{-1}\lceil \log_2 D \rceil$, which can be substantially different from the case without privacy constraints. We present several specific cases of Theorem 2.1 through corollaries that encapsulate its various implications, as discussed in Sections 2.1 and 2.2.

The upper bounds for both types of risk (as given in Theorems 3.2 and 3.4) are derived by constructing two estimators. One is proven optimal for global risk, while the other is optimal for pointwise risk. The construction of these estimators is detailed in Section 3. Notably, the optimal estimators for each risk type take distinct forms and employ different privacy mechanisms.

Both of these lower bounds require a different technique. For the global risk, the lower bounding technique is reminiscent of the score attack of [17, 18], which is a generalization of the tracing adversary argument of [11, 34]. We describe the technique in detail in Section 4.1. In case of the pointwise risk, we employ a coupling argument akin to [4, 49] in conjunction with Le Cam's two point method (see e.g. [52, 72]). The technique for the pointwise risk lower bound is described in Section 4.2. Whilst the techniques differ, a similarity is that they both account for the differences in the required levels of privacy between the servers, with the quantity $D > 0$ in (14) being the outcome of balancing a bias-variance trade-off, where the variance for each of the servers is either dominated by the (local) noise in the data itself or by the privacy requirement of the server.

2.1. *The homogeneous setting.* Let us start by studying the case where all machines have both an equal amount of observations, as well as privacy budgets. The following result describes the global risk behavior under DP constraints when the servers are homogeneous in both the number of observations, as well as the privacy constraints they adhere to.

COROLLARY 1. *Suppose that $n_j = n$, $\varepsilon_j = \varepsilon$, $\delta_j = \delta$ for $j = 1, \ldots, m$ and assume that $\delta \lesssim (\varepsilon^2/\sqrt{m})^{1+\kappa}$ for some $\kappa > 0$. Then, the global minimax risk over $\mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$ satisfies (4).*

Whenever $n\varepsilon^2 \ll (mn)^{\frac{1}{2\alpha+1}}$, we have that

$$\inf_{\hat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{f\in\mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f\|\hat{f} - f\|_2^2 \asymp M_{m,n}\,(mn)^{-\frac{2\alpha}{2\alpha+1}} \left(m^{\frac{1}{2\alpha+1}} n^{-\frac{2\alpha}{2\alpha+1}} \varepsilon^{-2}\right)^{\frac{2\alpha}{2\alpha+2}},$$

which indicates that the minimax estimation error becomes larger than the unconstrained minimax rate ($(mn)^{-\frac{2\alpha}{2\alpha+1}}$) by a factor of $(m^{\frac{1}{(2\alpha+1)}} n^{-\frac{2\alpha}{2\alpha+1}} \varepsilon^{-2})^{\frac{2\alpha}{2\alpha+2}}$ (ignoring the logarithmic

factor). This factor can be seen to capture the cost of privacy in terms of the global risk. A smaller ε results in an increase in minimax estimation error, where larger smoothness exacerbates the increase.

A second observation that can be made on the basis of the privacy cost factor, is the cost of distributing observations in a privacy setting. That is to say, if one distributes $N = mn$ observations across m machines, the task becomes more challenging as the N observations are spread over a greater number of machines, rather than having a large number of observations on a smaller number of machines. The relative cost of distributing observations is also revealed to be related to the smoothness, where a larger smoothness again exacerbates the relative cost of distributing data. This observation confirms a folklore understanding that it is easier to retain privacy within a larger crowd. Distributing data across more machines means that each machine needs to add additional noise, to compensate for an overall lack of observations. It also affirms that local differentially private methods perform relatively poorly in multiple observation settings and that applying a privacy constraint at an observation level is comparatively costly.

Classically, the pointwise risk is known to be subject to different phenomena than the global risk over the Besov spaces [15]. Writing $\nu = \alpha - 1/p$ and assuming $\alpha > 1/p$, it is known that the unconstrained pointwise minimax risk satisfies

$$\inf_{\hat{f}} \sup_{f \in \mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f |\hat{f}(x_0) - f(x_0)|^2 \asymp (mn)^{-\frac{2\nu}{2\nu+1}}. \tag{7}$$

Compared to the unconstrained global risk, this indicates that the estimation error at a point is subject to a fundamentally slower convergence rate than the global estimation minimax rate, where the ℓ_p-norm used to measure the smoothness of the Besov ellipsoid influences the minimax estimation performance. Roughly speaking, the "pointwise" integrability of the derivatives of the function underlying the data impacts the problem of estimation at a point, whilst the global risk remains unaffected. This effect disappears for Hölder alternatives, where $p = \infty$ and the minimax rate for the global risk and the pointwise risk coincide.

The main theorem on the minimax risk for pointwise estimation leads to the following result for the homogeneous setting.

COROLLARY 2. *Suppose that $n_j = n$, $\varepsilon_j = \varepsilon$, $\delta_j = \delta$ for $j = 1, \ldots, m$ and $\delta \ll (mn)^{-1}$. Then, for $x_0 \in [0,1]$, the pointwise minimax risk at x_0 over the class $\mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$ satisfies (5).*

The minimax rate for the pointwise risk seemingly takes on a similar form as that of the global risk and it coincides with the global risk whenever $p = \infty$. However, for finite values of p, the cost of privacy can be seen to differ. In particular, to attain the unconstrained optimal pointwise minimax rate (7), it can be seen that a relatively larger ε is needed, where a smaller value of p in fact exacerbates the demand. More precisely, whenever $(mn)^{\frac{1}{2\alpha+1}} \lesssim n\varepsilon^2 \ll (mn)^{\frac{1}{2\nu+1}}$, the pointwise risk suffers from the DP constraints, whereas the global risk performance is the same as in the problem without the DP constraints.

Whenever $n\varepsilon^2 \ll (mn)^{\frac{1}{2\nu+1}}$, comparing (5) to (7) shows that the minimax rate of the classical (unconstrained) pointwise risk increases by a factor of $(m^{\frac{1}{2\nu+1}} n^{-\frac{2\nu}{2\nu+1}} \varepsilon^{-2})^{\frac{2\nu}{2\nu+2}}$ (ignoring the logarithmic factor). This shows that the pointwise risk is subject to a similar cost-relationship as the global risk. What is similar is that more stringent privacy demands in terms of a smaller ε translate to an increased cost in terms of the pointwise risk. However, the relative increase in privacy cost resulting from a decrease in ε for the case of pointwise risk, is smaller than the relative increase in privacy cost of the global risk, where this discrepancy is further exacarbated for smaller values of p. This shows that stringent privacy demands are comparatively less costly for the pointwise risk.

On the other hand, the cost of distributing observations (i.e. increasing m when distributing $N = nm$ observations) is relatively larger for smaller values of p. That is to say, differentially private estimation in pointwise risk suffers less from stringent per machine privacy demands, while it suffers more from the fact that data is distributed before privacy preservation is applied. This surprising phenomenon shows that in a distributed setting with privacy constraints, the distribution of the data across servers impacts the rate differently depending on the inferential task at hand.

2.2. *The heterogeneous setting.* While the homogeneous setting described in the introduction serves to illustrate fundamental phenomena, real-world scenarios often involve heterogeneous data and privacy constraints across various data silos. In applications, data may not be uniformly distributed among different sources. For instance, consider cases where data is observed and processed locally, as in the context of hospitals. The results presented here highlight the optimal estimation under differential privacy in such a heterogeneous setting.

Theorems 4.1 and 3.2 describe the minimax rate for the global risk for the full spectrum of possibilities in terms of heterogeneous constraints. Similarly, Theorems 4.4 and 3.4 describe the minimax rate for the local risk in the heterogeneous setting. Here, for the sake of clarity of interpretation, we will focus on two different regimes of privacy budgets. For both regimes, whenever $\min_j \varepsilon \gtrsim 1/(mn)$, we require that $\min_j \delta_j \ll 1/(mn)^2$, which translates to δ having no further impact on the minimax performance except for incurring a logarithmic factor in case of the global risk. For the first regime, we shall consider privacy budgets where the no single server has much more data than the other servers, comparatively to the stringency in terms of the DP parameter ε. This amounts to

$$\left(\sum_{j=1}^{m} n_j^2 \varepsilon_j^2\right)^{\frac{1}{2\gamma+2}} \geq \max_j n_j \varepsilon_j^2, \tag{8}$$

where $N = \sum_{j=1}^m n_j$ and $\gamma = \alpha$ or $\gamma = \nu$ for the global and pointwise risk respectively. The following result describes the minimax rate in this regime for the global and pointwise risks.

COROLLARY 3. *Suppose that $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$ is such that $\sum_{j=1}^m n_j^2 \varepsilon_j^2 \to \infty$, for $j = 1, \ldots, m$ we have $\delta_j \lesssim (\varepsilon_j^2/\sqrt{m})^{1+\kappa}$ for some $\kappa > 0$ and (8) holds with $\gamma = \alpha$. Then, it holds that*

$$\inf_{\hat{f} \in \mathcal{M}(\boldsymbol{\epsilon}, \boldsymbol{\delta})} \sup_{f \in \mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \left\| \hat{f} - f \right\|_2^2 \asymp M_N \left(\sum_{j=1}^{m} n_j^2 \varepsilon_j^2\right)^{-\frac{2\alpha}{2\alpha+2}}.$$

for some $M_N \geq 1$ of the order at most $\log(N) \cdot \log(1/\min_j \delta_j)$.

If (8) holds for $\gamma = \nu$ and $\sum_{j=1}^m n_j \delta_j \to 0$, it holds that

$$\inf_{\hat{f} \in \mathcal{M}(\boldsymbol{\epsilon}, \boldsymbol{\delta})} \sup_{f \in \mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \left| \hat{f}(x_0) - f(x_0) \right|^2 \asymp M_N \left(\sum_{j=1}^{m} n_j^2 \varepsilon_j^2\right)^{-\frac{2\nu}{2\nu+2}}$$

for some $M_N \geq 1$ of the order at most $\log(N)$.

In such a setting, the behaviour in terms of the privacy cost is similar to that described by Corollaries 1 and 2. A first glance shows that in the distributed privacy setup, the problem is much more difficult compared to the problem without privacy constraints: the rate when no privacy constraints are in place, which is $N^{-\frac{2\alpha}{2\alpha+1}}$. Furthermore, the minimax rate shows

that, when N observations are divided over m machines somewhat equally, there is benefit in dividing over as few machines as possible and there is an additional benefit to having machines with a relatively large amount of data. The explanation for this is the same as described in the homogeneous case: it is easier to retain privacy within large local samples. When the samples are "spread thinly" across the servers, the cost of DP is larger. Between the pointwise risk and the global risk, the phenomenon of pointwise risk incurring relatively less cost when ε_j's are decreased compared to the global risk is also still observed when $p < \infty$, whilst the cost of distributing is relatively higher.

In the regime of (8), even though the privacy budgets vary between the servers, all the servers can be seen to provide a non-negligible contribution to the central estimator. Another regime which we highlight, is the case where some $j^* \in [m]$ the privacy budget satisfies

$$(n_{j^*}^2\varepsilon_{j^*}^2) \wedge \left(n_{j^*}^{\frac{2\gamma+4}{2\gamma+2}}\varepsilon_{j^*}^{\frac{2}{2\gamma+2}}\right) \wedge n_{j^*}^{\frac{2\gamma+2}{2\gamma+1}} \geq \sum_{[m]\setminus\{j^*\}} n_j^2\varepsilon_j^2, \tag{9}$$

where we consider $\gamma = \alpha$ or $\gamma = \nu$ for the global and pointwise risk respectively. This regime is in a sense the juxtaposition of (8). Where in (8), no server has a substantially better privacy budget compared to its number of observations, in the case of (9), there is (at least) one server with a substantially larger sample and/or a relatively better privacy budget than the other servers. The following result captures the minimax rate for the global and pointwise risks in such a regime.

COROLLARY 4. *Suppose that* $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$ *satisfies* $\delta_j \lesssim (\sqrt{n_j}\varepsilon_j^2/(n_{j^*}^{2/3}\varepsilon_{j^*}^{2/3}))^{1+\kappa}$ *for some* $\kappa > 0$ *and all* $j = 1, \ldots, m$, (9) *holds for* $\gamma = \alpha$ *and* $\varepsilon_{j^*} > (n_{j^*})^{-1}$. *Then, it holds that*

$$\inf_{\hat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{f\in\mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f\|\hat{f} - f\|_2^2 \asymp M_{m,n}\left(\left(n_{j^*}^2\varepsilon_{j^*}^2\right)^{-\frac{2\alpha}{2\alpha+2}} + (n_{j^*})^{-\frac{2\alpha}{2\alpha+1}}\right)$$

for some $M_N \geq 1$ *of the order at most* $\log(N)\cdot\log(1/\min_j\delta_j)$. *If* (9) *holds for* $\gamma = \nu$, *it holds that*

$$\inf_{\hat{f}\in\mathcal{M}(\boldsymbol{\varepsilon},\boldsymbol{\delta})} \sup_{f\in\mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f|\hat{f}(x_0) - f(x_0)|^2 \asymp M_{m,n}\left(\left(n_{j^*}^2\varepsilon_{j^*}^2\right)^{-\frac{2\nu}{2\nu+2}} + (n_{j^*})^{-\frac{2\nu}{2\nu+1}}\right)$$

for some $M_N \geq 1$ *of the order at most* $\log(N)$.

In the regime described by the theorem, certain servers have a large sample and relatively large privacy budget, compared to the "majority" of the other servers in the sense of (9). The minimax rate derived describes that in such settings these large sample/budget servers dictate the statistical accuracy of estimation. This is true both for the global, as well as the pointwise risk. In terms of optimal estimation procedures, the minimax rate can be achieved by only using estimators based on the data of the server(s) with relatively large samples and privacy budget, as the benefit of the servers with smaller samples and privacy budgets have an asymptotically negligible benefit.

3. Optimal Distributed (ε, δ)-DP Estimators. In this section, we present two estimators that attain the optimal rates as described by the theorems of the previous section. One estimator specifically targets the global risk, the other is constructed specifically to perform well in terms of the pointwise risk. Whilst it perhaps natural to estimate $f(x_0)$ using the global risk DP-estimator evaluated at x_0, the specific pointwise estimator we propose combines DP-estimates of $f(x_0)$ computed locally (i.e. estimators of $f(x_0)$ computed at each of the servers). This approach offers several benefits, such as an improved performance regardless of the value of $\delta_j \geq 0$. Both estimators are constructed using a wavelet basis.

Wavelets are known to have many favourable properties when using them for function estimation in classical settings, see for example [13, 26, 40]. Under DP constraints, wavelet constructions have other desirable properties: they allow for exact control of the estimator's *sensitivity* to changes in the data. Loosely speaking, this allows us to control the "influence" each individual observation has on the outcome of the estimator, whilst retaining the information the full sample has to a large extent.

3.1. *Wavelets and Besov spaces.* In the context of nonparametric regression, we aim to construct an optimal estimator for an unknown function f based on the distributed data. Here, we assume that f belongs to the Besov space $\mathcal{B}^{\alpha}_{p,q}$. Roughly stated, the Besov space $\mathcal{B}^{\alpha}_{p,q}$ containts functions having α bounded derivatives in L_p-space, with q giving a finer control of the degree of smoothness. We refer the reader to [69] for a detailed description.

Wavelet bases allow characterization of the Besov spaces, where α, p and q are parameters that capture the decay rate of wavelet basis coefficients. Before presenting the two optimal estimators for global and pointwise risk in Sections 3.2 and 3.3 respectively, we first briefly introduce wavelets and collect some properties used to define the Besov space. For a more detailed and elaborate introduction of wavelets in the context of Besov spaces, we refer to [43, 46].

In our work we consider the Cohen, Daubechies and Vial construction of compactly supported, orthonormal, A-regular wavelet basis of $L_2[0,1]$, see for instance [23]. First for any $A \in \mathbb{N}$ one can follow Daubechies' construction of the father $\phi(\cdot)$ and mother $\psi(\cdot)$ wavelets with A vanishing moments and bounded support on $[0, 2A-1]$ and $[-A+1, A]$, respectively, for which we refer to [24]. The basis functions are then obtained as

$$\big\{\phi_{l_0+1,m}, \psi_{lk} : m \in \{0, ..., 2^{l_0+1} - 1\}, \quad l \geq l_0 + 1, \quad k \in \{0, ..., 2^l - 1\}\big\},$$

with $\psi_{lk}(x) = 2^{l/2}\psi(2^l x - k)$, for $k \in [A-1, 2^l - A]$, and $\phi_{l_0+1,k}(x) = 2^{l_0+1}\phi(2^{l_0+1}x - m)$, for $m \in [0, 2^{l_0+1} - 2A]$, while for other values of k and m, the functions are specially constructed, to form a basis with the required smoothness property. In a slight abuse of notation, we shall denote the father wavelet by $\psi_{l_0 k} = \phi_{l_0+1,k}$ and represent any function $f \in L_2[0,1]$ in the form

$$f = \sum_{l=l_0}^{\infty} \sum_{k=0}^{2^l-1} f_{lk}\psi_{lk}, \tag{10}$$

where the $f_{lk} = \langle f, \psi_{lk} \rangle$ are called the *wavelet coefficients*. Note that in view of the orthonormality of the wavelet basis the L_2-norm of the function f is equal to

$$\|f\|_2^2 = \sum_{l=l_0}^{\infty} \sum_{k=0}^{2^l-1} f_{lk}^2.$$

Next we give definition of Besov spaces using wavelets. Let us define the norms

$$\|f\|_{\mathcal{B}^{\alpha}_{p,q}} := \begin{cases} \left(\sum\limits_{l=l_0}^{\infty} \left(2^{l(\alpha+1/2-1/p)} \left\| (f_{lk})_{k=0}^{2^l-1} \right\|_p \right)^q \right)^{1/q} & \text{for } 1 \leq q < \infty, \\ \sup\limits_{l \geq l_0} 2^{l(\alpha+1/2-1/p)} \left\| (f_{lk})_{k=0}^{2^l-1} \right\|_p & \text{for } q = \infty, \end{cases}$$

for $\alpha \in (0, A)$, $1 \leq q \leq \infty$, $2 \leq p \leq \infty$. Then, the Besov space $\mathcal{B}^{\alpha}_{p,q}[0,1]$ and Besov ball $\mathcal{B}^{\alpha,R}_{p,q}[0,1]$ of radius $R > 0$ can be defined as

$$\mathcal{B}^{\alpha}_{p,q}[0,1] = \{f \in L_2[0,1] : \|f\|_{\mathcal{B}^{\alpha}_{p,q}} < \infty\} \text{ and } \mathcal{B}^{\alpha,R}_{p,q}[0,1] = \{f \in L_2[0,1] : \|f\|_{\mathcal{B}^{\alpha}_{p,q}} \leq R\}, \tag{11}$$

respectively. The above definition of the Besov space and norm is equivalent to the classical one based on the weak derivatives of the function (see e.g. Chapter 4 in [39]).

For the construction of our estimators, we consider a A-smooth wavelet basis ($A > \alpha$) with a compactly supported mother wavelet ψ such that wavelets $\psi_{lk}(x) = 2^{l/2}\psi(2^l x - k)$ for $l \geq l_0$ and $k = 0, \ldots, 2^l - 1$ form an orthonormal basis for $\mathcal{B}^{\alpha}_{p,q}[0,1]$.

We will briefly describe in broad terms the idea behind using the wavelet transform (10) to construct the global and pointwise optimal estimators. Both estimators are based on wavelet approximations up until a limited resolution level. Besides the excellent approximation properties of wavelets in Besov spaces (see e.g. [39]), the first property ensures that a change in the data in terms of $X_i^{(j)}$ has a limited change in terms of the "size" the wavelet estimator. The second and third property yield a limited support that shrinks at higher "resolution levels" of the wavelet functions, which controls the number of wavelet coefficients affected by a change in $X_i^{(j)}$. Making sure that changes in individual datums have a limited effect on the shared transcript is crucial for assuring privacy. A further, more detailed description of how these properties interlink is given in the Sections 3.2 and 3.3 below.

3.2. *Constructing an optimal global estimator.* We now proceed to construct the estimator, utilizing the wavelet transform of (10), allowing the representation of a function f in L_2 as a linear combination of wavelet basis functions. We first introduce some notation. For $\tau > 0, x \in \mathbb{R}$, let $[x]_\tau$ denote x clipped at the threshold τ:

$$[x]_\tau := \begin{cases} \tau & \text{if } x > \tau, \\ x & \text{if } -\tau \leq x \leq \tau \\ -\tau & \text{otherwise.} \end{cases}$$

Given $L \in \mathbb{N}$ and $\tau > 0$, each machine $j = 1, \ldots, m$ computes the real numbers

$$\hat{f}^{(j)}_{lk;\tau} = \frac{1}{n_j} \sum_{i=1}^{n_j} \left[Y_i^{(j)} \right]_\tau \psi_{lk}\left(X_i^{(j)} \right), \tag{12}$$

for $l, k \in \mathbb{N}$ such that $l_0 \leq l \leq L$, $0 \leq k \leq 2^l - 1$. We will specify the exact choice of τ and L later. These numbers, which we denote as the vector

$$\hat{\boldsymbol{f}}^{(j)}_{L,\tau} := \left\{ \hat{f}^{(j)}_{lk;\tau} : k = 0, \ldots, 2^l - 1, l = l_0, \ldots, L \right\},$$

will form the statistic underlying our transcript. To assure privacy, we aim to communicate a noisy version of this vector. Adding additional noise leads to an estimator that is necessarily worse, adding noise of a large enough magnitude yields a final transcript satisfies the privacy guarantee of Definition 1.1. To control the magnitude of the noise that needs to be added, it is important to have a statistic that does not change too drastically when the underlying data is changed in one data point. We formalize this in terms of the *sensitivity* of the statistic $\hat{\boldsymbol{f}}^{(j)}_{L,\tau}$.

The following lemma controls the L_2-*sensitivity* of the statistic $\hat{\boldsymbol{f}}^{(j)}_{L,\tau}$, i.e. the difference in Euclidian distance when applied to two neighboring data sets.

LEMMA 3.1. *Let $Z^{(j)}$ and $\tilde{Z}^{(j)}$ any realizations of neighboring data sets. It holds that*

$$\left\| \hat{\boldsymbol{f}}^{(j)}_{L,\tau}(Z^{(j)}) - \hat{\boldsymbol{f}}^{(j)}_{L,\tau}(\tilde{Z}^{(j)}) \right\|_2 \leq c_\psi \frac{\tau\sqrt{2^L}}{n_j}$$

where c_ψ is a constant depending only on the choice of wavelet basis.

We provide a proof for the lemma in Section B.1.1. The limited L_2-sensitivity of the $\hat{\boldsymbol{f}}^{(j)}_{L,\tau}$ is a consequence of merging two elements in its construction. First of, clipping limits the change in (12) when $Y^{(j)}_i$ is exchanged for another data point $\tilde{Y}^{(j)}_i$. In the vector as a whole, the coordinate wise change is limited by the compact support of the wavelet basis. The essential feature of the wavelet basis here is that, even though the basis elements increase exponentially as the resolution levels l increases, their support shrinks proportionally, ensuring that each $X^{(j)}_i$ is in the support of only finitely many wavelets at each resolution level. That is, there are at most $c_A > 0$ number of basis functions ψ_{lk} with overlapping support at each resolution level l, where $c_A > 0$ depends on $A > \alpha$. This means that changing one datum in $Z^{(j)}$, say $(Y^{(j)}_i, X^{(j)}_i)$ to $(\tilde{Y}^{(j)}_i, \tilde{X}^{(j)}_i)$, has only a limited impact at the level of the transcript $\hat{\boldsymbol{f}}^{(j)}_{L,\tau}(Z^{(j)})$.

The bounded L_2-sensitivity means that the statistic $\hat{\boldsymbol{f}}^{(j)}_{L,\tau}(Z^{(j)})$, combined with additive, appropriately scaled Gaussian noise satisfies $(\varepsilon_j, \delta_j)$-differentially privacy. This result for mappings with bounded L_2-sensitivity is well known and a proof can be found in e.g. Appendix A of [31]. To be precise, the j-th server outputs $\tilde{T}^{(j)}_{lk;\tau} = \hat{f}^{(j)}_{lk;\tau} + W^{(j)}_{lk}$ for $k = 0, \ldots, 2^l - 1$, $l = l_0, \ldots, L$, where the coordinates of $\boldsymbol{W}^{(j)} := (W^{(j)}_{lk} : k = 0, \ldots, 2^l - 1, l = l_0, \ldots, L)$ are i.i.d. mean zero gaussian with variance $\frac{4\tau^2 2^L c_\psi^2 \log(2/\delta_j)}{n_j^2 \varepsilon_j^2}$. The constant $c_\psi := 2\sqrt{2}\sqrt{c_A}\|\psi\|_\infty$ matches the constant in Lemma 3.1. The addition of the gaussian noise ensures that the transcript

$$T^{(j)}_{L,\tau} := \left\{ T^{(j)}_{lk;\tau} \,:\, k = 0, \ldots, 2^l - 1, \, l = l_0, \ldots, L \right\} = \hat{\boldsymbol{f}}^{(j)}_{L,\tau}(Z^{(j)}) + \boldsymbol{W}^{(j)}$$

is $(\varepsilon_j, \delta_j)$-differentially private.

The final estimator of f is then obtained via a post-processing step in which each of the transcripts is reweighted, taking the heterogeneity between the servers into account. The choice of weight depends crucially on the local number of observations n_j and the local privacy constraint ε_j. Given the transcripts $T = (T^{(1)}_{L,\tau}, \ldots, T^{(m)}_{L,\tau})$, the final estimator takes the form of

$$\hat{f}_{L,\tau}(x) = \sum_{l=l_0}^{L} \sum_{k=0}^{2^l - 1} \left(\sum_{j=1}^{m} u_j T^{(j)}_{lk;\tau} \right) \psi_{k,l}(x),$$

where the weights are given by

$$u_j = \frac{v_j}{\sum_j v_j} \quad \text{with} \quad v_j = \left(n_j^2 \varepsilon_j^2\right) \wedge \left(n_j 2^L\right). \tag{13}$$

The following theorem captures the global risk attained by the estimator $\hat{f}_{L,\tau}$ resulting from the distributed $(\boldsymbol{\epsilon}, \boldsymbol{\delta})$-DP procedure outlined above, with optimal selection of L and a sufficiently large choice of τ. For the latter, a choice of $C_{\alpha,R} + \sqrt{(2\alpha+1)L}$ is adequate, where $C_{\alpha,R} > 0$ is a constant, as specified by Lemma B.6 or a larger constant.

The variance of the Gaussian noise vectors $\boldsymbol{W}^{(j)}$, which yield the privacy guarantee, increases with L. Consequently, the optimal choice of L is not just governed by the classical bias variance trade-off, but also by the trade-off in the additional noise required to guarantee privacy.

The optimal choice of L is taken as follows. Let $D > 0$ be the number solving the equation

$$D^{2\alpha+2} = \sum_{j=1}^{m} \left(n_j^2 \varepsilon_j^2\right) \wedge \left(n_j D\right). \tag{14}$$

Setting $L = (l_0+1) \vee \lceil \log_2(D) \rceil$ yields the optimal performance as described by the theorem below, in terms of a bias-variance-sensitivity trade-off. Furthermore, this performance turns out to be the theoretically best possible performance in a minimax sense, as established by the lower bound of Theorem 4.1 in Section 4.

THEOREM 3.2. *Set* $\tau = C_{\alpha,R} + \sqrt{(2\alpha+1)L}$ *and take* $L = (l_0+1) \vee \lceil \log_2(D) \rceil$, *where* $D > 0$ *is the solution to* (14).

Then, the L_2*-risk of the distributed* $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$*-DP protocol* $\hat{f}_{L,\tau}$ *satisfies*

$$\sup_{f \in \mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \left\| \hat{f}_{L,\tau} - f \right\|_2^2 \leq C_\psi \log(N) 2^{-2L\alpha} \log(2/\delta'),$$

where $\delta' = \min_{i \in [m]} \delta_i$ *and* C_ψ *denotes a constant depending on* ψ.

We briefly comment on the derived result. We first note that the choice of wavelet basis (in particular the father wavelet ψ) influences the constants in the theorem, but not the convergence rate. The rate attained by the choice of L as directed by (14) yields optimal rate as given in Corollary 1 in case of homogeneous servers, the optimal rate of Corollary 3 in case the privacy budgets satisfy (8) or the optimal rates of Corollary 4 in case the budget satisfies (9).

3.3. *Constructing an optimal estimator of* f *at a point.* We now turn to the task of estimating the unknown function $f \in \mathcal{B}_{p,q}^{\alpha,R}$ at a given point $x_0 \in (0,1)$. That is to say, we will construct an estimator $\hat{f}$ such that $\mathbb{E}_f(\hat{f}(x_0) - f(x_0))^2$ achieves the optimal rates as predicated by Corollaries 2, 3 and 4.

A natural estimator of $f(x_0)$ is to use the global plug-in estimator of the previous section, $\hat{f}_{L,\tau}(x_0)$, with $\hat{f}_{L,\tau}$ as constructed in the previous section. However, for estimation at a point as goal of inference, we instead opt for estimating $f(x_0)$ locally. Similarly to the pre-processing step in Section 3.2, the preliminary local estimator is to be perturbed with noise to ensure it satisfies the DP constraint of Definition 1.1. Adding Laplacian noise turns out to suffice to ensure $(\varepsilon_j, 0)$-DP, which is a stronger guarantee than $(\varepsilon_j, \delta_j)$-DP. Another advantage to this approach compared to the plug-in estimator $\hat{f}_{L,\tau}(x_0)$ is that the procedure derived below has a $\log(N)$-factor improved rate compared to the plug-in estimator.

As a first step in constructing the estimator of $f(x_0)$, we consider for $L \in \mathbb{N}$ and $\tau > 0$ the first wavelet coefficients $\hat{\boldsymbol{f}}_{L,\tau}^{(j)}$ as computed in (12). On the j-th server, we construct the estimator corresponding to $\hat{\boldsymbol{f}}_{L,\tau}^{(j)}$, which we then evaluate in the point x_0,

$$\hat{f}_{L,\tau}^{(j)}(x_0) \equiv \hat{f}_{L,\tau}^{(j)}(x_0 | Z^{(j)}) := \sum_{l=l_0}^{L} \sum_{k=0}^{2^l-1} \hat{f}_{lk;\tau}^{(j)} \psi_{lk}(x_0). \tag{15}$$

In order to create a $(\varepsilon_j, 0)$-DP transcript, we add Laplacian noise to $\hat{f}_{L,\tau}^{(j)}(x_0)$ directly. Laplacian noise performs well for statistics with small L_1-*sensitivity*, i.e. the change in L_1-norm when one datum is changed underlying the statistic. The L_1-sensitivity scales poorly in the dimension of the statistic compared to the L_2-sensitivity dictating the noise of the Gaussian mechanism of Section 3.2. Essentially, the estimator proposed in (15) has large sensitivity only for observations that are "close" to x_0. To see this, note that $\psi_{lk} \lesssim 2^{l/2}$ and that the latter estimator can be written as

$$\hat{f}_{L,\tau}^{(j)}(x_0 | Z^{(j)}) = \sum_{l=l_0}^{L} \sum_{k \in K_l(x_0)} \hat{f}_{lk;\tau}^{(j)} \psi_{lk}(x_0),$$

where $K_l(x_0) := \{k : \psi_{lk}(x_0) \neq 0\}$. Using this fact, the lemma below gives an exact bound on the L_1-sensitivity of the functional $\hat{f}^{(j)}_{L,\tau}(x_0)$.

LEMMA 3.3. *Let $Z^{(j)}$ and $\tilde{Z}^{(j)}$ any realizations of neighboring data sets. It holds that*

$$\left\| \hat{f}^{(j)}_{L,\tau}(x_0|Z^{(j)}) - \hat{f}^{(j)}_{L,\tau}(x_0|\tilde{Z}^{(j)}) \right\|_1 \leq c'_\psi \frac{\tau 2^L}{n_j}, \tag{16}$$

where $c''_\psi := 2c_A\|\psi\|^2_\infty$ is a constant depending only on the choice of wavelet basis.

This bound on the L_1-sensitivity yields that as a privacy mechanism it suffices to add Laplace noise with variance $\frac{c'_\psi \tau 2^L}{n_j \varepsilon_j}$ to the functional $\hat{f}^{(j)}_{L,\tau}(x_0)$. That is, the transcript $T^{(j)}_{lk,\tau}$ for $j \in [m]$ given by

$$T^{(j)}_{L,\tau} = \hat{f}^{(j)}_{L,\tau}(x_0) + W^{(j)}, \quad \text{where} \quad W^{(j)} \overset{i.i.d}{\sim} \text{Lap}\left(0, \frac{c'_\psi \tau 2^L}{n_j \varepsilon_j}\right)$$

is $(\varepsilon_j, 0)$-DP (see e.g. [31]). With each j-th server transmitting $T^{(j)}_{L,\tau}$, the estimator computed in the central server is given by $\hat{f}(x_0) = \sum_{j=1}^m u_j T^{(j)}_{L,\tau}$.

It remains to determine the optimal choice of L. Similarly as in in the case of the estimator of global risk as presented in Section 3.2, there is a trade-off between bias, variance and sensitivity, where the L_1-sensitivity can be seen to have a dependence on L. The explanation for this, is that even though the functional $\hat{f}^{(j)}_{L,\tau}(x_0)$ is unidimensional, the wavelet resolution level L still determines how a change in an individual datum can potentially change the value of a local estimator defined in (15).

Here, the choice of L is governed by $L = (l_0 + 1) \vee \lceil \log_2(D) \rceil$, where $D > 0$ be the number solving the equation

$$D^{2\nu+2} = \sum_{j=1}^m \left(n_j^2 \varepsilon_j^2\right) \wedge (n_j D)\,. \tag{17}$$

The following theorem describes the performance of the pointwise estimator on the basis of the $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$-DP transcript $T = (T^{(1)}_{L,\tau}, \ldots, T^{(m)}_{L,\tau})$ for a sufficiently large choice of τ, such as $\tau = C_{\alpha,R} + \sqrt{(2\alpha+1)L}$, with $C_{\alpha,R} > 0$ as given by Lemma B.6.

THEOREM 3.4. *Set $\tau = C_{\alpha,R} + \sqrt{(2\alpha+1)L}$ and take $L = (l_0 + 1) \vee \lceil \log_2(D) \rceil$, where D is governed by (17).*

Then, the pointwise ℓ_2-risk of the distributed $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$-DP protocol $\hat{f}_{L,\tau}$ satisfies

$$\sup_{f \in \mathcal{B}^{\alpha,R}_{p,q}} \mathbb{E}_f (\hat{f}(x_0) - f(x_0))^2 \leq C_\psi \log(N) 2^{-2L\nu}.$$

for a constant $C_\psi > 0$ depending only on the choice of wavelet basis.

The rate attained by the choice of L as directed by (17) does not just yield the best possible bias-variance-sensitivity trade-off for the estimator class under consideration, but it turns out to be minimax optimal (up to a log factor) as established in the lower bound of Theorem 4.4. It consequently yields the optimal rate as given in Corollary 2 in case of homogeneous servers, the optimal rate of Corollary 3 in case the privacy budgets satisfy (8) or the optimal rates of Corollary 4 in case the budget satisfies (9). As is the case with Theorem 3.2, the choice of wavelet basis influences the constants in the theorem, but not the convergence rate.

4. Minimax Lower Bounds. Theorems 3.2 and 3.4 provide the rates of convergence for the proposed estimators of f and $f(x_0)$, respectively. In this section we shall show that these rates of convergence are indeed optimal among all estimators by establishing two matching minimax lower bounds, up to logarithmic factors, for global and pointwise estimation. These results affirm the optimality of the estimators presented in Section 3.

The lower bounds are presented in Theorems 4.1 and 4.4 for the global and pointwise risks, respectively. The derivation of each lower bound relies on entirely distinct techniques, elaborated in Section 4.1 and Section 4.2.

4.1. *Minimax lower bounds under heterogeneous distributed DP constraints.* The following theorem states a lower bound on the minimax risk for global estimation.

THEOREM 4.1. *Let $D > 0$ be the solution to (14) and assume that $\delta_j < (n_j^{1/2}\varepsilon_j^2(D \vee 1)^{-1})$ for some $\kappa > 0$ and all $j \in [m]$. Then, we have the following lower bound on the minimax risk:*

$$\inf_{\hat{f} \in \mathcal{M}(\boldsymbol{\varepsilon}, \boldsymbol{\delta})} \sup_{f \in \mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \|\hat{f} - f\|_2^2 \gtrsim D^{-2\alpha} \wedge 1. \tag{18}$$

Before outlining the proof, we briefly note that this lower bound matches that of the upper bound of Theorem 3.2 (up to a log-factor), for the choice $L = (l_0 + 1) \vee \lceil \log_2(D) \rceil$ in the estimator under consideration in Section 3.2. The theorem affirms that, up to a log factor, the proposed estimator attains the best rate among all privacy constrained estimators.

Next, we discuss the most important steps in the proof here, whilst leaving the technical details to the appendix. To lower bound the global risk, we first restrict to a finite-dimensional sub-model of the Besov space $\mathcal{B}_{p,q}^{\alpha,R}$. To align notation with the previous section, we shall use the wavelet basis from before to do so. Given $L \in \mathbb{N}$, we consider the finite-dimensional subspace

$$\mathcal{B}_{p,q}^{\alpha,R,L} := \left\{ f \in \mathcal{B}_{p,q}^{\alpha,R} : f = \sum_{k=0}^{2^L-1} f_{Lk}\psi_{Lk},\ f_{Lk} \in [-2^{-L(\alpha+1/2)}R, 2^{-L(\alpha+1/2)}R] \right\}.$$

Let $\boldsymbol{\psi}(X)$ denote the 2^L dimensional vector $\{\psi_{Lk}(X)\}_{k=1}^{2^L}$ and define

$$\boldsymbol{S}_f\left(Z_i^{(j)}\right) := \sigma^{-1}\left(Y_i^{(j)} - \sum_{k=0}^{2^L-1} f_{Lk}\psi_{Lk}\left(X_i^{(j)}\right)\right)\boldsymbol{\psi}\left(X_i^{(j)}\right). \tag{19}$$

The random vector $\boldsymbol{S}_f(Z_i^{(j)})$ can be seen as an "score function" of the i-th observation on the $j-$th server, within the finite dimensional sub-model. Similarly, consider the "score function" for local data $Z^{(j)}$ on the jth server; $\boldsymbol{S}_f(Z^{(j)}) := \sum_{i=1}^{n_j} \boldsymbol{S}_f(Z_i^{(j)})$. Furthermore, let $\boldsymbol{C}_f(T^{(j)})$ denote the 2^L dimensional matrix

$$\mathbb{E}\,\mathbb{E}\left[\boldsymbol{S}_f\left(Z^{(j)}\right) \mid T^{(j)}\right]\mathbb{E}\left[\boldsymbol{S}_f\left(Z^{(j)}\right) \mid T^{(j)}\right]^T. \tag{20}$$

We shall write $\boldsymbol{C}_f(Z^{(j)})$ for the unconditional version covariance matrix of $\boldsymbol{S}_f(Z^{(j)})$and let $\boldsymbol{C}_{f,i}^{(j)} = \mathbb{E}\left[\boldsymbol{S}_f(Z_i^{(j)})\boldsymbol{S}_f(Z_i^{(j)})^T\right]$, such that $\boldsymbol{C}_f = \sum_{i=1}^{n_j} \boldsymbol{C}_{f,i}^{(j)}$.

Using the Van-Trees inequality (with a prior as specified later on in the section), we obtain an expression in terms of the sum-of-traces of the matrices in display (20), i.e. the covariance of the score function $\boldsymbol{S}_f(Z^{(j)})$, conditionally on the released transcripts.

As the conditional expectation contracts the L_2-norm, we have the "data processing" bound $\boldsymbol{C}_f(T^{(j)}) \leq \boldsymbol{C}_f(Z^{(j)})$, which in turn implies that

$$\mathrm{Tr}(\boldsymbol{C}_f(T^{(j)})) \leq \mathrm{Tr}(\boldsymbol{C}_f(Z^{(j)})). \tag{21}$$

The right-hand side can be bounded by $2^L n_j$ by direct calculation, which we defer to Section B.2. These standard bounds do not take the privacy constraints into account and would lead to the unconstrained minimax rate.

To capture the loss of information stemming from the DP constraint of Definition 1.1, a more sophisticated data processing argument is required. This bring us to one of the key technical innovations of the paper, which comes in the form of a data-processing inequality (Lemma 4.2 below) for the conditional covariance given a $(\varepsilon_j, \delta_j)$-differentially private transcripts of linear functionals of the data such as the score $\boldsymbol{S}_f(Z_i^{(j)})$. The lemma can be seen as a geometric version of the "score attack" lower bound of [18]. Combining this data processing step with the linearity of the trace accommodates for the heterogeneity between the servers.

LEMMA 4.2. *Let $\delta_j \log(1/\delta_j) < n_j^{1/2}\varepsilon_j^2 (D \vee 1)^{-1}$ for $j = 1, \dots, m$. There exists a universal constant $C > 0$ such that*

$$\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{C}_f(T^{(j)}))\right] \leq C n_j \varepsilon_j \sqrt{\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{C}_f(T^{(j)}))\right]} \sqrt{\lambda_{\max}(\boldsymbol{C}_{f,i})} + C\delta_j \left(2^L n_j^{1/2} \log(1/\delta_j) + n_j\right).$$

In Section B.2 of the appendix, we show that the largest eigenvalue of $\boldsymbol{C}_{f,i}$; $\lambda_{\max}(\boldsymbol{C}_{f,i})$, is bounded, from which it follows from the $\mathbb{E}\left[\mathrm{Tr}(\boldsymbol{C}_f(T^{(j)}))\right] \lesssim n_j^2 \varepsilon_j^2$ uniformly for $f \in \mathcal{B}_{p,q}^{\alpha,R,L}$ whenever δ_j is of smaller than $n_j^{1/2}\varepsilon_j^2 D^{-1}$.

With the two bounds on the trace of $\boldsymbol{C}_f(T^{(j)})$ in hand, we now lower bound global estimation risk using the Van-Trees inequality. The Van-Trees inequality provides an expression in terms of the trace of a certain covariance matrix, which is the conditional covariance of a linear functional of the data. Combined with the data processing inequalities, the linearity of the trace accommodates for the heterogeneity between the servers.

In order to apply the Van-Trees inequality, we first define a prior such that the worst-case global risk is lower bounded by the corresponding Bayes risk. To that extent, we define a prior Π that is supported on $\mathcal{B}_{p,q}^{\alpha,R,L}$. Given the resolution level $L \in \mathbb{N}$, we draw f_{Lk} independently from the probability distribution Π_{Lk}, defined through an appropriately rescaled version of the density $t \mapsto \cos^2(\pi t/2)\mathbb{1}_{|t|\leq 1)}$ such that has its support equal to $[-2^{-L(\alpha+1/2)}R, 2^{-L(\alpha+1/2)}R]$ for $k = 0 \dots, 2^L - 1$ and set $f_{lk} = 0$ otherwise. For this choice of prior, the Van-Trees inequality of [38] yields the following lemma, for which we defer the details of the proof to Section B.2 in the appendix.

LEMMA 4.3. *It holds that $\sup_{f\in\mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f\|\hat{f} - f\|_2^2$ is lower bounded by the Bayes risk $\int \mathbb{E}_f\|\hat{f} - f\|_2^2 d\Pi(f)$, which is further lower bounded as follows*

$$\int \mathbb{E}_f\|\hat{f} - f\|_2^2 d\Pi(f) \geq \frac{2^{2L}}{\sup_{f\in\mathcal{B}_{p,q}^{\alpha,R,L}} \sum_{j=1}^m \mathrm{Tr}(\boldsymbol{C}_f(T^{(j)})) + \pi^2 2^{L(2\alpha+2)}}.$$

Combining the upper bound on the trace of $\boldsymbol{C}_f(T^{(j)})$ of (21) and Lemma 4.2, we have, by Lemma 4.3, that

$$\sup_{f\in\mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f\|\hat{f} - f\|_2^2 \gtrsim \frac{2^{2L}}{\sum_{j=1}^m n_j^2\varepsilon_j^2 \wedge n_j 2^L + \pi^2 2^{L(2\alpha+2)}}.$$

We obtain the desired lower bound by choosing an L that maximizes the lower bound. Setting $L = (l_0+1) \vee \lceil \log_2(D) \rceil$ can be seen to do so by the relationship (14), which proves Theorem 4.1.

4.2. *Lower bound for the pointwise risk.* In this section, we derive the minimax lower bound for the pointwise risk. We first present the lower bound as the main result of the section in the form of Theorem 4.4, after which we discuss its proof. The theorem tells us that the pointwise risk estimator proposed in Section 3.3 performs optimally in terms of achieving the minimax privacy constrained rate up to a logarithmic factor.

THEOREM 4.4. *Let $D > 0$ be the number solving the equation*

$$D^{2\nu+2} = \sum_{j=1}^{m} \left(n_j^2 \varepsilon_j^2\right) \wedge (n_j D). \tag{22}$$

Assume furthermore that $\sum_j n_j \delta_j \to 0$. Then, for any $x_0 \in (0,1)$, the minimax pointwise risk is lower bounded as follows:

$$\sup_{f \in \mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \left(\hat{f}(x_0) - f(x_0)\right)^2 \gtrsim D^{-2\nu} \wedge 1.$$

Whenever $\left(\sum_{j=1}^m n_j^2 \varepsilon_j^2\right)^{\frac{1}{2\nu+2}} \geq \max_j n_j \epsilon_j^2$, the right hand side is further bounded from below by $\left(\sum_{j=1}^m n_j^2 \varepsilon_j^2\right)^{-\frac{2\nu}{2\nu+2}} \wedge 1$.

The proof of the theorem is based around the Le Cam two point method, which is a common approach to lower bounding the pointwise risk, see for example [72]. However, to capture the effect of the transcripts satisfying the DP constraint of Definition 1.1, we introduce a coupling argument in conjunction.

We briefly sketch the two point method and coupling argument here, leaving the technical details to the appendix. Take any function $f \in \mathcal{B}_{p,q}^{\alpha}$ such that $\|f\|_{\mathcal{B}_{p,q}^{\alpha}} = R' < R$ and a compactly supported function $g \in \mathcal{B}_{p,q}^{\alpha}$ such that $\|g\|_{\mathcal{B}_{p,q}^{\alpha}} \leq R - R'$ and $g(0) > 0$. Define a third function

$$\tilde{f}(t) := \gamma_D^{-1} g(\beta_D (t - x_0)) + f(t),$$

where $\gamma_D := c_0^{-1} D^{\nu}$ and $\beta_D = \gamma_D^{1/\nu}$, where we recall that $\nu = \alpha - \frac{1}{p}$. By e.g. Lemma 1 from [15], $\|f\|_{\mathcal{B}_{p,q}^{\alpha}} \leq R$.

Let $(Y_i^{(j)}, X_i^{(j)}) \sim P_f$ and $(\tilde{Y}_i^{(j)}, \tilde{X}_i^{(j)}) \sim P_{\tilde{f}}$ for individual observations generated according to (1) with either f or $\tilde{f}$ the true underlying regression function respectively. We construct a coupling between P_f and $P_{\tilde{f}}$ such that $(Y_i^{(j)}, X_i^{(j)})$ and $(\tilde{Y}_i^{(j)}, \tilde{X}_i^{(j)})$ are equal with probability proportional to $\sigma^{-1}\|\tilde{f} - f\|_1$, which forms the content of the following lemma.

LEMMA 4.5. *There exists a joint distribution $P_{f,\tilde{f}}$ of $\left((Y_i^{(j)}, X_i^{(j)}), (\tilde{Y}_i^{(j)}, \tilde{X}_i^{(j)})\right)$ such that*

$$\rho := P_{f,\tilde{f}}\left(\left(Y_i^{(j)}, X_i^{(j)}\right) \neq \left(\tilde{Y}_i^{(j)}, \tilde{X}_i^{(j)}\right)\right) \leq \frac{c}{\sigma}\|\tilde{f} - f\|_1, \tag{23}$$

for a universal constant $c > 0$.

We prove the above lemma in Section C.3. Loosely speaking, the quantity ρ captures the difficulty of distinguishing individual observations from P_f of those generated from $P_{\tilde{f}}$.

Consider now transcripts $T = (T^{(1)}, \ldots, T^{(m)})$ each satisfying the DP constraint of Definition 1.1 with a privacy budget $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$, and let $\mathbb{P}_f$ denote the joint law of transcripts and the $N = \sum_{j=1}^m n_j$ observations generated from P_f. Let $\mathbb{P}_f^T$ denote the push-forward measure of the transcript, i.e. its marginal distribution given that the data is generated by P_f. Similarly, let $\mathbb{P}_{\tilde{f}}$ denote the joint law of T with the data generated from $P_{\tilde{f}}$ and let $\mathbb{P}_{\tilde{f}}^T$ denote the corresponding marginal distribution of T. With the coupling of Lemma 4.5 in hand, we derive the following lemma.

LEMMA 4.6. *For any subset $S \subseteq [m]$,*

$$\left\|\mathbb{P}_f^T - \mathbb{P}_{\tilde{f}}^T\right\|_{\mathrm{TV}} \leq \sqrt{2}\sqrt{\sum_{j\in S} \bar{\varepsilon}_j \left(e^{\bar{\varepsilon}_j} - 1\right) + \sum_{j\in S^c} n_j D_{\mathrm{KL}}(P_f; P_{\tilde{f}})} + 4\sum_{j\in S} e^{\bar{\varepsilon}_j} n_j \delta_j \rho, \tag{24}$$

where $\bar{\varepsilon}_j = 6 n_j \varepsilon_j \rho$, ρ as defined in (23).

We defer a proof of the lemma to Section C.3 of the appendix. The lemma allows analysis of the contributions of the separate the servers, accounting for the heterogeneity in the privacy budgets (ϵ_j, δ_j) and the differing number of observations. Roughly speaking, for servers with relatively large privacy budgets, their contribution to the estimator is to be captured by $n_j D_{\mathrm{KL}}(P_f; P_{\tilde{f}}^T)$, which does not involve the privacy budget all together. Servers for which the privacy budget is more stringent, contribute with the (potentially) smaller quantity $\bar{\varepsilon}_j$, where ρ corresponds to the probability in (23), established in the coupling relationship of Lemma 4.5.

The optimal division into these stringent and non-stringent privacy budgets is made by taking

$$S = \left\{ j \in [m] : \varepsilon_j \leq \sqrt{D/n_j} \right\},$$

in the sense that this choice of S minimizes the right-hand side of (24). With this choice of S, the bound on ρ established in Lemma 4.5 and the fact that by the construction of $\tilde{f}$ we have $\|\tilde{f} - f\|_1 = \gamma_D^{-1}\beta_D^{-1}\|g\|_1$ with $\gamma_D := c_0^{-1} D^{\nu}$ and $\beta_D = \gamma_D^{1/\nu}$, we obtain that

$$\sum_{j\in S} \bar{\varepsilon}_j \left(e^{\bar{\varepsilon}_j} - 1\right) \lesssim \sum_{j\in S} n_j^2 \varepsilon_j^2 \rho^2 \leq \gamma_D^{-2}\beta_D^{-2}\|g\|_1^2 \sum_{j\in S} n_j^2 \varepsilon_j^2 = c_0^{2+2/\nu} D^{-2-2\nu} \|g\|_1^2 \sum_{j\in S} n_j^2 \varepsilon_j^2.$$

The bound $D_{\mathrm{KL}}(P_{\tilde{f}}; P_f) \lesssim \sigma^{-2}\|\tilde{f} - f\|_2^2$, which is obtained through standard calculations, combined with the fact that $\|\tilde{f} - f\|_2^2 \lesssim \gamma_D^{-2}\beta_D^{-1}\|g\|_2^2$ by construction,

$$\sum_{j\in S^c} n_j D_{\mathrm{KL}}(P_f; P_{\tilde{f}}) \lesssim c_0^{1+2/\nu} D^{-2-2\nu} \|g\|_2^2 \sum_{j\in S^c} n_j D.$$

In Section C.3 of the appendix, it is shown that if $\sum_{j\in S} n_j \delta_j = o(1)$ as is assumed in the theorem, it holds that $\sum_{j\in S} e^{\bar{\varepsilon}_j} n_j \delta_j \rho = o(1)$ for the particular choice of S as well. Per the choice of the set S and D satisfying (17), we obtain that for $c_0 < 1$,

$$\left\|\mathbb{P}_f^T - \mathbb{P}_{\tilde{f}}^T\right\|_{\mathrm{TV}} \leq C c_0^{1+2/\nu} + o(1).$$

By taking the constant c_0 sufficiently small, the conclusion of Theorem 4.4 follows by Le Cam's two point method (see e.g. Lemma 1 in [72]), which then yields that

$$\sup_{f\in\mathcal{B}_{p,q}^{\alpha,R}} \mathbb{E}_f \left(\hat{f}(x_0) - f(x_0)\right)^2 \gtrsim \left(\tilde{f}(x_0) - f_0(x_0)\right)^2 \gtrsim \gamma_n^{-2} g^2(0) \gtrsim D^{-2\nu},$$

proving the theorem.

5. Discussion. The findings in the present paper highlight the trade-off between statistical accuracy and privacy preservation within the context of federated nonparametric regression. The results under the heterogeneous setting quantify the degree to which the individual DP constraints, as well as the degree to which observations are distributed among the different servers, impact the statistical performance as measured by the minimax risk. Furthermore, we find that the influence of the privacy constraints on the optimal performance depends on the inferential task at hand, with global estimation of an unknown function being subject to a different performance impact than estimation of a function at a point. For each of these inferential tasks, we provide an estimation procedure that attains the optimal statistical performance up to a logarithmic factor.

One promising direction for future research is the exploration of adaptive estimation in the federated learning framework. While our paper characterizes the statistical performance for nonparametric regression in a heterogeneous setting, the estimation procedures in this paper assume knowledge about the regularity of the underlying function. However, in many real-world applications, the regularity is unknown and estimators that can adapt to the true underlying regularity are required to attain the best possible performance. while such adaptive techniques exist for problems without privacy constraints (see e.g. [15, 26]), the theoretical (im)possibilities of adaptation under privacy constraints are relatively understudied. Such adaptive techniques might serve dual purposes: they could potentially refine the statistical accuracy while also optimizing the DP constraints for individual servers, especially when the constraints are dynamic or based on real-time needs. We leave this for future work.

Another promising avenue for future research is nonparametric hypothesis testing. It is well known that testing in nonparametric settings is subject to different phenomena than estimation, see for example [45]. However, under privacy constraints, the theoretical best possible performance in nonparametric hypothesis testing is not well understood. Addressing this question complements our understanding of the estimation problem.

Funding. The research of Tony Cai was supported in part by NIH grants R01-GM123056 and R01-GM129781.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal Federated Learning for Nonparametric Regression with Heterogenous Distributed Differential Privacy Constraints"

In the supplement to this paper, we present the detailed proofs for the main theorems in the paper "Optimal Federated Learning for Nonparametric Regression with Heterogenous Distributed Differential Privacy Constraints".

REFERENCES

[1] ACHARYA, J., BONAWITZ, K., KAIROUZ, P., RAMAGE, D. and SUN, Z. (2020). Context Aware Local Differential Privacy. In *Proceedings of the 37th International Conference on Machine Learning* (H. D. III and A. SINGH, eds.). *Proceedings of Machine Learning Research* **119** 52–62. PMLR.

[2] ACHARYA, J., CANONNE, C. L. and TYAGI, H. (2020). Distributed Signal Detection under Communication Constraints. In *Proceedings of Thirty Third Conference on Learning Theory* (J. ABERNETHY and S. AGARWAL, eds.). *Proceedings of Machine Learning Research* **125** 41–63. PMLR.

[3] ACHARYA, J., LIU, Y. and SUN, Z. (2023). Discrete Distribution Estimation under User-level Local Differential Privacy. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* (F. RUIZ, J. DY and J.-W. VAN DE MEENT, eds.). *Proceedings of Machine Learning Research* **206** 8561–8585. PMLR.

[4] ACHARYA, J., SUN, Z. and ZHANG, H. (2018). Differentially Private Testing of Identity and Closeness of Discrete Distributions. In *Advances in Neural Information Processing Systems* (S. BENGIO, H. WALLACH, H. LAROCHELLE, K. GRAUMAN, N. CESA-BIANCHI and R. GARNETT, eds.) **31**. Curran Associates, Inc.

[5] ARACHCHIGE, P. C. M., BERTOK, P., KHALIL, I., LIU, D., CAMTEPE, S. and ATIQUZZAMAN, M. (2019). Local differential privacy for deep learning. *IEEE Internet of Things Journal* **7** 5827–5842.

[6] BARNES, L. P., CHEN, W.-N. and ÖZGÜR, A. (2020). Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory* **1** 645–659.

[7] BARNES, L. P., HAN, Y. and OZGUR, A. (2020). Lower bounds for learning distributions under communication constraints via Fisher information. *Journal of Machine Learning Research* **21** 1–30.

[8] BASSILY, R., SMITH, A. and THAKURTA, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science* 464–473. IEEE.

[9] BEAUFAYS, F., RAO, K., MATHEWS, R. and RAMASWAMY, S. (2019). Federated learning for emoji prediction in a mobile keyboard.

[10] BRAVERMAN, M., GARG, A., MA, T., NGUYEN, H. L. and WOODRUFF, D. P. (2016). 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* 1011–1020.

[11] BUN, M., ULLMAN, J. and VADHAN, S. (2014). Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing* 1–10.

[12] BUTUCEA, C., DUBOIS, A., KROLL, M. and SAUMARD, A. (2020). Local differential privacy: Elbow effect in optimal density estimation and adaptation over Besov ellipsoids. *Bernoulli* **26** 1727 – 1764. https://doi.org/10.3150/19-BEJ1165

[13] CAI, T. T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *The Annals of Statistics* **27** 898–924.

[14] CAI, T. T. (2002). On block thresholding in wavelet regression: Adaptivity, block size, and threshold level. *Statistica Sinica* **12** 1241-1274.

[15] CAI, T. T. (2003). Rates of convergence and adaptation over Besov spaces under pointwise risk. *Statistica Sinica* 881–902.

[16] CAI, T. T., CHAKRABORTY, A. and VUURSTEEN, L. (2023). Supplement to "Optimal Federated Learning for Nonparametric Regression with Heterogenous Distributed Differential Privacy Constraints".

[17] CAI, T. T., WANG, Y. and ZHANG, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics* **49** 2825–2850.

[18] CAI, T. T., WANG, Y. and ZHANG, L. (2023). Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152*.

[19] CAI, T. T. and WEI, H. (2022). Distributed nonparametric regression: Optimal rate of convergence and cost of adaptation. *The Annals of Statistics* **50** 698-725.

[20] CAI, T. T. and WEI, H. (2023). Distributed Gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *Journal of Machine Learning Research* **to appear**.

[21] CANONNE, C. L. and SUN, Y. (2023). Private Distribution Testing with Heterogeneous Constraints: Your Epsilon Might Not Be Mine. *arXiv preprint arXiv:2309.06068*.

[22] CHAUDHURI, S. and COURTADE, T. A. (2023). Mean Estimation Under Heterogeneous Privacy: Some Privacy Can Be Free. *arXiv preprint arXiv:2305.09668*.

[23] COHEN, A., DAUBECHIES, I. and VIAL, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and computational harmonic analysis*.

[24] DAUBECHIES, I. (1992). *Ten lectures on wavelets*. SIAM.

[25] DING, B., KULKARNI, J. and YEKHANIN, S. (2017). Collecting telemetry data privately. *Advances in Neural Information Processing Systems* **30**.

[26] DONOHO, D. L. and JOHNSTONE, I. M. (1998). Minimax estimation via wavelet shrinkage. *The annals of Statistics* **26** 879–921.

[27] DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2013). Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science* 429–438. IEEE.

[28] DUCHI, J. C., JORDAN, M. I. and WAINWRIGHT, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association* **113** 182–201.

[29] DUCHI, J. C., JORDAN, M. I., WAINWRIGHT, M. J. and ZHANG, Y. (2014). Optimality guarantees for distributed statistical estimation. *arXiv:1405.0782 [cs, math, stat]*. arXiv: 1405.0782.

[30] DWORK, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming* 1–12. Springer.

[31] DWORK, C., ROTH, A. et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* **9** 211–407.

[32] DWORK, C. and SMITH, A. (2010). Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality* **1**.

[33] DWORK, C., SMITH, A., STEINKE, T. and ULLMAN, J. (2017). Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application* **4** 61–84.

[34] DWORK, C., SMITH, A., STEINKE, T., ULLMAN, J. and VADHAN, S. (2015). Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science* 650–669. IEEE.

[35] DWORK, C., TALWAR, K., THAKURTA, A. and ZHANG, L. (2014). Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing* 11–20.

[36] ERLINGSSON, U., PIHUR, V. and KOROLOVA, A. (2014). RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. *CCS '14* 1054–1067. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/2660267.2660348

[37] FALLAH, A., MAKHDOUMI, A., MALEKIAN, A. and OZDAGLAR, A. (2023). Optimal and differentially private data acquisition: Central and local mechanisms. *Operations Research*.

[38] GILL, R. D. and LEVIT, B. Y. (1995). Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli* 59–79.

[39] GINE, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781107337862

[40] HALL, P., KERKYACHARIAN, G. and PICARD, D. (1999). On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica* 33–49.

[41] HAN, Y., ÖZGÜR, A. and WEISSMAN, T. (2018). Geometric lower bounds for distributed parameter estimation under communication constraints. In *Conference On Learning Theory* 3163–3188. PMLR.

[42] HARD, A., RAO, K., MATHEWS, R., RAMASWAMY, S., BEAUFAYS, F., AUGENSTEIN, S., EICHNER, H., KIDDON, C. and RAMAGE, D. (2018). Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*.

[43] HÄRDLE, W., KERKYACHARIAN, G., PICARD, D. and TSYBAKOV, A. (2012). *Wavelets, approximation, and statistical applications* **129**. Springer Science & Business Media.

[44] IBRAGIMOV, I. and KHASMINSKII, R. (1997). *Some Estimation Problems in Infinite Dimensional Gaussian White Noise* In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* 259–274. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4612-1880-7_16

[45] INGSTER, Y. I. and SUSLINA, I. A. (2003). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. *Lecture Notes in Statistics* **169**. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-21580-8

[46] JOHNSTONE, I. M. (2019). *Function Estimation and Gaussian Sequence Models*. Unpublished manuscript.

[47] KAMATH, G., LI, J., SINGHAL, V. and ULLMAN, J. (2019). Privately learning high-dimensional distributions. In *Conference on Learning Theory* 1853–1902. PMLR.

[48] KAMATH, G., SINGHAL, V. and ULLMAN, J. (2020). Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory* 2204–2235. PMLR.

[49] KARWA, V. and VADHAN, S. (2017). Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*.

[50] KONEČNÝ, J., MCMAHAN, H. B., RAMAGE, D. and RICHTÁRIK, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

[51] KROLL, M. (2021). On density estimation at a fixed point under local differential privacy. *Electronic Journal of Statistics* **15** 1783 – 1813. https://doi.org/10.1214/21-EJS1830

[52] LECAM, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics* 38–53.

[53] LEVY, D., SUN, Z., AMIN, K., KALE, S., KULESZA, A., MOHRI, M. and SURESH, A. T. (2021). Learning with user-level privacy. *Advances in Neural Information Processing Systems* **34** 12466–12479.

[54] LI, T., SAHU, A. K., TALWALKAR, A. and SMITH, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine* **37** 50–60.

[55] LIU, Y., SURESH, A. T., YU, F. X. X., KUMAR, S. and RILEY, M. (2020). Learning discrete distributions: user vs item-level privacy. *Advances in Neural Information Processing Systems* **33** 20965–20976.

[56] NARAYANAN, S. (2022). Private high-dimensional hypothesis testing. In *Conference on Learning Theory* 3979–4027. PMLR.

[57] NARAYANAN, S., MIRROKNI, V. and ESFANDIARI, H. (2022). Tight and robust private mean estimation with few users. In *International Conference on Machine Learning* 16383–16412. PMLR.

[58] NGUYEN, A., DO, T., TRAN, M., NGUYEN, B. X., DUONG, C., PHAN, T., TJIPUTRA, E. and TRAN, Q. D. (2022). Deep federated learning for autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)* 1824–1830. IEEE.

[59] RAGINSKY, M. (2016). Strong data processing inequalities and \Phi-Sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory* **62** 3355–3389.

[60] RODRIGUEZ, I. M., SEXTON12, W. N., SINGER, P. E. and VILHUBER, L. The modernization of statistical disclosure limitation at the US Census Bureau.

[61] SART, M. (2023). Density estimation under local differential privacy and Hellinger loss. *Bernoulli* **29** 2318 – 2341. https://doi.org/10.3150/22-BEJ1543

[62] SMITH, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing* 813–822.

[63] SZABO, B. and VAN ZANTEN, H. (2020). Adaptive distributed methods under communication constraints. *The Annals of Statistics* **48** 2347–2380.

[64] SZABÓ, B. and VAN ZANTEN, H. (2022). Distributed function estimation: Adaptation using minimal communication. *Mathematical Statistics and Learning* **5** 159–199.

[65] SZABÓ, B., VUURSTEEN, L. and VAN ZANTEN, H. (2022). Optimal distributed composite testing in high-dimensional Gaussian models with 1-bit communication. *IEEE Transactions on Information Theory* **68** 4070–4084.

[66] SZABÓ, B., VUURSTEEN, L. and VAN ZANTEN, H. (2023). Optimal high-dimensional and nonparametric distributed testing under communication constraints. *The Annals of Statistics* **51** 909–934.

[67] TEAM, A. et al. (2017). Learning with privacy at scale. *Apple Mach. Learn. J* **1** 1–25.

[68] THORISSON, H. (2000). *Coupling, Stationarity, and Regeneration*. *Probability and Its Applications*. Springer New York.

[69] TRIEBEL, H. (1992). *Theory of Function Spaces II*. *Monographs in mathematics*. Springer.

[70] WASSERMAN, L. and ZHOU, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association* **105** 375–389.

[71] YE, M. and BARG, A. (2018). Optimal Schemes for Discrete Distribution Estimation Under Locally Differential Privacy. *IEEE Transactions on Information Theory* **64** 5662-5676. https://doi.org/10.1109/TIT.2018.2809790

[72] YU, B. (1997). *Assouad, Fano, and Le Cam* In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* 423–435. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4612-1880-7_29

[73] ZHU, Y. and LAFFERTY, J. (2018). Distributed Nonparametric Regression under Communication Constraints. In *Proceedings of the 35th International Conference on Machine Learning* (J. DY and A. KRAUSE, eds.). *Proceedings of Machine Learning Research* **80** 6009–6017. PMLR.