Differentially Private Functional Linear regression: Optimal Algorithms and Theoretical Guarantees

Yicheng Li and T. Tony Cai
Tsinghua University and University of Pennsylvania

September 3, 2025

Abstract

We study functional linear regression under differential privacy in both fully observed and discretely observed settings. For the fully observed case, we propose a private algorithm that integrates private covariance estimation with noisy gradient descent, and establish its convergence rate. We further derive a matching minimax lower bound, demonstrating that the proposed estimator is minimax-optimal up to a logarithmic factor. In the discrete setting, we analyze the trade-offs among privacy, statistical accuracy, and discretization error, revealing a notable phase transition phenomenon. Simulations and real-data applications confirm the effectiveness of our approach in preserving privacy while maintaining statistical accuracy.

Differential privacy in functional linear regression presents unique challenges due to the infinite-dimensional structure of functional data, which complicates sensitivity analysis, noise calibration, and algorithm design. Standard privacy mechanisms developed for finite-dimensional settings are inadequate in this context. To address these difficulties, we use basis expansions to reduce the problem to high-dimensional linear regression and introduce a preconditioning strategy that ensures robustness, even under ill-conditioned covariance structures. This framework broadens the applicability of differential privacy to complex functional models and ill-posed settings, advancing privacy-preserving methods in infinite-dimensional data analysis.

1 Introduction

In modern data analysis—where vast quantities of data are generated, collected, and processed—preserving individual privacy has become a central and urgent concern. The proliferation of data in sensitive domains such as healthcare, finance, education, and online behavior demand statistical methods that ensure both accuracy and privacy protection.

In healthcare, for example, electronic health records and genomic datasets provide powerful resources for personalized medicine and disease risk prediction. However, these data often contain sensitive personal and biological details, and any disclosure can lead to discrimination, stigmatization, or breaches of medical confidentiality. In finance, transaction records, credit histories, and market data are critical for risk modeling and fraud detection, yet they also reveal highly personal or proprietary information. Similarly, in education, student performance data can support personalized instruction and policy evaluation but may expose vulnerable populations if not properly protected.

In all of these contexts, the ability to perform statistical analysis and machine learning under formal privacy constraints is not just a technical challenge—it is an ethical imperative. As data becomes increasingly central to both scientific discovery and decision-making in society, privacy-preserving analysis methods are needed to ensure that data utility is not achieved at the expense of individual rights.

Among the various privacy-preserving methodologies developed to address this challenge, differential privacy (DP) has emerged as a leading and mathematically principled framework. It provides formal, quantifiable guarantees that the inclusion or exclusion of any single individual's data has a limited impact on the outcome of the analysis [Dwork et al., 2006, Dwork and Roth, 2014]. This property is especially vital in applications involving sensitive data such as medical records, financial transactions, or behavioral data, where even indirect disclosure can result in serious ethical, legal, and societal consequences [Abadi et al., 2016, Hawes, 2020, Adnan et al., 2022].

At the core of DP is the introduction of carefully calibrated random noise into statistical procedures [Dwork and Roth, 2014]. While this protects individual-level information, it also introduces a trade-off between privacy and statistical accuracy—commonly referred to as the "cost of privacy" [Cai et al., 2021, 2023]. Understanding and optimizing this trade-off is a central goal of ongoing research and has led to a rich body of work developing DP

algorithms for a wide range of statistical and machine learning tasks. These developments span a wide range of statistical and machine learning problems, including but not limited to: mean estimation [Barber and Duchi, 2014, Kamath et al., 2019, 2021, Alabi et al., 2023, Biswas et al., 2020], covariance estimation [Cai et al., 2024c, Amin et al., 2019, Dong et al., 2022], linear and nonparametric regression [Cai et al., 2021, Liu et al., 2023, Cai et al., 2024b], ranking [Cai et al., 2025], statistical inference [Dwork et al., 2021, Kim and Schrab, 2024, Cai et al., 2024a], and deep learning [Abadi et al., 2016, Phan et al., 2016].

While the theory of DP has been actively and relatively well studied in finite-dimensional settings, new challenges are emerging in more complex data environments. In particular, functional data—where each observation is a curve or a function rather than a finite-dimensional vector—are becoming increasingly prevalent due to advances in biomedical imaging, chemometrics, wearable technologies, and econometrics [Ramsay and Silverman, 2005, Hsing and Eubank, 2015]. These data types are often rich in structure and high in resolution, offering valuable insights into temporal or spatial processes, but they also raise new questions regarding how to quantify and preserve privacy in infinite-dimensional spaces.

Within this context, functional linear regression (FLR) has emerged as a foundational modeling framework, allowing researchers to explore the relationship between scalar outcomes and function-valued predictors. FLR is widely used in applications ranging from growth curve analysis to spectroscopic data interpretation. Unlike classical linear regression, which operates in a finite-dimensional Euclidean space, FLR is situated in an infinite-dimensional Hilbert space, where predictors are entire functions, such as time series or spatial trajectories. This shift in structure brings substantial theoretical and practical complications, especially when privacy guarantees must be imposed. Standard DP techniques, designed for vector-valued data, are often ill-suited for this setting, requiring new approaches to sensitivity analysis, noise calibration, and algorithmic design that respect

both the geometry of functional spaces and the principles of DP.

In this paper, we study functional linear regression (FLR) under differential privacy (DP) constraints. While the non-private theory of FLR is well-established [Yuan and Cai, 2010, Cai and Yuan, 2012, Zhou et al., 2023], ensuring DP in this setting presents unique challenges due to the infinite-dimensional nature of functional data. Addressing this gap involves overcoming two key technical hurdles:

- *Designing noise mechanisms for infinite-dimensional data:* Traditional DP mechanisms—such as Gaussian or Laplace noise—rely on well-defined sensitivity measures. The absence of a natural, bounded sensitivity notion for function-valued inputs makes it nontrivial to adapt existing privacy frameworks.

- *Handling the ill-posed nature of functional data:* Functional covariates often exhibit non-isotropic, ill-conditioned covariance structures with rapidly decaying eigenvalues. This amplifies the complexity of balancing privacy and accuracy, demanding new theoretical tools and methodological advances.

We develop a rigorous theoretical framework for differentially private functional linear regression, establishing minimax-optimal convergence rates under privacy constraints. Our proposed estimation and prediction procedures adapt to the structure of functional data while ensuring differential privacy. These results advance the growing field at the intersection of functional data analysis and privacy-preserving statistics, with implications for both theory and practice.

1.1 Problem Formulation

In modern statistical analysis, FLR serves as a key framework for modeling relationships between a functional covariate $X(t)$, observed over an interval T (e.g., $[0, 1]$), and a scalar

response y. The model is expressed as:

$$y = \int_T X(t)\beta^*(t)\,\mathrm{d}t + \varepsilon, \tag{1}$$

where $\beta^*(t)$ is the unknown functional coefficient to be estimated, $X(t)$ is a mean-zero Gaussian process with a continuous covariance function $k(s,t) = \mathbb{E}[X(s)X(t)]$, and ε is a mean-zero noise term independent of $X(t)$. The goal is to estimate the functional coefficient $\beta^*(t)$ or the associated functional $\zeta^*(X) = \int_T X(t)\beta^*(t)\,\mathrm{d}t$ based on observed data. In practice, $X(t)$ is typically observed at discrete points $\{t_{il}\}_{l=1}^m$ for each subject $i \in [n]$, resulting in the observed data: $z_{il} = X_i(t_{il}) + \varepsilon_{il},\ l \in [m]$, where ε_{il} are i.i.d. sub-Gaussian noises. The performance of an estimator $\hat{\zeta}(\cdot)$ is measured by the excess risk $\mathcal{R}(\hat{\zeta}, \zeta^*) = \mathbb{E}\left[\hat{\zeta}(X_{\text{new}}) - \zeta^*(X_{\text{new}})\right]^2$ over new observation X_{new}.

In many applications, such as healthcare and finance, functional data often contain sensitive personal information, raising significant privacy concerns. To address these challenges, DP has emerged as a leading framework for ensuring that statistical analyses safeguard individual data while still yielding meaningful insights. Originally introduced by Dwork et al. [2006], DP provides a rigorous mathematical foundation for quantifying privacy guarantees. The central idea is to ensure that the inclusion or exclusion of any single individual's data does not substantially affect the output of an algorithm, thereby protecting individuals from potential inference by adversaries [Dwork and Roth, 2014].

In this paper, we mainly adopt the ρ-zero concentrated DP framework [Dwork and Rothblum, 2016], which offers a rigorous and flexible privacy guarantee. Hereafter, we say two data sets S and S' are "adjacent" if they differ in only one element. We also recall that the α-Rényi divergence between two random variables X and Y associated with probability measures P and Q is defined as $D_\alpha(X\|Y) = \frac{1}{\alpha-1}\log \mathbb{E}\left(\frac{\mathrm{d}P}{\mathrm{d}Q}(X)\right)^\alpha$ for $\alpha > 1$.

Definition 1.1 (ρ-zCDP)**.** A randomized algorithm $M : \mathcal{D} \to \mathcal{A}$ satisfies ρ-zero concentrated DP (ρ-zCDP) if for all adjacent $S, S' \in \mathcal{D}$, $D_\alpha(M(S)\|M(S')) \le \rho\alpha$, $\forall \alpha \in (1, \infty)$,

where $D_\alpha(X\|Y)$ is taken only over the randomness of M conditioned on the data.

This Rényi divergence-based definition of DP ensures that an algorithm's output distribution changes minimally when a single data point is modified, thereby providing a privacy guarantee. The parameter ρ governs the privacy budget, with smaller values indicating stronger protection. A widely used alternative is (ϵ, δ)-DP [Dwork et al., 2006, Dwork and Roth, 2014].

Definition 1.2 ((ϵ, δ)-DP)**.** A randomized algorithm $M : \mathcal{D} \to \mathcal{A}$ satisfies (ϵ, δ)-DP if for all adjacent $S, S' \in \mathcal{D}$, and for all measurable $A \in \mathcal{A}$, we have $\mathbb{P}\{M(S) \in A\} \leq e^\epsilon \mathbb{P}\{M(S') \in A\} + \delta$, where $\mathbb{P}$ denotes the probability over the randomness of M.

While (ϵ, δ)-DP is more commonly cited in the literature, ρ-zCDP provides advantages such as tighter privacy budget composition and is particularly well-suited for iterative algorithms. The two measures are related: satisfying ρ-zCDP implies (ϵ, δ)-DP for appropriately chosen ϵ and δ, see Lemma 2.3 for details.

1.2 Our Contribution

In this paper, we extend the DP framework to FLR, advancing privacy-preserving analysis to FLR. Our work develops a novel approach that accommodates both fully observed functional covariates and discretely observed data, tackling real-world scenarios where functional data are sampled at discrete points [Cai and Yuan, 2011, Zhou et al., 2023]. By constructing algorithms that achieve minimax-optimal rates of convergence up to logarithmic factors, we establish fundamental trade-offs between privacy and statistical accuracy in infinite-dimensional data analysis. For the fully observed case, our theoretical analysis confirms that the proposed algorithm attains minimax optimality under DP constraints, shedding light on the inherent balance between privacy preservation and estimation accuracy. In the discrete observation setting, we identify a phase transition phenomenon [Cai

and Yuan, 2011]: as the number of sampling points increases, the discretization error diminishes, enabling the method to match the statistical performance of the fully observed case. This finding provides new insights on the sampling density required to optimize privacy and accuracy simultaneously.

A key component of our framework is a differentially private linear regression algorithm designed to handle general covariance structures. Unlike conventional methods that often assume well-conditioned covariance matrices [Cai et al., 2021], our approach employs a preconditioning technique with a dependency only logarithmic in the condition number of the covariance matrix. By integrating this preconditioner with noisy gradient descent, the algorithm remains robust even in highly ill-conditioned scenarios where eigenvalues decay significantly. This enhancement broadens the applicability of DP to realistic functional data settings and holds potential value for other statistical problems beyond this context.

Our analysis is based on ρ-zCDP, under which we derive sharp upper bounds on generalization error and matching minimax lower bounds. Notably, the privacy cost aligns directly with the budget ρ, in contrast to traditional (ϵ, δ)-DP, where a $\log(1/\delta)$ gap typically exists. This highlights the tighter privacy-utility trade-offs enabled by our approach.

The proposed method is computationally efficient and easy to implement. To complement the theory, we perform extensive simulations and apply our method to two benchmark functional datasets: the Tecator dataset [Febrero-Bande and Oviedo de la Fuente, 2012], which involves predicting fat, water, and protein content from meat absorbance spectra; and the wheat dataset [Kalivas, 1997], which uses near-infrared spectra to estimate protein content. Our results demonstrate that the proposed method achieves strong predictive performance under moderate privacy budgets, with accuracy improving as privacy constraints are relaxed. The Tecator results highlight consistent gains across multiple response variables, while the wheat dataset shows competitive performance even with sparse observations. Together, these findings confirm that our approach balances privacy and utility well

in practical functional data applications.

Collectively, our work provides a unified framework for differentially private FLR, delivering both theoretical insights and practical tools that advance privacy-preserving techniques for complex, infinite-dimensional data models. These contributions pave the way for future research into privacy-aware methodologies tailored to functional and high-dimensional statistical models.

1.3 Related Work

Differential privacy DP has emerged as a foundational framework for protecting individual privacy in data analysis [Dwork et al., 2006, Dwork and Roth, 2014]. While the most widely used formulation is (ϵ, δ)-DP [Dwork and Roth, 2014], several alternative definitions have been proposed to address specific challenges in various settings. Notably, concentrated DP and its variants offer refined privacy guarantees with improved composition properties [Dwork and Rothblum, 2016, Bun and Steinke, 2016]. DP is typically achieved by introducing random noise into statistical procedures, with common mechanisms including the Laplace mechanism and the Gaussian mechanism [Dwork and Roth, 2014]. These mechanisms provide formal protection by ensuring that the output distribution remains stable under small changes to the input data, thereby limiting the influence of any single individual's information.

Statistical analysis of DP algorithms DP has been widely applied to statistical and machine learning tasks, including mean estimation [Kamath et al., 2019, 2021, Alabi et al., 2023], linear regression [Cai et al., 2021, 2023, Talwar et al., 2015, Liu et al., 2023, Varshney et al., 2022], and other models [Cai et al., 2024b,c]. For linear regression under nearly isotropic covariates, Cai et al. [2021] established a minimax-optimal DP algorithm using noisy gradient descent. However, its optimality fails for general, non-isotropic covariance

structures. More recent works have extended DP linear regression to handle arbitrary covariances, but with limitations: some incur polynomial dependence on the condition number [Biswas et al., 2020, Varshney et al., 2022, Liu et al., 2023], while others achieve optimality only at the cost of exponential runtime [Liu et al., 2022, Asi et al., 2023].

Functional Linear Regression FLR is a well-established model in functional data analysis, designed to investigate the relationship between functional covariates and a scalar response [Ramsay and Silverman, 2005, Hsing and Eubank, 2015]. In the classical setting where functional covariates are fully observed, a rich body of work has established minimax-optimal rates for both estimation and prediction [Cai and Yuan, 2012, Yuan and Cai, 2010, Hall and Horowitz, 2007]. More recently, the FLR framework has been extended to accommodate more complex settings, including discretely observed functional data and other generalized regression models [Cai and Yuan, 2011, Zhang and Wang, 2016, Zhou et al., 2023, Wang et al., 2022, Yao et al., 2005].

Despite these advances, privacy concerns in FLR have received little attention. In many applications, such as health monitoring, wearable devices, and financial transactions, functional data often contain sensitive personal information that must be protected. This highlights the urgent need for privacy-preserving methods tailored to functional data, an area that remains largely unexplored and is the focus of the present work. Our work aims to develop the first theoretical and algorithmic framework for FLR under DP constraints.

1.4 Organization of the Paper

The remainder of the paper is organized as follows. In Section 2, we introduce differentially private methods for linear regression with general covariance structures, which serve as the building blocks of our FLR algorithm. Section 3 presents the proposed differentially private FLR estimators for both fully and discretely observed functional data. Numerical

studies are reported in Section 4. For reasons of space, the proofs of the main results and technical lemmas are deferred to the supplement, where we also provide additional numerical experiments and supporting technical details.

Notations We use $c, C, \ldots$ for positive constants that may vary by line. Write $a \lesssim b$ if $a \leq Cb$ for some C, and $a \asymp b$ if both $a \lesssim b$ and $b \lesssim a$. Let $[n] = \{1, \ldots, n\}$ for integer n. Let $L^2(T)$ (or L^2) be the space of square-integrable functions on T, with inner product $\langle \cdot, \cdot \rangle_{L^2}$; ℓ^2 is the space of square-summable sequences with inner product $\langle \cdot, \cdot \rangle_{\ell^2}$. For vector v, let $v_{\leq p}$ and $v_{>p}$ denote its first p and remaining entries, respectively. For matrix A, let $A_{\leq p}$ and $A_{>p}$ be its corresponding principal submatrices. Write $A \preceq B$ if $B - A$ is positive semi-definite, and $A \prec B$ if positive definite. Let $\|A\|_\infty$ denote the operator norm.

2 Differentially Private Linear Regression under General Covariance

FLR can be viewed as an infinite-dimensional linear regression problem. A key component in analyzing it under DP is the study of a finite-dimensional model with an ill-conditioned covariance structure. In this section, we address this problem, which is of independent interest. Specifically, we consider:

$$y = \langle u, \theta^* \rangle + \varepsilon, \tag{2}$$

where $u \in \mathbb{R}^p$ is the covariate, $\theta^* \in \mathbb{R}^p$ the unknown parameter, and ε the noise. Given n i.i.d. samples (u_i, y_i), our goal is to construct a differentially private estimator for θ^*.

In standard differentially private linear regression (e.g., Cai et al. [2021]), the covariance matrix Σ of u is typically assumed to be well-conditioned. In contrast, FLR often involves a decaying spectrum, making Σ ill-conditioned and standard methods suboptimal. Addressing this is crucial for functional data applications.

We propose a differentially private algorithm that handles general, possibly ill-conditioned, covariance structures with non-isotropic and decaying eigenvalues. While recent works [Liu et al., 2023, Varshney et al., 2022] relax the identity covariance assumption, they suffer from either poor privacy-accuracy trade-offs or exponential runtime. In contrast, our method achieves minimax optimality (up to logarithmic factors) with polynomial-time complexity.

The key idea is to privately estimate Σ and construct a *preconditioner matrix* A so that transformed covariates $v = Au$ have approximately identity covariance. We then apply private regression on v, improving both privacy and accuracy. By addressing the challenges of general covariance structures, our method provides a foundation for privacy-preserving inference in infinite-dimensional models such as FLR, where standard assumptions fail.

2.1 General DP Mechanisms

We begin by introducing core mechanisms for achieving DP, which underpin many privacy-preserving algorithms [Dwork and Roth, 2014]. The standard approach is to add carefully calibrated noise to the output of a function with bounded sensitivity. For ρ-zCDP, this is typically done via the Gaussian mechanism, as shown in the following lemma [Dwork and Roth, 2014, Bun and Steinke, 2016].

Lemma 2.1 (Gaussian Mechanism)**.** *Let $f : \mathcal{D} \to \mathbb{R}^d$ be an algorithm. Suppose that $\sup_{D,D' \text{ adjacent}} \|f(D) - f(D')\|_2 \leq \Delta$. Then, adding Gaussian random noise $w \sim N(0, \sigma^2 I_d)$ with $\sigma^2 = \frac{\Delta^2}{2\rho}$, the algorithm $M_f(D) = f(D) + w$ is ρ-zCDP.*

Furthermore, the composition property facilitates the construction of complex differentially private algorithms by combining simpler ones, as formalized in the next lemma [Dwork and Roth, 2014, Bun and Steinke, 2016].

Lemma 2.2 (Composition)**.** *If $M : \mathcal{D} \to \mathcal{Y}$ is ρ-zCDP and $f : \mathcal{Y} \to \mathcal{Z}$ is an arbitrary algorithm, then $f \circ M$ is also ρ-zCDP. If $M_1, \ldots, M_T$ are $\rho_1, \ldots, \rho_T$-zCDP and M is a*

function of $M_1, \ldots, M_T$*, then* M *is* ρ*-zCDP for* $\rho = \sum_t \rho_t$*.*

The following lemma shows that ρ-zCDP implies the more commonly used (ϵ, δ)-DP under certain conditions [Dwork and Roth, 2014, Bun and Steinke, 2016].

Lemma 2.3. *If* M *satisfies* $(\epsilon, 0)$*-DP, then* M *is* $(\epsilon^2/2)$*-zCDP. Conversely, If* M *satisfies* ρ*-zCDP, then* M *satisfies* $\left(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta\right)$*-DP for every* $\delta \in (0, 1)$*. Particularly, for* $\epsilon \in (0, 1]$ *and* $\delta \in (0, e^{-1}]$*,* ρ*-zCDP with* $\rho \leq \epsilon^2 (8 \log(1/\delta))^{-1}$ *implies* (ϵ, δ)*-DP.*

Although (ϵ, δ)-DP remains the more standard framework, ρ-zCDP offers distinct advantages for designing intricate algorithms, notably through its single-parameter control of the privacy budget, which often yields tighter guarantees. Accordingly, this paper adopts ρ-zCDP as the primary privacy measure, with corresponding (ϵ, δ)-DP results derivable as needed using the relationship above.

2.2 Covariance Matrix Estimation

The first step of our approach is to obtain a differentially private covariance estimator $\tilde{\Sigma}$ which closely approximates the true covariance matrix $\Sigma = \mathbb{E}uu^\top$ in a relative sense. Specifically, we aim to satisfy the double-sided bound $cI_p \preceq \Sigma^{-\frac{1}{2}} \tilde{\Sigma} \Sigma^{-\frac{1}{2}} \preceq CI_p$ for some constants c, C.

A key challenge is that the naive approach of adding isotropic noise to enforce DP disrupts the covariance structure. If the condition number of Σ is large, this method fails to achieve the desired bound. While existing works [Kamath et al., 2019, Biswas et al., 2020] explore differentially private covariance estimation under similar conditions, we introduce an alternative approach tailored to our needs, enabling us to handle more general settings relevant to FLR.

Our approach leverages an iterative refinement process: given an initial preconditioner (or covariance estimate), we construct a differentially private improvement that enhances

the estimate by a constant factor. By repeating this process, we generate a sequence of differentially private, regularized preconditioners A_t that ultimately satisfy our requirement. This iterative strategy shares similarities with Kamath et al. [2019] but incorporates explicit regularization and a simplified algorithm. The procedure is detailed in Algorithm 1 (one-step improvement) and Algorithm 2.

Algorithm 1 DP Covariance Estimation with Preconditioner: $\mathsf{DPCovEstP}(U; A_0, L, \rho)$

Input: Data matrix $U = (u_1, \ldots, u_n) \in \mathbb{R}^{p\times n}$, positive definite preconditioner $A_0 \succ 0$, truncation level $L > 0$, privacy budget ρ.

Output: ρ-zCDP covariance estimator $\tilde{\Sigma}$.

Compute $v_i = A_0 u_i$ for $i = 1, \ldots, n$, and set $S = \{i \in [n] \mid \|v_i\|_2 \le L\}$.

Compute $\hat{V} = \frac{1}{n}\sum_{i\in S} v_i v_i^\top$ and $\tilde{V} = \hat{V} + \sigma N$, where N is a $p \times p$ symmetric matrix with i.i.d. standard Gaussian entries in the upper triangular part and $\sigma^2 = \frac{2L^4}{\rho n^2}$.

Return $\tilde{\Sigma} = A_0^{-1}\tilde{V}A_0^{-1}$.

Algorithm 2 DP Iterative Preconditioner Estimation: $\mathsf{DPPrecondEstItr}(U; T, L, \rho)$

Input: Data matrix $U = (u_1, \ldots, u_n) \in \mathbb{R}^{p\times n}$, iteration number T, truncation level L, privacy budget ρ.

Output: ρ-zCDP regularized preconditioner A_T.

Initialize $A_0 = I_p$, $\lambda_0 = 1$.

for $t = 1$ to T **do**

Set $\lambda_t = \lambda_{t-1}/2 = 2^{-t}$.

Compute $\tilde{\Sigma}_t = \mathsf{DPCovEstP}(U; A_{t-1}, L, \rho/T)$ and $A_t = \left(\tilde{\Sigma}_t + \lambda_t\right)^{-1/2}$.

end for

Return A_T.

We first establish the privacy guarantee of the covariance estimator. The proof is deferred to the appendix.

Proposition 2.4. Both Algorithm 1 and Algorithm 2 are ρ-zCDP.

We assume a standard sub-Gaussian condition on the covariates [Biswas et al., 2020, Cai et al., 2021], with sub-Gaussian norm defined as $\|X\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \le 2\}$.

Assumption 2.5. The covariance matrix Σ is invertible and the normalized covariate $\Sigma^{-\frac{1}{2}}u$ is K-sub-Gaussian vector, that is, $\left\|\left\langle \Sigma^{-\frac{1}{2}}u, w\right\rangle\right\|_{\psi_2} \le K\|w\|_2, \quad \forall w \in \mathbb{R}^p$. Moreover, $\operatorname{Tr}\Sigma \le C_{\mathrm{tr}}$ for some constant C_{tr}.

Theorem 2.6 (Differentially Private Exact Preconditioner)**.** *Under Assumption 2.5, let* $\lambda_{\min} = \lambda_{\min}(\Sigma)$. *Suppose that* $n^2 \ge C_1\rho^{-1}p^3\log\lambda_{\min}^{-1}$, $n \ge C_2 p$ *hold for some constants* $C_1, C_2 > 0$. *Then, by applying Algorithm 2 with parameters* $T \asymp \log_2\lambda_{\min}^{-1}$, $L \asymp \sqrt{p}$, *with probability at least* $1-C(n+\log\lambda_{\min}^{-1})\exp(-cp)$, *the output* A_T *satisfies* $\frac{1}{3}I_p \preceq A_T\Sigma A_T \preceq 2I_p$.

Theorem 2.6 shows that a differentially private preconditioner A_T can be obtained when the sample size satisfies $n^2 \gtrsim \rho^{-1}p^3\log\lambda_{\min}^{-1}$, comparable to prior results [Biswas et al., 2020, Kamath et al., 2019]. Notably, the algorithm depends only logarithmically on the minimum eigenvalue $\lambda_{\min}$, making it well-suited for covariance matrices with decaying spectra, as commonly seen in FLR.

2.3 Linear Regression with General Covariance Matrix

Let us now turn to the linear regression model in (2). The key idea is that, given a preconditioner A such that $V \coloneqq A\Sigma A \approx I_p$, we can transform the covariates and parameters via $v = Au$ and parameter $\eta = A^{-1}\theta$. Under this transformation, the model becomes $y = \langle v, \eta^*\rangle + \varepsilon$. This allows us to apply a noisy gradient descent algorithm, such as the one proposed in Cai et al. [2021], using the transformed data v. However, the transformation requires us to use a modified projection step—rather than standard clipping—to ensure the privacy guarantee is preserved.

Let us denote by $\mathrm{Trunc}_{L_\mathrm{y}}(x)$ the truncation operator that truncates the value of x to $[-L_\mathrm{y}, L_\mathrm{y}]$ and by $\mathrm{Proj}_{\mathcal{C}}(x)$ the projection operator onto the convex set $\mathcal{C}$, namely $\mathrm{Proj}_{\mathcal{C}}(x) = \arg\min_{y\in\mathcal{C}} \|x-y\|_2$. We introduce the following preconditioned differentially private linear regression algorithm described in Algorithm 3 and Algorithm 4.

Algorithm 3 Preconditioned Differentially Private Gradient Descent

Input: Data $(u_1, y_1), \ldots, (u_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, preconditioner $A \succ 0$, truncation levels $L, L_\mathrm{y} > 0$, feasibility parameter R, iteration number T, step size r, privacy budget ρ.

Output: ρ-zCDP estimator $\tilde{\theta}$.

Compute $v_i = Au_i$ for $i = 1, \ldots, n$, and set $S = \{i \in [n] : \|v_i\|_2 \le L\}$, convex set $\mathcal{C} = \{x \in \mathbb{R}^p : \|Ax\|_2 \le R\}$.

Initialize $\eta_0 = \mathbf{0} \in \mathbb{R}^p$, $\sigma_\mathrm{w}^2 = \frac{2L^2 L_\mathrm{y}^2 T}{\rho n^2}$.

for $t = 0$ to $T-1$ **do**

Compute $\tilde{\eta}_{t+1} = \mathrm{Proj}_{\mathcal{C}}(\tilde{\eta}'_{t+1})$, where $\tilde{\eta}'_{t+1} = \tilde{\eta}_t - \frac{r}{n}\sum_{i\in S} \mathrm{Trunc}_{L_\mathrm{y}}(v_i^\top \tilde{\eta}_t - y_i)v_i + r\sigma_\mathrm{w} w_t$, $w_t \sim N(0, I_p)$.

end for

Return $A\tilde{\eta}_T$.

First, we state the privacy guarantee of the proposed algorithms. The proof follows standard arguments and is deferred to the supplementary material.

Proposition 2.7. Algorithm 3 is ρ-zCDP.

Regarding the generalization guarantee, we introduce the following assumptions on the source condition and the noise, both of which are standard in the linear regression literature.

Assumption 2.8 (Source condition)**.** There exists some constant R such that $\|\theta^*\|_2 \le R$.

Assumption 2.9 (Noise)**.** The noise ε is σ^2-sub-Gaussian and there exists some $\mu \ge 0$ such that $\|\mathbb{E}(\varepsilon u)\|_2 \le \mu$.

Algorithm 4 Differentially Private Linear Regression

Input: Data $(u_1, y_1), \ldots, (u_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$, truncation levels $L, L_{\mathrm{y}} > 0$, feasibility parameter R, iteration numbers T_1, T_2, step size r, privacy budget ρ.

Output: ρ-zCDP estimator $\tilde{\theta}$.

Apply Algorithm 2 to half of the samples with privacy budget $\rho/2$ and iteration number T_1 to obtain preconditioner $\tilde{A}$.

Apply Algorithm 3 to the remaining samples with preconditioner $\tilde{A}$ and privacy budget $\rho/2$ to compute $\tilde{\theta}$.

Return $\tilde{\theta}$.

Now, the following theorem shows the error bound of the output of Algorithm 3 provided that a suitable preconditioner is given.

Theorem 2.10. *Under Assumptions 2.5, 2.8, and 2.9, suppose that the condition* $\frac{1}{3} I_p \preceq V := A\Sigma A \preceq 2 I_p$ *holds. Consider Algorithm 3 with parameters* $L \asymp \sqrt{p}$, $L_{\mathrm{y}} \asymp \sqrt{(R^2 + \sigma^2) \log n}$, *and* $\sigma_{\mathrm{w}}^2 = \frac{2L^2 L_{\mathrm{y}}^2}{(\rho/T) n^2} \asymp \frac{p(R^2+\sigma^2)(\log n)^2}{\rho n^2}$. *Then, provided that* $n \geq Cp$, *with probability at least* $1 - Cn\exp(-p) - Cn\exp(-C\log n)$, *the output* $\tilde{\theta}$ *satisfies*

$$\left\| \Sigma^{1/2} (\tilde{\theta} - \theta^*) \right\|_2^2 \lesssim \frac{p\sigma^2}{n} + p\sigma_{\mathrm{w}}^2 + \mu^2. \tag{3}$$

The proof of Theorem 2.10 is based on analyzing the gradient dynamics of the algorithm while carefully controlling the noise introduced by both the statistical variation and the privacy mechanism at each iteration. The complete proof is presented in the supplementary material. By combining Theorem 2.10 with Theorem 2.6, we derive the overall error bound for the differentially private linear regression algorithm.

Corollary 2.11. *Under Assumptions 2.5, 2.8, and 2.9, suppose that the conditions*

$$n^2 \geq C_1 \rho^{-1} p^3 \log \lambda_{\min}(\Sigma)^{-1}, \qquad n \geq C_2 p \tag{4}$$

are satisfied for some constants $C_1, C_2 > 0$. *Then, applying Algorithm* 4 *with parameters specified in Theorem* 2.6 *and Theorem* 2.10*, the output* $\tilde{\theta}$ *satisfies*

$$\left\|\Sigma^{\frac{1}{2}}(\tilde{\theta} - \theta^*)\right\|_2^2 \lesssim \frac{p\sigma^2}{n} + \frac{p^2(\log n)^2}{\rho n^2} + \mu^2$$

with probability at least $1 - C(n + \log \lambda_{\min}^{-1}) \exp(-cp) - Cn \exp(-C \log n)$.

Using Lemma 2.3, we translate the result to the (ϵ, δ)-DP setting. For completeness, we present the (ϵ, δ)-DP version below.

Corollary 2.12 ((ϵ, δ)-DP version of Corollary 2.11)**.** *Let* $\epsilon \in (0, 1]$ *and* $\delta \in (0, e^{-1}]$. *Under the same conditions as in Corollary* 2.11*, suppose that*

$$n^2 \geq C_1 \epsilon^{-2} p^3 \log \lambda_{\min}(\Sigma)^{-1} \log(1/\delta), \qquad n \geq C_2 p$$

hold for some constants $C_1, C_2 > 0$. *Then, by applying Algorithm* 4 *with* $\rho = \epsilon^2 (8 \log(1/\delta))^{-1}$*, the output* $\tilde{\theta}$ *satisfies*

$$\left\|\Sigma^{\frac{1}{2}}(\tilde{\theta} - \theta^*)\right\|_2^2 \lesssim \frac{p\sigma^2}{n} + \frac{p^2(\log n)^2 \log(1/\delta)}{\epsilon^2 n^2} + \mu^2.$$

The following theorem gives the minimax lower bound for the ρ-zCDP algorithms.

Theorem 2.13. *Let* $\rho \in (0, 1]$. *Denote by* $\mathcal{P}$ *the collection of all distributions of* (u, y) *satisfying Assumptions 2.5, 2.8, and 2.9. Suppose that* $n \gtrsim \rho^{-1} p^3$. *Then,*

$$\inf_{M \in \mathcal{M}_\rho} \sup_{\mathcal{P}} \mathbb{E}\left\|\Sigma^{\frac{1}{2}}(M(S) - \beta)\right\|_2^2 \gtrsim \sigma^2 \left(\frac{p}{n} + \frac{p^2}{\rho n^2}\right). \tag{5}$$

The proof of Theorem 2.13 builds on the fundamental idea of constructing a tracing attack, as introduced in Cai et al. [2021]. However, we present a novel analysis that directly addresses ρ-zCDP algorithms, enabling a more precise characterization of the lower bound. The complete proof is provided in the supplementary material.

2.4 Discussion

We now highlight key implications of our results.

Theorem 2.10 and Corollary 2.11 provide error bounds for our differentially private linear regression algorithm under general covariance. When the noise is independent of the covariate (i.e., $\mu = 0$), the bound in Corollary 2.11 matches the minimax lower bound in Theorem 2.13 up to logarithmic factors. To our knowledge, this is the first polynomial-time algorithm to achieve minimax-optimal error (up to log factors) for private linear regression with general covariance, improving on prior results [Cai et al., 2021, Liu et al., 2023, Varshney et al., 2022].

In contrast, existing polynomial-time methods [Liu et al., 2023, Varshney et al., 2022] incur polynomial dependence on the condition number, while statistically optimal algorithms [Liu et al., 2022, Asi et al., 2023] require exponential time. While our result serves as a building block for private functional regression, it is also of independent interest and broadly applicable.

Our analysis is conducted under the ρ-zCDP framework, but results can be translated to the (ϵ, δ)-DP setting via Corollary 2.12, introducing a multiplicative $\log(1/\delta)$ factor in the bounds. Under (ϵ, δ)-DP, Cai et al. [2021] established a lower bound of order $\frac{p^2}{\epsilon^2 n^2}$, but matching upper bounds typically include extra $\log(1/\delta)$ terms, leaving a gap.

By working directly with ρ-zCDP and refining the lower bound in Theorem 2.13, we close this gap and establish tight privacy-utility trade-offs without additional logarithmic penalties, yielding a sharper characterization of privacy cost under ρ-zCDP.

3 Differentially Private Functional Linear Regression

We now turn to the FLR model (1) with the aim to estimate the effect of a functional covariate X on a scalar response y under DP constraints. Let $\{(X_i(\cdot), y_i)\}_{i=1}^n$ be n i.i.d.

samples from the model (1). We distinguish between the following observation settings for the functional covariate $X(\cdot)$:

- *Full observation*: The entire path of the functional covariate $X(\cdot)$ is observed.
- *Discrete observation*: The $X(\cdot)$ is observed only at a finite number of time points with additive noise:

$$z_{il} = X_i(t_{il}) + \varepsilon_{il}, \quad i = 1, \ldots, n, \quad l = 1, \ldots, m, \tag{6}$$

  where $(t_{il})_{i,l}$ are observation points and ε_{il} are i.i.d. mean-zero σ_{x}^2-sub-Gaussian noises. The observed samples are $\{((z_{il}, t_{il})_{l=1}^m, y_i)\}_{i=1}^n$. We assume mutual independence among the process $X(\cdot)$, the observation points $(t_{il})_{i,l}$, and the noise terms $(\varepsilon_{il})_{i,l}$. We consider two designs for the observation points in this setting:

  - *Independent design*: The observation points $(t_{il})_{i,l}$ are i.i.d. samples drawn from the uniform distribution on the interval T.
  - *Common design*: The observation points are the same across all samples, i.e., $t_{1l} = \cdots = t_{ml} = t_l$ for $l = 1, \ldots, m$, where t_l are m common observation points drawn uniformly from $T = [0, 1]$.

  For DP, we treat the collection $((z_{il}, t_{il})_{l=1}^m, y_i)$ as a single data point. Thus, for adjacent datasets, one replaces the whole collection of discrete observations with another one.

As in linear regression, prediction is an important goal of FLR. That is, given a new independent input X_{new} from the same distribution as X, we wish to estimate the conditional mean $\zeta^*(X_{\mathrm{new}}) \coloneqq \int_T X_{\mathrm{new}}(t)\beta^*(t)\mathrm{d}t$. The performance of an estimator is measured by the excess risk

$$\mathcal{R}(\hat{\zeta}, \zeta^*) \coloneqq \mathbb{E}\left[\hat{\zeta}(X_{\mathrm{new}}) - \zeta^*(X_{\mathrm{new}})\right]^2 = \mathbb{E}\left[y - \hat{\zeta}(X_{\mathrm{new}})\right]^2 - \mathbb{E}\left[y - \zeta^*(X_{\mathrm{new}})\right]^2. \tag{7}$$

To construct differentially private estimators for the FLR model, let us first fix a neutral orthonormal basis $(e_j)_{j\geq 1}$ over $L^2(T)$, which can be chosen as the Fourier basis or the wavelet basis. Then, the functional coefficient can be expanded under this basis as $\beta^*(t) = \sum_{j\geq 1} \beta_j^* e_j(t)$, and we denote the coefficient vector $\boldsymbol{\beta}^* = (\beta_j^*)_{j\geq 1}$. We can also introduce the (infinite-dimensional) covariance matrix $\Sigma = (\Sigma_{jj'})$ defined as $\Sigma_{jj'} = \int_{T\times T} k(t,t')e_j(t)e_j(t')\mathrm{d}t\mathrm{d}t'$. Introducing the score vector $\boldsymbol{u} = \boldsymbol{u}(X) = (\langle X, e_1\rangle_{L^2}, \langle X, e_2\rangle, \dots)$, since $X(\cdot)$ is a mean-zero and Gaussian process with covariance function $k(s,t)$, we have $\boldsymbol{u} \sim N(0,\Sigma)$. Now, the FLR model can be written as an infinite-dimensional linear model:

$$y = \int_T X(t)\beta^*(t)\mathrm{d}t = \langle \boldsymbol{u}, \boldsymbol{\beta}^*\rangle_{\ell^2} + \varepsilon. \tag{8}$$

Suppose we have an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$, we can define the functional estimator $\hat{\zeta}(X) = \left\langle \boldsymbol{u}(X), \hat{\boldsymbol{\beta}} \right\rangle_{\ell^2}$. Then, the generalization error can be written as $\mathcal{R}(\hat{\zeta}, \zeta^*) = \left\| \Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right\|_2^2$. We note that $\left\| \Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right\|_2^2$ measures the prediction error, while $\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \right\|_2^2$ captures the estimation error of the functional coefficient. This work focuses on prediction error, but the same algorithm applies to estimation error with appropriate parameter scaling.

With (8), we can further consider a truncated model with first p basis functions

$$y = \left\langle \boldsymbol{u}_{\leq p}, \boldsymbol{\beta}^*_{\leq p} \right\rangle_{\mathbb{R}^p} + \varepsilon_p, \quad \varepsilon_p = \varepsilon + \left\langle \boldsymbol{u}_{>p}, \boldsymbol{\beta}^*_{>p} \right\rangle. \tag{9}$$

The truncation reduces the infinite-dimensional problem to a finite-dimensional one, so we can run the differentially private linear regression algorithm on the truncated finite dimensional model. Moreover, if we only have discrete observation, we can replace $\langle X, e_j\rangle_{L^2}$ by its empirical version. This leads to the following Algorithm 5, whose privacy guarantee can be obtained by the composition theorem.

The privacy guarantee of Algorithm 5 is easily deduced from that of Algorithm 4 and the composition theorem.

Proposition 3.1. Algorithm 5 is ρ-zCDP.

Algorithm 5 Differentially Private FLR via Truncation

Input: Functional covariates $X_1(\cdot), \ldots, X_n(\cdot)$ or discrete observations $\{(z_{il}, t_{il}), i \in [n], l \in [m]\}$, responses $y_1, \ldots, y_n$, truncation level $p \in \mathbb{N}$, privacy budget ρ, and parameters $L, L_{\mathrm{y}}, R, T_1, T_2, r$ as specified in Algorithm 4.

Output: ρ-zCDP functional coefficient estimator $\tilde{\boldsymbol{\beta}}$.

Compute $u_{ij} = \langle X_i, e_j \rangle_{L^2}$ or $u_{ij} = \frac{1}{m} \sum_{l=1}^{m} z_{il} e_j(t_{il})$ for $i \in [n]$ and $j \in [p]$; set $u_i = (u_{i1}, \ldots, u_{ip})^\top$.

Apply Algorithm 4 to the model (9) with data $\{(u_i, y_i)\}_{i=1}^n$ to obtain estimator $\tilde{\boldsymbol{\beta}}_{\leq p}$.

Return $\tilde{\boldsymbol{\beta}}$ by padding zeros to $\tilde{\boldsymbol{\beta}}_{\leq p}$.

3.1 Full Observation

Now let us consider the full observation setting. To establish the generalization error of the estimator in the FLR model, let us introduce the following function classes $\mathcal{F}(r, R, C)$ for $r \geq 1$ and $R, C > 0$, where we denote $\boldsymbol{\beta} = (\beta_j)_{j \geq 1}$:

$$\mathcal{F}(r, R, C) = \left\{ \beta(t) = \sum_{j \geq 1} \beta_j e_j(t) : \|\boldsymbol{\beta}\|_2 \leq R,\ \langle \boldsymbol{\beta}_{>p}, \Sigma_{>p} \boldsymbol{\beta}_{>p} \rangle \leq C p^{-r}, \forall p \geq 1 \right\}. \quad (10)$$

The function class $\mathcal{F}(r, R, C)$ characterizes the relative smoothness of β^* and Σ via tail decay under the chosen basis. It aligns with classical FLR assumptions [Cai and Yuan, 2012, Yuan and Cai, 2010] and holds with $r = \gamma + \alpha$ when the following decay conditions are met:

Condition 3.2. There are constants $\gamma > 1$, $\alpha > 0$ and $C_{\mathrm{eig}}, C_{\beta^*} > 0$ such that

(i) Smoothness of $X(\cdot)$: $\|\Sigma_{>p}\|_\infty = \lambda_{\max}(\Sigma_{>p}) \leq C_{\mathrm{eig}} p^{-\gamma}, \quad \forall p \geq 1.$

(ii) Smoothness of β^*: $\sum_{j \geq 1} j^\alpha (\beta_j^*)^2 \leq C_{\beta^*}$.

In particular, if we take the basis as the Fourier basis, the decay condition (ii) is satisfied when the functional coefficient β^* belong to Sobolev spaces $H^{\frac{\alpha}{2}}(T)$, and (i) is satisfied

when the covariance function $k(\cdot,\cdot)$ is diagonalized by the Fourier basis and the function $t \mapsto k(t,t)$ lies in $H^{\frac{\gamma-1}{2}}(T)$. Moreover, if the basis is taken as the eigenfunctions of the covariance operator, (i) is just the eigenvalue decay condition commonly used in the FLR literature [Cai and Yuan, 2012, Yuan and Cai, 2010]. In addition, we also have to introduce the following assumption on the regularity of the functional covariance.

Assumption 3.3. There exist positive constants c_{eig} and κ such that for all $p \geq 1$, $\lambda_{\min}(\Sigma_{\leq p}) \geq c_{\text{eig}} p^{-\kappa}$.

Assumption 3.3 assumption ensures that the covariance matrix Σ is well-posed in the sense that the minimal eigenvalues only decay in a polynomial rate. It can be seen as a lower-bound counterpart of (i) in Condition 3.2. If the basis is taken as the eigenfunctions of the covariance operator, (i) in Condition 3.2 and Assumption 3.3 are satisfied for $\kappa = \gamma$ by the common power-law decay condition that $\lambda_j(\Sigma) \asymp j^{-\gamma}$. However, in general cases, we may only have $\kappa \geq \gamma$. Intuitively, the deviation between κ and γ can be seen as the misalignment between the basis taken and the covariance structure.

With the previous assumptions, we can now provide the generalization error of the differentially private FLR algorithm in the full observation setting.

Theorem 3.4. *Consider the FLR model under Assumption 3.3 with full observation. Apply Algorithm* 5 *with truncation level* $p \asymp \min\left(n^{\frac{1}{r+1}}, (\rho n^2)^{\frac{1}{2+r}} (\log n)^{-\frac{2}{2+r}}\right)$, *and other parameters as specified in Corollary* 2.11. *Then, the resulting* ρ*-zCDP estimator* $\tilde{\zeta}(X)$ *satisfies*

$$\sup_{\beta^* \in \mathcal{F}(r,R^*,C^*)} \mathbb{E}_{\beta^*} \mathcal{R}(\tilde{\zeta}, \zeta^*) \lesssim n^{-\frac{r}{r+1}} + (\rho n^2)^{-\frac{r}{2+r}} (\log n)^{\frac{2r}{2+r}}. \tag{11}$$

The (ϵ, δ)-DP version of Theorem 3.4 is given in the following corollary.

Corollary 3.5 ((ϵ,δ)-DP version of Theorem 3.4)**.** *Let* $\epsilon \in (0,1]$ *and* $\delta \in (0, e^{-1}]$. *Under the same conditions as in Theorem* 3.4, *set* $\rho = \epsilon^2 (8\log(1/\delta))^{-1}$ *and apply Algorithm* 5 *as*

specified in Theorem 3.4 *to obtain the* (ϵ, δ)*-DP estimator* $\tilde{\zeta}(X)$*. Then,*

$$\sup_{\beta^* \in \mathcal{F}(r,R^*,C^*)} \mathbb{E}_{\beta^*} \mathcal{R}(\tilde{\zeta}, \zeta^*) \lesssim n^{-\frac{r}{r+1}} + (\epsilon^2 n^2)^{-\frac{r}{2+r}} (\log n)^{\frac{2r}{2+r}} (\log(1/\delta))^{\frac{r}{2+r}}. \tag{12}$$

The proof of Theorem 3.4, given in the supplementary material, proceeds by truncating the basis expansion and reducing the problem to the linear model in (9). The error term ε_p is correlated with $u^*_{\leq p}$, necessitating the inclusion of μ in Assumption 2.9. The truncation level p balances approximation, estimation, and privacy errors.

We next provide a minimax lower bound for FLR under DP, measured by ρ-zCDP. The proof of Theorem 3.6, deferred to the supplementary material, builds on the idea in Theorem 2.13 and techniques from Cai et al. [2023]: we construct a tracing attack in a finite-dimensional setting and optimize the truncation level p to obtain the bound.

Theorem 3.6. *Consider the FLR model with full observations. Let* $\mathcal{M}_\rho$ *be the class of all* ρ*-zCDP algorithms. Then,*

$$\inf_{\hat{\zeta} \in \mathcal{M}_\rho} \sup_{\beta^* \in \mathcal{F}(r,R^*,C^*)} \mathbb{E} \left[\hat{\zeta}(X) - \zeta^*(X) \right]^2 \gtrsim n^{-\frac{r}{r+1}} + (\rho n^2)^{-\frac{r}{2+r}}. \tag{13}$$

Theorem 3.4 and Theorem 3.6 together show that our differentially private estimator is minimax-optimal up to $\log n$ factors. The generalization error consists of a statistical term $n^{-r/(r+1)}$—matching the non-private rate [Cai and Yuan, 2012, Yuan and Cai, 2010]—and a privacy term $(\rho n^2)^{-r/(2+r)}$, which reflects the cost of privacy [Cai et al., 2021, 2023]. The dominant term depends on ρ: if $\rho \lesssim n^{-r/(r+2)}$, the privacy error dominates; otherwise, the statistical term prevails.

Our estimator employs basis truncation followed by private linear regression. Classical methods based on splines or RKHS [Cai and Yuan, 2012, Yuan and Cai, 2010, Zhou et al., 2023] are challenging to privatize due to the infinite-dimensional nature of functional data. Standard mechanisms, such as the Gaussian mechanism, are often inapplicable or yield loose guarantees in this setting. Truncation reduces the problem to a finite-dimensional

one, but the resulting covariates are non-isotropic. Theorem 2.10, developed in Section 2, is a key tool for ensuring privacy under such general covariance structures.

3.2 Discrete Observation

Let us now consider the discrete observation setting. In this case, first we have to introduce the following regularity assumption.

Assumption 3.7 (Regularity)**.** There exist constants $\alpha_0 > 1/2$ and $C_{\mathrm{X}} > 0$ such that $\mathbb{E}\|X\|^2_{W^{\alpha_0,2}} < C_{\mathrm{X}}$. Additionally, there exists a constant $C_{\mathrm{reg}} > 0$ such that

$$\sup_{t\in T} \sum_{j=1}^{p} e_j(t)^2 \leq C_{\mathrm{reg}} p, \quad \forall p \geq 1. \tag{14}$$

The first condition in Assumption 3.7 is common in literature concerning discretely observed functional data [Cai and Yuan, 2011, Wang et al., 2022]. This condition can in fact be weakened to that the sup-norm $\|X\|_\infty$ is sub-Gaussian. The second condition (14) is a regularity condition on the basis functions. In particular, it is satisfied if the basis functions uniformly bounded, namely $\sup_j \|e_j\|_\infty < \infty$, so the Fourier basis satisfies this condition. Moreover, since this condition only requires the sum of the basis functions to be bounded, the wavelet basis also satisfies this condition as the supports of the wavelet functions only have a finite overlap.

Theorem 3.8. *Consider the FLR model under Assumptions 3.3 and 3.7. Let $\kappa' = \kappa$ for the independent design case and $\kappa' = \kappa + 2$ for the common design case. Apply Algorithm* 5 *with truncation level $p \asymp \min(p_*(n), p_{\mathrm{pri}}(n), p_{\mathrm{dis}}(n))$, where*

$$p_*(n) = n^{\frac{1}{r+1}}, \quad p_{\mathrm{pri}}(n) = (\rho n^2)^{\frac{1}{2+r}} (\log n)^{-\frac{2}{2+r}}, \quad p_{\mathrm{dis}}(n) = (m/\log n)^{1/\kappa'}. \tag{15}$$

Then, the resulting ρ-zCDP estimator $\tilde{\zeta}(\cdot)$ satisfies

$$\sup_{\beta^* \in \mathcal{F}(r,R^*,C^*)} \mathbb{E}_{\beta^*} \mathcal{R}(\tilde{\zeta}, \zeta^*) \lesssim \underbrace{n^{-\frac{r}{r+1}}}_{\text{statistical}} + \underbrace{(\rho n^2)^{-\frac{r}{2+r}} (\log n)^{\frac{2r}{2+r}}}_{\text{privacy}} + \underbrace{m^{-\frac{r}{\kappa'}} (\log n)^{\frac{r}{\kappa'}}}_{\text{discretization}}.$$

Corollary 3.9 ((ϵ, δ)-DP version of Theorem 3.8)**.** *Let $\epsilon \in (0, 1]$ and $\delta \in (0, e^{-1}]$. Under the same conditions as in Theorem* 3.8*, set $\rho = \epsilon^2 \left(8 \log\left(\frac{1}{\delta}\right)\right)^{-1}$ and apply Algorithm* 5 *with the truncation level specified in* (15) *to obtain the (ϵ, δ)-DP estimator $\tilde{\zeta}(X)$. Then, with probability at least $1 - Cn^{-2}$, the estimator satisfies*

$$\sup_{\beta^* \in \mathcal{F}(r, R^*, C^*)} \mathbb{E}_{\beta^*} \mathcal{R}(\tilde{\zeta}, \zeta^*) \lesssim$$

$$\underbrace{n^{-\frac{r}{r+1}}}_{\text{statistical}} + \underbrace{(\epsilon^2 n^2)^{-\frac{r}{2+r}} (\log n)^{\frac{2r}{2+r}} \left(\log\left(\frac{1}{\delta}\right)\right)^{\frac{r}{2+r}}}_{\text{privacy}} + \underbrace{m^{-\frac{r}{\kappa'}} (\log n)^{\frac{r}{\kappa'}} \left(\log\left(\frac{1}{\delta}\right)\right)^{\frac{r}{\kappa'}}}_{\text{discretization}}.$$

The proof of Theorem 3.8, deferred to the supplementary material, follows Theorem 3.4 with an additional analysis of the discretization error.

Ignoring logarithmic factors, Theorem 3.8 shows that the statistical and privacy errors match those in the full observation setting, while the discretization error is bounded by $m^{-r/\kappa}$ (independent design) or $m^{-r/(\kappa+2)}$ (common design). Since $r/\kappa > r/(\kappa + 2)$, the independent design yields smaller error, aligning with non-private results [Cai and Yuan, 2011]. When $m \geq \min(n^{\kappa'/(r+1)}, (\rho n^2)^{\kappa'/(r+2)})$, the discretization error becomes negligible, allowing the method to recover the minimax rate of the full observation setting—exhibiting a phase transition phenomenon also also observed in the non-private literature [Cai and Yuan, 2011, Zhou et al., 2023]. This trade-off can be further understood via the three critical truncation levels defined in (15):

- $p_*(n)$ balances approximation error $\lesssim p^{-r}$ and estimation error $\lesssim p/n$ (non-private),
- $p_{\text{pri}}(n)$ balances approximation and privacy error $\lesssim p^2/(\rho n^2)$,
- $p_{\text{dis}}(n)$ ensures reliable estimation of $\Sigma_{\leq p}$ from discrete observations, requiring $p \lesssim m^{1/\kappa}$ in the independent design case.

To demonstrate the phase transition phenomenon, let us consider the independent design case with Condition 3.2 and $\kappa = \gamma$. We present the phase diagram in Figure 1 on page

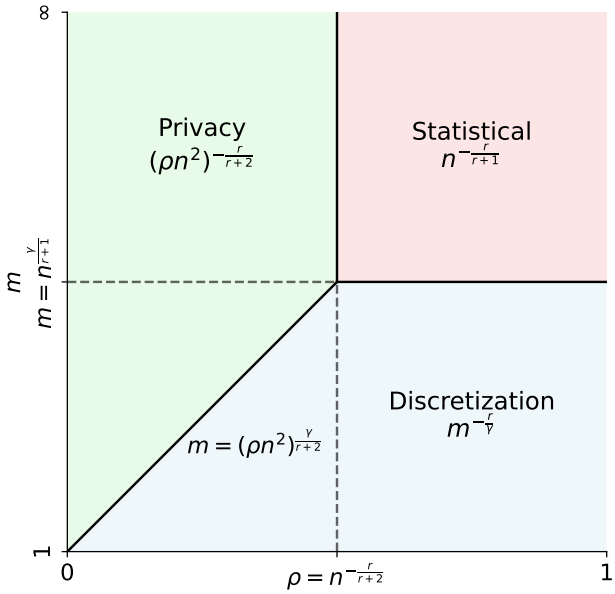

Figure 1: Phase diagram for the FLR model with independent design under DP. Horizontal axis (ρ): privacy budget, increasing from left to right; vertical axis (m): number of discrete observations, increasing from bottom to top.

The phase diagram is divided into regions dominated by each source of error: the statistical error, the privacy error, and the discretization error.

26. The phase diagram is divided into regions dominated by each source of error:

- *Statistical region:* Dominated by the statistical error $n^{-\frac{r}{r+1}}$, occurring when $m \geq n^{\frac{\gamma}{r+1}}$ and $\rho \geq n^{-\frac{r}{r+2}}$.
- *Privacy region:* Dominated by the privacy error $(\rho n^2)^{-\frac{r}{2+r}}$, occurring when $\rho \leq n^{-\frac{r}{r+2}}$ and $m \geq (\rho n^2)^{\frac{\gamma}{r+2}}$.
- *Discretization region:* Dominated by discretization error $m^{-\frac{r}{\kappa}}$ or $m^{-\frac{r}{\kappa+2}}$ (for independent and common designs, respectively), occurring when $m \leq \min\left(n^{\frac{\kappa}{r+1}}, (\rho n^2)^{\frac{\kappa}{r+2}}\right)$.

Transitions between these regions correspond to key thresholds:

- privacy-statistical transition at $\rho = n^{-\frac{r}{r+2}}$.
- discretization-statistical transition at $m = n^{\frac{\gamma}{r+1}}$.
- discretization-privacy transition at $m = (\rho n^2)^{\frac{\gamma}{r+2}}$.

This phase diagram captures the interplay among statistical accuracy, privacy, and discretization, aligning with discretization-statistical trade-offs in the non-private literature [Cai and Yuan, 2011, Zhou et al., 2023] and privacy-statistical trade-offs in the DP literature [Cai et al., 2021, 2023], offering a unified perspective for designing differentially private FLR under discrete observation.

4 Numerical Results

4.1 Simulation Data

We begin by evaluating the proposed differentially private algorithm in a simulation setting. The eigen-basis is defined as $e_1(t) = 1$ and $e_j(t) = \sqrt{2}\cos((j-1)\pi t)$ for $j \geq 2$. The covariance of the functional covariate $X(\cdot)$ is specified as a diagonal matrix with the j-th diagonal element $\lambda_j = j^{-\gamma}$, where $\gamma > 1$. The true functional coefficient $\boldsymbol{\beta}^*$ is set as $\beta_j^* = j^{-\frac{\alpha+1}{2}}$ for $\alpha > 0$, satisfying Condition 3.2. For the discrete observation case in (6), we introduce Gaussian noise with variances $\sigma = 0.5$ and $\sigma_{\mathrm{x}} = 0.1$. We explore various configurations of the eigenvalue decay γ, smoothness parameter α, sample size n, and number of discrete observations m. Algorithm hyperparameters are set as $L = \sqrt{p}$, $L_y = \sqrt{\log n}$, $T = \log n$, with the truncation level p selected via cross-validation.

In Figure 2 on page 28, we present the generalization error, measured as mean squared error (MSE), of the proposed differentially private method as a function of the privacy budget ρ. The plot reveals that the generalization error decreases as ρ increases, stabilizing when ρ is sufficiently large. This illustrates the trade-off between privacy and accuracy, where stricter privacy constraints lead to higher generalization error. Moreover, the stabilization of the error at large ρ indicates that the statistical error dominates the privacy error in this regime. In the left plot, with a fixed number of discrete sample points m, the generalization error decreases as the sample size n grows. However, when both the

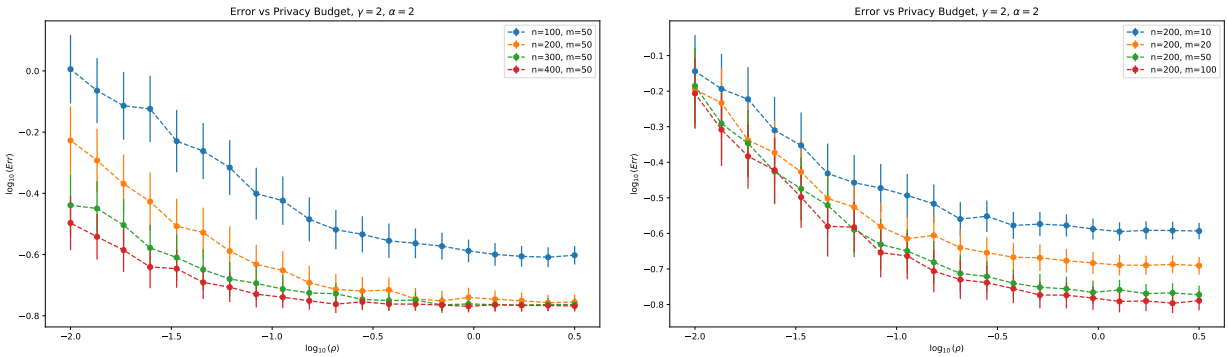

Figure 2: Generalization error of the proposed differentially private algorithm with respect to the privacy budget ρ under $\gamma = 2$, $\gamma = 2$. The error bar represents the standard deviation (scaled by 0.2 for readability) over 500 replications. Left: Fix $m = 50$ and vary n; Right: Fix $n = 200$ and vary m.

privacy budget and n are large, further increases in n or ρ yield limited improvement, signaling a phase transition to a region dominated by discretization error. In the right plot, the generalization error improves with an increasing number of discrete observations m. However, this improvement plateaus beyond a certain threshold of m, indicating a regime where statistical error prevails. These findings are consistent with our theoretical results. Due to the page limit, additional numerical results demonstrating the interplay of the parameters α, γ, m and comparison with non-private methods are provided in the supplementary material.

4.2 Real Data

We now apply our method to two benchmark functional datasets: the Tecator dataset [Febrero-Bande and Oviedo de la Fuente, 2012], which involves predicting fat, water, and protein content from meat absorbance spectra; and the wheat dataset [Kalivas, 1997], which uses near-infrared spectra to estimate protein content.

4.2.1 Tecator Dataset

The Tecator dataset [Febrero-Bande and Oviedo de la Fuente, 2012] is a widely referenced benchmark in functional data analysis. This dataset originates from a study on meat samples and has been extensively used in the analysis of functional data. It consists of 215 meat samples, each characterized by its absorbance spectrum measured over 100 wavelengths in the range of 850 nm to 1050 nm. These absorbance curves, treated as functional predictors, represent continuous spectral information that reflects the chemical properties of the samples. The corresponding response variables are the percentages of fat, water, and protein content, determined through laboratory analysis.

In the numerical experiments, we adopt the same Fourier basis as in the simulation setting. We apply the FLR model separately to each response variable (protein, water, and fat content). The dataset is split randomly, with 80% of the samples allocated to the training set and the remaining 20% to the test set. Hyperparameters for the algorithm are selected via grid search with cross-validation to ensure optimality. Results are presented in Table 1 and Figure 3 on page 30. These findings highlight the trade-off between privacy and accuracy: the method achieves strong performance with a sufficiently large privacy budget, but stringent privacy constraints lead to a notable increase in prediction error.

Response		$\rho = 0.5$	$\rho = 1$	$\rho = 5$	$\rho = \infty$
Protein	MSE	7.70 (1.16)	5.75 (1.54)	4.01 (0.66)	1.45 (0.29)
	R^2	0.14 (0.11)	0.33 (0.16)	0.51 (0.12)	0.82 (0.04)

Table 1: Prediction error and R^2 score (and standard deviations) of the differentially private FLR method on the Tecator dataset over the protein response variables with different privacy budgets ρ. Results are averaged over 200 bootstrap replications.

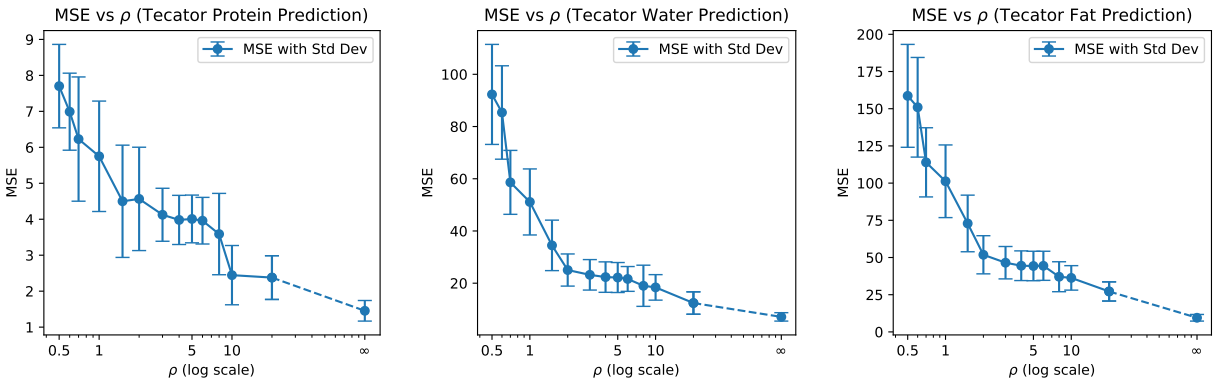

Figure 3: Prediction error of the proposed differentially private algorithm on the Tecator dataset with respect to the privacy budget ρ. The error bar represents the standard deviation over 200 bootstrapped replications.

4.2.2 The Wheat Data

The wheat dataset [Kalivas, 1997] is a widely used benchmark in spectroscopic calibration. We apply our method to this dataset and compare it with the non-private approaches of Zhou et al. [2023]. The data consist of near-infrared spectra measured at 701 equally spaced wavelengths from 1100 nm to 2500 nm, serving as functional predictors. The response variable is the protein content, a key indicator of nutritional quality in wheat.

We compare our proposed differentially private methods with the non-private methods for FLR in Zhou et al. [2023]. Following the experimental setup in Zhou et al. [2023], we use the first 80 samples for training and the remaining 20 for testing. To simulate sparse observation, we randomly select m frequencies from the 701 available. Prediction results are reported in Table 2 on page 31. While stronger privacy constraints increase prediction error, our method remains competitive with non-private baselines, demonstrating its effectiveness in real-world functional data settings.

m	$\rho = 0.1$	$\rho = 0.5$	$\rho = 1$	**PACE**	**POOL**
10	1.72 (1.64)	1.02 (0.59)	0.93 (0.42)	1.52 (9.11)	0.35 (0.07)
30	1.42 (1.36)	0.91 (0.49)	0.85 (0.34)	0.36 (0.24)	0.28 (0.07)

Table 2: Comparison of proposed DP FLR with different ρ's and other non-private methods (PACE and POOL). The results for the non-private methods are reported in Zhou et al. [2023] and the DP methods are computed in the same setting as the non-private methods.

5 Discussion

This work contributes to the growing body of research at the intersection of functional data analysis and privacy-preserving statistics. Unlike conventional finite-dimensional settings, functional data are characterized by rich structures and infinite-dimensional feature spaces, requiring fundamentally different approaches to privacy mechanism design and sensitivity calibration. Our method addresses these difficulties through a dimension-reducing truncation and a preconditioning step that mitigates the ill-conditioning commonly observed in functional covariates.

A key insight of our analysis is that DP in functional models demands careful coordination between statistical estimation and privacy budget allocation. In the fully observed case, our minimax-optimality result under ρ-zCDP provides a benchmark for future development of private inference tools in infinite-dimensional spaces. In the discrete observation case, our identification of a phase transition phenomenon clarifies the conditions under which discretization error becomes negligible, allowing for practical guidance on sampling density in privacy-aware functional studies.

Looking ahead, there are several promising directions for future work. First, while our study focuses on FLR, the methodology may extend naturally to generalized functional regression models or functional additive models under privacy constraints. Second, in many

real-world applications, functional data are observed with measurement noise. Extending our framework to accommodate both privacy and functional measurement error is a non-trivial but important extension. Third, our current analysis assumes independence across individuals; developing DP methods for longitudinal or correlated functional data remains an open and compelling challenge. Finally, integrating our approach into federated learning frameworks—where functional data are distributed across institutions—could further enhance privacy guarantees while enabling collaborative analysis at scale.

We hope that our findings stimulate further exploration into privacy-preserving functional data methods and contribute to the development of theoretically sound and practically viable tools for analyzing sensitive, high-dimensional data.

Data Availability Statement

The real datasets used in this study are publicly available in R packages: the Tecator dataset through `fda.usc` [Febrero-Bande and Oviedo de la Fuente, 2012] and the wheat dataset through `OHPL` [Kalivas, 1997].

References

Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, Vienna Austria, October 2016. ACM. ISBN 978-1-4503-4139-4. doi: 10.1145/2976749.2978318.

Mohammed Adnan, Shivam Kalra, Jesse C. Cresswell, Graham W. Taylor, and Hamid R. Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific reports*, 12(1):1953, 2022. doi: 10.1038/s41598-022-05539-7.

Daniel Alabi, Pravesh K. Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately Estimating a Gaussian: Efficient, Robust, and Optimal. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, STOC 2023, pages 483–496,

New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 978-1-4503-9913-5. doi: 10.1145/3564246.3585194.

Kareem Amin, Travis Dick, Alex Kulesza, Andres Munoz, and Sergei Vassilvitskii. Differentially Private Covariance Estimation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back, February 2023.

Rina Foygel Barber and John C. Duchi. Privacy and Statistical Risk: Formalisms and Minimax Bounds, December 2014.

Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, 33:14475–14485, 2020.

Mark Bun and Thomas Steinke. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds, May 2016.

T. Tony Cai and Ming Yuan. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics*, 39(5), October 2011. ISSN 0090-5364. doi: 10.1214/11-AOS898.

T. Tony Cai and Ming Yuan. Minimax and Adaptive Prediction for Functional Linear Regression. *Journal of the American Statistical Association*, 107(499):1201–1216, September 2012. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2012.716337.

T. Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5), October 2021. ISSN 0090-5364. doi: 10.1214/21-AOS2058.

T. Tony Cai, Yichen Wang, and Linjun Zhang. Score Attack: A Lower Bound Technique for Optimal Differentially Private Learning, March 2023.

T. Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Federated Nonparametric Hypothesis Testing with Differential Privacy Constraints: Optimal Rates and Adaptive Tests, June 2024a.

T. Tony Cai, Abhinav Chakraborty, and Lasse Vuursteen. Optimal Federated Learning for Nonparametric Regression with Heterogeneous Distributed Differential Privacy Constraints, June 2024b.

T. Tony Cai, Dong Xia, and Mengyue Zha. Optimal differentially private pca and estimation for spiked covariance matrices, January 2024c.

T. Tony Cai, Abhinav Chakraborty, and Yichen Wang. Optimal Differentially Private Ranking from Pairwise Comparisons, July 2025.

Wei Dong, Yuting Liang, and Ke Yi. Differentially Private Covariance Revisited, September 2022.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork and Guy N. Rothblum. Concentrated Differential Privacy, March 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006. doi: 10.1007/11681878_14.

Cynthia Dwork, Weijie J. Su, and Li Zhang. Differentially Private False Discovery Rate Control, July 2021.

Manuel Febrero-Bande and Manuel Oviedo de la Fuente. Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, 51(4):1–28, 2012. doi: 10.18637/jss.v051.i04.

Peter Hall and Joel L. Horowitz. Methodology and convergence rates for functional linear regression. 2007. doi: 10.1214/009053606000000957.

Michael B. Hawes. Implementing differential privacy: Seven lessons from the 2020 United States Census. *Harvard Data Science Review*, 2(2):4, 2020.

Tailen Hsing and Randall Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley, 1 edition, May 2015. ISBN 978-0-470-01691-6 978-1-118-76254-7. doi: 10.1002/9781118762547.

John H Kalivas. Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 37(2):255–259, 1997. doi: 10.1016/S0169-7439(97)00038-5.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 1853–1902. PMLR, June 2019.

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private Mean Estimation of Heavy-Tailed Distributions, February 2021.

Ilmun Kim and Antonin Schrab. Differentially Private Permutation Tests: Applications to Kernel Methods, January 2024.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 1167–1246. PMLR, June 2022.

Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Suggala. Label robust and differentially private linear regression: Computational and statistical efficiency. *Advances in Neural Information Processing Systems*, 36:23019–23033, December 2023.

NhatHai Phan, Yue Wang, Xintao Wu, and Dejing Dou. Differential privacy preservation for deep auto-encoders: An application of human behavior prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. doi: 10.1609/aaai.v30i1.10165.

Jim Ramsay and B. W. Silverman. *Functional Data Analysis (Springer Series in Statistics)*. Springer, 2nd edition, 2005. ISBN 978-0-387-40080-8.

Kunal Talwar, Abhradeep Guha Thakurta, and Li Zhang. Nearly Optimal Private LASSO. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (Nearly) optimal private linear regression via adaptive clipping, July 2022.

Daren Wang, Zifeng Zhao, Yi Yu, and Rebecca Willett. Functional linear regression with mixed predictors, August 2022.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590, 2005. doi: 10.1198/016214504000001745.

Ming Yuan and T. Tony Cai. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, December 2010. ISSN 0090-5364, 2168-8966. doi: 10.1214/09-AOS772.

Xiaoke Zhang and Jane-Ling Wang. From sparse to dense functional data and beyond. 2016. doi: 10.1214/16-AOS1446.

Hang Zhou, Fang Yao, and Huiming Zhang. Functional linear regression for discretely observed data: From ideal to reality. *Biometrika*, 110(2):381–393, 2023. doi: 10.1093/biomet/asac053.