Private Statistical Survey Avoiding Privacy Composition in the Real World

Department of Statistical Science
School of Multidisciplinary Sciences
The Graduate University for Advanced Studies, SOKENDAI
Hajime Ono

March 2023

Contents

1 Introduction **5**

2 Preliminaries **7**
2.1 Convergences 7
2.2 Estimation Problem 8
2.3 Minimax risk analysis 9
2.4 QMLE and its asymptotic normality 11

3 A Practical LDP Quasi-MLE **13**
3.1 Introduction 13
3.2 Preliminaries 15
3.2.1 Local Differential Privacy 15
3.2.2 Quasi-Maximum Likelihood Estimator 16
3.2.3 Quantile Regression 17
3.3 Proposed Protocol 18
3.3.1 Regression with Public X 18
3.3.2 Regression with Private X 21
3.3.3 Remark and Limitation 23
3.4 Example: Quantile Regression 23
3.4.1 With Public X 24
3.4.2 With Private X 24
3.5 Numerical Evaluation 25
3.6 Conclusion 25
3.7 Pseudo-code 26
3.8 Mathematical Notes 27
3.8.1 for Section 3.3.1 27
3.8.2 for Section 3.3.2 28
3.8.3 for Section 3.4 30
3.9 Comparison with Non-private Estimator 35
3.10 Additional Numerical Evaluation 36
3.10.1 Evaluation of Private X 36
3.10.2 Evaluation of Effect of Truncation 37
3.10.3 Comparison with Non-private Estimator 37

4 Local Privacy in the Presence of Unexpected Values **41**
4.1 Introduction 41
4.2 Background 43
4.2.1 Local Differential Privacy 43

4.3 Local Privacy with Unexpected Values 45
4.4 Lower Bound of Estimation Problem 47
4.4.1 Abstract Framework 48
4.4.2 Simple erasure of out-of-domain values 49
4.4.3 Stochastic erasure as ρ^{pre} and ρ^{pos} 50
4.5 Examples 51
4.5.1 First Example 51
4.5.2 Second Example 52
4.6 Conclusion 53
4.7 Proof of Proposition 8 53
4.8 Proof of Proposition 10 54

5 Inconsistency Due to Synthetic-data Use **57**
5.1 Introduction 57
5.2 Analytic Target 59
5.3 Minimax lower bound analysis 60
5.4 An Example of Inconsistency 63
5.5 Numerical Experiments 67
5.5.1 Fitting to a Gaussian Model 68
5.5.2 Fitting to a Laplace Model 69
5.6 Discussion and Future Work 70
5.7 Related Work 71
5.8 Conclusion 71

6 Conclusion **73**

7 Acknowledge **75**

Chapter 1

Introduction

In recent years, data collection and use have become increasingly popular. Since data records are often tied to real individuals, it is necessary to consider the privacy of the data providers when using the data. One promising approach to balance privacy protection and data utilization is to publish perturbed data or statistics instead of raw data. Differential privacy [Dwork et al., 2006, Dwork and Roth, 2014] is a quantitative definition of privacy for such a perturbation strategy. The definition requires a data curator to perturb the publication such that an adversary cannot distinguish two neighboring datasets using the perturbed publication. Moreover, the definition regards a perturbation mechanism as safer if the adversary is less likely to distinguish two neighborhood databases. There are many differentially private algorithms, ranging from basic ones [Dwork and Roth, 2014] to complex ones such as deep learning [Abadi et al., 2016]. Differential privacy has been deployed in the real world. For example, the U.S. Census Bureau adopts differentially private perturbation mechanisms when it publishes statistics of the census [Abowd, 2018].

However, differentially private publications of statistics cannot control privacy risks when a curator publishes a large number of statistics. In the real world, many researchers access and analyze some popular datasets, and eventually publish their findings in research papers. Official microdata are an example of such popular data, which consist of highly sensitive records. As the curator publishes statistics, the privacy risk accumulates. The accumulation of privacy risk is called *privacy composition* and has been studied in [McSherry, 2009, Kairouz et al., 2015, Abadi et al., 2016]. Even if each publication strictly controls the privacy risk, the accumulated privacy risk caused by multiple publications can be unboundedly large. This issue also occurs on federated learning [Kairouz et al., 2021] which is a distributed machine learning framework. In the framework, clients who possess a local dataset repeatedly communicate with a central server to update a statistical estimation. Even if clients perturb their submissions to prevent direct disclosure of their local dataset, privacy risk accumulates communication by communication. Can we avoid the accumulation of privacy risk by multiple publications while maintaining the utility of data?

One possible solution to avoid privacy composition is to use local perturbation methods such that data providers perturb their data before supplying it to a data curator. Local differential privacy (LDP) [Kasiviswanathan et al., 2011, Duchi et al., 2013] is a quantitative definition of privacy achieved by such

a local perturbation method. Originally, local perturbation strategies and LDP are studied to ensure that user privacy is protected even if data curators are adversarial. Notably, Google and Apple have conducted statistical surveys that guarantee user privacy based on this definition [Erlingsson et al., 2014, Apple Differential Privacy Team, 2017]. Data collected while satisfying LDP automatically satisfies DP. The perturbed data can be further used without privacy composition.

Although a data collection method satisfying LDP promises strict privacy protection, the requirement by LDP raises issues concerning privacy and data utility. The LDP definition requires a data provider to perturb her record so as to be indistinguishable from the other candidate records in the domain. To satisfy the requirement, perturbation mechanisms often assume a known data domain that is finite or bounded. However, since the LDP system model allows no participant to have a complete picture of the raw data, the assumption that the data domain is known in advance is unrealistic.

When a perturbation mechanism receives an undesirable value, the mechanism can output an invalid value or nothing. Undesirable values include extremely large values, non-responses, and unintended error messages. By observing the abnormal behavior, the curator can infer that the user supplied an abnormal value. The lack of knowledge of data decreases data utility. For example, the data curator tends to fit the data to a misspecified model.

To handle the issue, we propose an LDP protocol for Quasi-MLE using truncation. Truncation is a technique that projects real values into a bounded interval. Quasi-MLE is an estimator for a model parameter and works even if we misspecified the model. We analyze the QMLE's asymptotic behavior. The analysis helps a curator to understand the data without directly observing the data. The contribution corresponding to this paragraph has been published in a conference proceeding [Ono et al., 2022].

Although truncation is helpful for handling extremely large values, it cannot cope with other undesirable values. Since it is necessary to implement a secure exception-handling mechanism to handle various unexpected inputs, we have proposed a modified LDP that includes this exception-handling mechanism. We also analyzed the benefits of including the exception-handling mechanism.

Another possible way to avoid privacy composition is the use of synthetic data that mimics the statistical properties of the original data. Synthetic data are not necessarily discussed in relation to DP, but, in recent years, a framework has been established to quantitatively discuss the degree of protection in relation to DP [Neunhoeffer et al., 2021].

However, it is not obvious that estimators evaluated using synthetic data are always useful as those of population statistics. We identify sufficient conditions under which estimators evaluated using synthetic do not match the population statistics that we truly wish to estimate. We also show that there may be problems that satisfy sufficient conditions.

This thesis is organized as follows. In Chapter 2, we introduce some knowledge that is necessary to read this thesis. In Chapter 3, we study a locally private quasi-MLE that is feasible in the real world. In Chapter 4, we study the privacy risk in the presence of unexpected values. In Chapter 5, we study the inconsistency of estimators caused by the use of synthetic data. In Chapter 6, we offer the conclusion of this thesis.

Chapter 2

Preliminaries

In this chapter, we introduce some basic concepts regarding multiple chapters. We will describe some basic concepts regarding a single chapter in the chapter. We denote the expectation of $f(X)$ as $P(f(X))$ to clarify the distribution P that the random variable X follows. We refer to the j-th element of a vector v by $[v]_j$.

2.1 Convergences

We use several different concepts of convergence of probabilistic measures, convergence in total variation, and convergence in distribution.

We first introduce *weak convergence*. We say that a sequence $\{X_n\}_{n=1}^{\infty}$ of random variables weakly converges or converges in distribution to a random variable X if

$$P_n(X_n \leq x) \to P(X \leq x)$$

where P_n and P are the distributions that X_n and X follow, respectively. The definition of weak convergence does not require that the density functions agree with each other. There are several equivalent definitions of weak convergence.

Lemma 1 (Portmanteau, Lemma 2.2 of [Vaart, 2000])**.** *For any random vectors* $\mathbf{X}_n$ *and* $\mathbf{X}$*, the following statements are equivalent.*

1. $P(\mathbf{X}_n \leq \mathbf{x}) \to P(\mathbf{X} \leq \mathbf{x})$ *for all continuity points of* $\mathbf{x} \mapsto P(\mathbf{X} \leq \mathbf{x})$*;*
2. $\mathbb{E}[f(\mathbf{X}_n)] \to \mathbb{E}[f(\mathbf{X})]$ *for any bounded, continuous function* f*;*
3. $\mathbb{E}[f(\mathbf{X}_n)] \to \mathbb{E}[f(\mathbf{X})]$ *for any bounded, Lipschitz function* f*;*
4. $\liminf \mathbb{E}[f(\mathbf{X}_n)] \geq \mathbb{E}[f(\mathbf{X})]$ *for any non-negative, continuous function* f*;*
5. $\liminf P(\mathbf{X}_n \in G) \geq P(\mathbf{X} \in G)$ *for every open set* G*;*
6. $\limsup P(\mathbf{X}_n \in F) \leq P(\mathbf{X} \in F)$ *for every closed set* F*;*
7. $P(\mathbf{X}_n \in B) \to P(\mathbf{X} \in B)$ *for any Borel set* B *with* $P(\mathbf{X} \in \delta B) = 0$*, where* δB *is the boundary of* B*.*

□

Statement (2) implies that weak convergence does not guarantee that $\mathbb{E}[f(\mathbf{X}_n)] \to \mathbb{E}[f(\mathbf{X})]$ for an unbounded or discontinuous function f.

Second, we introduce *convergence in probability*. We say that a sequence $\{X_n\}_{n=1}^{\infty}$ of random variables converges to X in probability if, for any $\epsilon > 0$,

$$\Pr(d(X_n, X) > \epsilon) \to 0.$$

We say that sequence $\{X_n\}_{n=1}^{\infty}$ *converges to* X *almost surely* when

$$\Pr(\lim(d(X_n, X)) = 0) = 1.$$

Finally, we will use *convergence in total variation*, which is denoted

$$\|P_n - P\|_{\mathrm{TV}} \to 0.$$

2.2 Estimation Problem

An estimation problem is a problem of estimating an unknown population parameter using data samples from the population. Here, we describe the standard estimation problem in which analysts can access the raw data directly.

We first describe data generation and estimation. Let $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$ be a family of probability distributions, indexed by $\Theta \subset \mathbb{R}^k$ where k is a natural number, on a measurable space $(\mathcal{X}, \mathcal{A})$. An unknown distribution $P_\theta \in \mathcal{P}$ generates n records independently, and we denote the records by $X_1, ..., X_n$. We treat $X_1, ..., X_n$ as random variables and denote their realizations by $x_1, ..., x_n$. Let $D_n = \{X_1, ..., X_n\}$, where we say that data D_n has size n. An analyst observes D_n and estimates θ or P_θ. For the estimation, the analyst uses an estimator $\hat{\theta}_n : \mathcal{X}^n \to \Theta; D_n \mapsto \hat{\theta}_n(D_n)$.

Next, we define maximum risk, a performance measure of estimators. To define risk, we use a semi-distance ρ on Θ and a weighting function w. $\rho : \Theta \times \Theta \to [0, +\infty)$ is a function satisfying symmetry $\rho(\theta, \theta') = \rho(\theta', \theta)$, triangle inequality $\rho(\theta, \theta'') \leq \rho(\theta, \theta') + \rho(\theta', \theta'')$, and $\rho(\theta, \theta) = 0$ for any $\theta, \theta', \theta'' \in \Theta$. ρ is a measure of how far apart the two elements of Θ are. The weighting function w, representing how much we penalize the semi-distance, is a function satisfying that

$$w : [0, \infty) \to [0, \infty) \text{ is monotone increasing, } w(0) = 0, \text{ and } w \neq 0, \tag{2.1}$$

where $w \neq 0$ means that the function w is not a constant function outputting 0. For example, $w : t \mapsto t$ and $w : t \mapsto t^2$. Given ρ and w, the performance of an estimator $\hat{\theta}_n$ of θ is measured by the maximum risk of this estimator on Θ:

$$\sup_{\theta \in \Theta} P_\theta(w(\rho(\hat{\theta}_n(D_n), \theta))).$$

Due to the monotonicity of w, $\rho(\theta', \theta) \leq \rho(\theta'', \theta)$ always implies $w(\rho(\theta', \theta)) \leq w(\rho(\theta'', \theta))$.

We here explain why we defined semi-distance ρ and weighting function w separately. Semi-distance ρ is useful in a mathematical proof. Especially, the triangle inequality plays an important role in analyzing minimax risk. However,

some important objective functions used in machine learning or statistics are not semi-distances. An example of a popular objective function not satisfying the triangle inequality is the squared loss function. We use weighting functions for both the convenience of quasi-distance and the generality of the theory. With $\rho(\theta, \theta') = |\theta - \theta|$ and $w(t) = t^2$, we can handle the squared loss function.

Minimax risk, which is a measure of the difficulty of the estimation problem, is defined by taking the infimum of the maximum risk over estimators:

$$\mathcal{R}_n^* \equiv \inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta(w(\rho(\hat{\theta}_n(D_n), \theta))).$$

The definition does not depend on any specific estimator, distribution, and data. Roughly speaking, the minimax risk is the risk of the best estimator in its worst case.

We finally define the consistency of an estimator. We say that an estimator $\hat{\theta}_n$ is consistent if the estimation $\hat{\theta}_n(D_n)$ converges to θ in probability for any P_θ. We also say that an estimator is inconsistent if the estimator is not consistent.

2.3 Minimax risk analysis

In this section, we describe a strategy for analyzing lower bounds of minimax risk. The standard strategy is to reduce the estimation problem to a hypothesis testing problem [Tsybakov, 2009].

We first derive a lower bound, which is characterized by the probability of estimation $\hat{\theta}_n(D_n)$ being far from true parameter θ, of the risk. For any weighting function w and any $\delta > 0$ such that $w(\delta) > 0$, we have, by the Markov inequality,

$$P_\theta(w(\rho(\hat{\theta}_n(D_n), \theta))) \geq w(\delta) P_\theta(\rho(\hat{\theta}_n(D_n), \theta) \geq \delta).$$

Thus, we focus on the probability $P_\theta(\rho(\hat{\theta}_n(D_n), \theta) \geq \delta)$.

Next, we lower bound $P_\theta(\rho(\hat{\theta}_n(D_n), \theta) \geq \delta)$ by the error probability of a finite hypotheses test. It is clear that

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta(\rho(\hat{\theta}_n(D_n), \theta) \geq \delta) \geq \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, ..., \theta_M\}} P_\theta(\rho(\hat{\theta}_n(D_n), \theta) \geq \delta) \tag{2.2}$$

for any finite subset $\{\theta_0, ..., \theta_M\}$ of Θ. We will call *hypotheses* the $M+1$ elements $\theta_0, ..., \theta_M$ of Θ chosen to obtain lower bounds on the minimax risk and call a *test* any $\mathcal{A}$-measurable function $\psi : \mathcal{X} \to \{0, ..., M\}$. We denote the distributions corresponding to θ_j for $j = 0, ..., M$ by P_j. Next, we restrict the hypotheses to make the analysis simple. We choose the hypotheses such that

$$\rho(\theta_j, \theta_{j'}) \geq 2\delta, \quad \forall j', j : j' \neq j.$$

Then, for any estimator $\hat{\theta}_n$,

$$P_j(\rho(\hat{\theta}_n(D_n), \theta_j) \geq \delta) \geq P_j(\psi^*(D_n; \hat{\theta}_n) \neq j), \quad j = 0, ..., M, \tag{2.3}$$

where $\psi^* : \mathcal{X}^n \to \{0, ..., M\}$ is the *minimum distance test* defined by

$$\psi^*(D_n; \hat{\theta}_n) = \arg\min_{0 \leq j \leq M} \rho(\hat{\theta}_n(D_n), \theta_j).$$

Equation (2.3) is obtained by the triangle inequality and the property of ψ^*. Since a similar proof will appear in Section 5.3, we omit the proof here. Combining (2.2) and (2.3), we have

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \Theta} P_\theta(\rho(\hat{\theta}_n(D_n), \theta) \geq \delta) \geq \inf_{\hat{\theta}_n} \max_{\theta \in \{\theta_0, \ldots, \theta_M\}} P_\theta(\rho(\hat{\theta}_n(D_n), \theta) \geq \delta) \geq p_{\mathrm{e},M},$$

where $p_{\mathrm{e},M}$ is the error probability of the best test:

$$p_{\mathrm{e},M} \equiv \inf_{\psi} \max_{0 \leq j \leq M} P_j(\psi \neq j).$$

The infimum is taken over all tests $\{\psi : \mathcal{X}^n \to \{0, ..., M\}\}$.

Our next interest is the error probability $p_{\mathrm{e},M}$. For the discussion in this thesis, $M = 1$ is sufficient. Thus, in the remainder of this thesis, we consider the two-hypotheses case. To lower bound the error probability, we introduce the following useful lemma.

Lemma 2 (Theorem 2.2 of [Tsybakov, 2009])**.** *Let P_0 and P_1 be two probability measures on $(\mathcal{X}, \mathcal{A})$. Suppose that there exists a real value α such that*

$$\|P_1 - P_0\|_{TV} \equiv \sup_{S \in \mathcal{A}} |P_1(S) - P_0(S)| \leq \alpha < 1.$$

Then the following relation holds:

$$p_{\mathrm{e},1} \geq \frac{1 - \alpha}{2}.$$

□

This lemma implies that we obtain a lower bound of the error probability if we obtain an upper bound of the total variation between the hypotheses.

Summarizing this section, we obtain the following theorem, which shows a minimax lower bound.

Theorem 1. *Let $\rho : \Theta \times \Theta \to [0, +\infty)$ be a semi-distance, and let $w : [0, \infty) \to [0, \infty)$ be a monotone increasing function such that $w(0) = 0$ and $w \neq 0$. Suppose that there exist $P_0, P_1 \in \mathcal{P}$ such that $\|P_1^n - P_0^n\|_{TV} \leq \alpha$ and $\rho(\theta_0, \theta_1) \geq 2\delta$ with positive values δ and α such that $w(\delta) > 0$ and $\alpha < 1$. Then we have*

$$\mathcal{R}_n^* \geq w(\delta) \frac{1 - \alpha}{2}. \tag{2.4}$$

□

By choosing P_0, P_1 concretely, we can find a minimax lower bound. In Section 5.3, we analyze the optimal convergence rate in our problem by modifying this theorem.

Finally, we introduce Pinsker's inequality, which is a classical and helpful inequality. For any distributions P and Q, we have

$$\|P - Q\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}(P \| Q).$$

This inequality allows us to use an upper bound of KL divergence instead of that of total variation in Theorem 1.

2.4 QMLE and its asymptotic normality

So far, we have described the evaluation of $\hat{\theta}_n$ and the lower bound for the minimax risk. In this section, we introduce concrete implementations and the properties of an estimator $\hat{\theta}_n$, a quasi-maximum likelihood estimator (QMLE), which is a standard method for parametric model fitting.

Given data D_n and the model family $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$, the QMLE is given as the maximizer of the log-likelihood function defined as

$$L_n(\theta; D_n) \equiv \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i), \tag{2.5}$$

where p_θ is the Radon–Nikodym density [Athreya and Lahiri, 2006] of P_θ with an appropriate measure. Strictly speaking, (2.5) is not always defined and maximized. To ensure that (2.5) can be defined and maximized, we make the following assumptions.

Assumption 1. *The independent random vectors $X_i, i = 1, ..., n$, have identical joint distribution function P on $(\mathcal{X}, \mathcal{A})$, a measurable Euclidean space, with measurable Radon–Nikodym density $p = dP/d\nu$, where ν is an appropriate measure on $(\mathcal{X}, \mathcal{A})$.* □

Assumption 2. *The family of distribution functions P_θ has Radon–Nikodym densities $p_\theta = dP_\theta/d\nu$ which are measurable in $(\mathcal{X}, \mathcal{A})$ for every $\theta \in \Theta$ and continuous in θ for every $x \in \mathcal{X}$.* □

Assumption 3. *(a) $P \log p$ exists and $|\log p_\theta(x)| \leq m(x)$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$, where m is integrable with respect to P; (b) $P \log p_\theta$ has a unique maximum at $\theta = \theta_* \in \Theta$.* □

Assumptions 1 and 2 require that both P and P_θ be regular distributions. Assumption 3 is used to ensure that the sequence of estimators converges to a point θ_* consistently.

We next describe the matrices which characterize the convergence of the estimator in distribution. When the partial derivatives exist and a realization $(x_1, ..., x_n)$ of data is given, we define matrices $A_n(\theta)$ and $B_n(\theta)$ as those whose elements are

$$[A_n(\theta)]_{j,j'} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log p_\theta(x_i)}{\partial [\theta]_j \partial [\theta]_{j'}}$$

$$\text{and} \quad [B_n(\theta)]_{j,j'} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \log p_\theta(x_i)}{\partial [\theta]_j} \frac{\partial \log p_\theta(x_i)}{\partial [\theta]_{j'}} \quad \text{for } j, j' = 1, ..., k.$$

If expectations also exist, we define $k \times k$ matrices $A(\theta; P)$ and $B(\theta; P)$ as those whose elements are

$$[A(\theta; P)]_{j,j'} = P\left(\frac{\partial^2 \log p_\theta}{\partial [\theta]_j \partial [\theta]_{j'}} \right) \quad \text{and} \quad [B(\theta; P)]_{j,j'} = P\left(\frac{\partial \log p_\theta}{\partial [\theta]_j} \frac{\partial \log p_\theta}{\partial [\theta]_{j'}} \right)$$

for $j, j' = 1, ..., k$. For simplicity of notation, $A(\theta; P)$ and $B(\theta; P)$ are simply denoted as $A(\theta)$ and $B(\theta)$ when there is no room for misunderstanding. When their inverses exist, we define

$$C_n(\theta) = A_n(\theta)^{-1} B_n(\theta) A_n(\theta)^{-1} \quad \text{and} \quad C(\theta; P) = A(\theta)^{-1} B(\theta) A(\theta)^{-1}.$$

We assume the existence of the expectations and make some additional technical assumptions.

Assumption 4. *$\partial \log p_\theta(x)/\partial[\theta]_j, j = 1, ..., k$, are measurable of x for each $\theta \in \Theta$ and continuously differentiable functions of θ at all $x \in \mathcal{X}$.* □

Assumption 5. *$|\partial^2 \log p_\theta/\partial[\theta]_j \partial[\theta]_{j'}|$ and $|\partial \log p_\theta/\partial[\theta]_j \cdot \partial \log p_\theta/\partial[\theta]_{j'}|, j, j' = 1, ..., k$ are dominated by functions integrable with respect to P for all $x \in \mathcal{X}$ and $\theta \in \Theta$.* □

Assumption 6. *(a) θ_* is interior of Θ; (b) $B(\theta_*)$ is nonsingular; (c) θ_* is a regular point of $A(\theta)$.* □

Under these assumptions, we have asymptotic normality.

Theorem 2 (Asymptotic Normality, Theorem 3.2 of [White, 1982])**.** *Given Assumptions 1 to 6,*

$$\sqrt{n}(\hat{\theta}_n(D_n) - \theta_*) \to \mathcal{N}(0, C(\theta_*)).$$

Moreover, $C_n(\hat{\theta}_n)$ converges to $C(\theta_*)$ almost surely, element by element. □

We make some remarks on this theorem. First, this theorem says that the distribution of $\hat{\theta}_n(D_n)$ weakly converges to a certain normal distribution. A stronger convergence is not guaranteed. Second, though this theorem implies that, for some fixed P, the normal distribution that $\hat{\theta}_n$ weakly converges to is uniquely determined, it does not imply that such P is unique. Multiple QMLE sequences of $\{\hat{\theta}_n(D_n)\}$ for different data distributions can weakly converge to the same normal random variable. Roughly speaking, QMLE works as a function from the data distributions to the normal distributions, and the function is not injective.

Chapter 3

A Practical LDP Quasi-MLE

3.1 Introduction

Locally perturbations by each data provider satisfying local differential privacy ensure strong privacy protection and give a data curator an alternate database, in which privacy composition does not occur when we utilize it multiple times. In exchange for that strong protection, it becomes extremely difficult for the data curator to learn the properties of the data. If the curator wants to know its properties with confidence, the curator must use some LDP statistical tools.

LDP versions of many statistical tools have been developed, including t-tests [Ding et al., 2018] and chi-squared tests [Gaboardi and Rogers, 2018]. An LDP quasi-maximum likelihood estimator (QMLE) can also be included among these tools. QMLE is an estimator of a parameter likely approximating a distribution F generating a set of observations $D_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, from model family $\{F_\theta : \theta \in \Theta\}$. The likelihood of parameter θ is evaluated using the log-likelihood function $\ell(\theta; D_n) = \sum_{i=1}^{n} \log f_\theta(\mathbf{x}_i)/n$, and QMLE $\hat{\theta}_n$ is defined as the maximizer of $\ell(\theta; D_n)$. MLE is a special case in which there is a correct model: $F \in \{F_\theta : \theta \in \Theta\}$. Since no one observes the raw data under the LDP constraint, it is too optimistic to assume that we can specify a family including the true distribution. In this thesis, we mainly consider QMLE rather than MLE. Under regularity conditions, QMLE has asymptotic normality. By understanding its normality, the curator is able to determine how likely and by how much the estimator is to deviate from the optimal point. Moreover, with the asymptotic normality, we can perform the Wald test, which is an important application [Vaart, 2000].

Bhowmick et al. [2018] provided a framework for LDP M-estimators, which is a superclass of LDP QMLEs. It approximates the maximizer of an objective function with stochastic gradient descent. They showed that the covariance matrix of the normal distribution on which the estimator converges agrees with minimax optimal ones up to a constant.

However, the existing protocol may be difficult to deploy for a large-scale system in the real world due to the following three problems: (i) it requires a long waiting time for users, (ii) it is communication inefficient, and (iii) it requires

finiteness of the derivative of the objective function. The existing protocol is interactive wherein the communication of the ith user depends on those of the previous $i-1$ users. Though this interactivity gives more accurate statistics [Smith et al., 2017], it causes a long waiting time for users when millions of users are involved in the protocol. Communication efficiency is a non-ignorable problem for large-scale implementation, especially on Edge or IoT devices. When the parameter is d-dimensional and each component of the parameter uses float as a data type, each user submits $32d$ bits. It is also of great practical importance to be able to apply to unbounded domain data. The LDP constraints require a user to perturb her record so as to be indistinguishable from the other candidate records in the domain. An unbounded domain makes it difficult to satisfy this requirement since no one knows how many candidate values exist in the domain.

We provide low-user-side-cost protocols that involve no waiting time, require no boundedness assumption, and avoid high communication costs for QMLEs of regression. In this thesis, we focus on regression which is a wide and important class. To eliminate waiting time, we abandon interactivity. Although less accurate than interactive methods, our protocol has a significant advantage in that the execution time on the user side is constant regardless of the number of users. To remove the boundedness assumption, we incorporate truncation into the protocol. This simple technique makes it possible for the protocol to perform safely even when the record domain is unbounded. For communication efficiency, we adopt the one-bit submission strategy whereby a record is stochastically quantized into a binary value [McGregor et al., 2010, Seide et al., 2014, Bassily and Smith, 2015, Ding et al., 2018, Wang et al., 2018]. This strategy significantly reduces the communication cost. See Table 3.1 for a quick comparison of the communication costs and waiting time.

As the main contributions of this chapter, (i) we give consistency and asymptotic normality theorem with their sufficient conditions for our QMLEs, and (ii) we make explicit the limitations of the scope of our theoretical analysis. The asymptotic normality is useful for curators to adequately decide sample size n and privacy parameter ϵ. The sufficient conditions for our consistent and normality theorems are conditions on the model family and the true distribution. The curator should check the conditions for the model family when selecting the family. On the other hand, no one can evaluate the conditions on the true distribution. We recommend that the curator should carefully consider these conditions with the help of experts.

To discuss the sufficient conditions for our theorems on a concrete problem,

Table 3.1: Comparison of communication costs in number of submitting bits and waiting time of the protocols of the existing protocol [Bhowmick et al., 2018] and our protocol in two scenarios where explanatory variables $\mathbf{X}$ are public and private, d is the dimension of parameter, k is number of explanatory variables, and n is the number of users.

Id	Scenario	Server	User	Wait
Bhowmick2018	$\mathbf{X}$ pub	$32(k+d)$	$32d$	$O(n)$
	$\mathbf{X}$ pri	$32d$	$32d$	
Ours	$\mathbf{X}$ pub	$32k$	1	$O(1)$
	$\mathbf{X}$ pri	0	$k+1$	

we consider α-quantile linear regression [Davino et al., 2013]. With this example, we can see that it is not so difficult to make a model family satisfying the conditions. Given $\alpha \in (0, 1)$, coefficients estimation for α-quantile regression is one of the standard statistical data analyses and QMLE is one of the solutions. For explanatory variables $\mathbf{X}$ on $\mathbb{R}^d$ and objective variable Y on $\mathbb{R}$, the goal of the α-quantile regression is to find coefficient $\beta \in \mathcal{B} \subset \mathbb{R}^d$ such that the inner product $\beta^\top \mathbf{X}$ well approximates the α-quantile of the distribution of Y, i.e., $\inf\{y | \Pr(Y \leq y|\mathbf{X}) > \alpha\}$. If we consider asymmetric Laplace distributions as the model family, this problem is a likelihood-maximizing problem. With this example, we are able to confirm that the conditions regarding the model family are easily satisfied. In addition, using real data, we observe the asymptotic behavior of our QMLE. The observations imply that the Frobenius norm of empirical covariance matrix shrinks in proportion to $1/n$ as expected in the asymptotic normality theorem.

We mention some related works. LDP regression by non-interactive algorithms has been studied in the context of LDP empirical risk minimization e.g., [Smith et al., 2017, Zheng et al., 2017, Wang et al., 2018, 2019, 2021]. Their targets are not analyses of asymptotic normality but seeking smaller risk. The studies for non-local differentially privately M-estimators took different ways from us [Smith, 2011, Chaudhuri and Hsu, 2012, Avella-Medina, 2020]. Due to the difference in the privacy models, we do not compare our results with theirs. Bhowmick et al. [2018] showed asymptotic normality of their estimator relying on Polyak and Juditsky [1992] 's asymptotic-normality proof for the estimators obtained by stochastic gradient descent. Since we do not use stochastic gradient descent, we prove our theorem by a different method.

The remainder of the chapter is organized as follows: In Section 3.2, we introduce the notation used in this chapter and some of the basic concepts. In Section 3.3, we describe our protocols for building QMLEs. In Section 3.4, we discuss QMLE for α-quantile regression as an illustrative application of the protocol. In Section 3.5, we report the results of a numerical experiment with real data. In Section 3.6, we offer concluding remarks.

3.2 Preliminaries

We begin by defining some of the notation used in the chapter. We denote by 0_d the d-length zero vector. When we take expectation while emphasizing the distribution F, we use $Fg = \mathbb{E}_{X \sim F}[g(X)]$ where g is a function.

3.2.1 Local Differential Privacy

Local differential privacy is a rigorous privacy definition for distributed statistical analyses. The definition requires each user to protect her sensitive record individually by stochastic perturbation. In particular, we consider the case in which users receive no feedback from the curator. LDP in such a situation is called non-interactive LDP; in this thesis, we refer to non-interactive LDP simply as LDP.

We can now formally define LDP. Assume there are n users, each of whom possesses a sensitive record R_i for $i = 1, \cdots, n$. Let $\mathcal{R}$ be the domain of the records. Assume that there is also a curator who will perform a statistical anal-

ysis on the users' records and that each user will submit her perturbed record to the curator. We can define the perturbation as a conditional distribution $Q(\cdot|R=r)$ and LDP as a property of Q. Perturbation $Q(\cdot|R=r)$ is a distribution on set $\mathcal{Z}$.

Definition 1 (ϵ-LDP)**.** *Given $\epsilon > 0$, distribution Q is ϵ-locally differentially private if, for any $r, r' \in \mathcal{R}$,*

$$\sup_{\mathcal{S}\in\sigma(\mathcal{Z})} Q(S|R=r) \leq e^{\epsilon} Q(S|R=r'),$$

where $\sigma(\mathcal{Z})$ is a σ-algebra on $\mathcal{Z}$.

This definition requires that the conditional distributions $Q(\cdot|r)$ and $Q(\cdot|r')$ are not so different from each other for any pair r, r' of records in $\mathcal{R}$. The ϵ represents the similarity of the conditional distributions. A smaller ϵ implies stricter privacy protection but less information of the outputs. ϵ thus controls the trade-off between privacy protection and utility.

We use the bit flip [Ding et al., 2018] for the concrete implementation of conditional distribution Q satisfying ϵ-LDP. The bit flip stochastically maps a finite continuous interval $[c_{\mathrm{l}}, c_{\mathrm{u}}]$, where c_{l} and c_{u} are some real constants such that $c_{\mathrm{l}} < c_{\mathrm{u}}$, into discrete binary values $\{z_-, z_+\}$. Then, for any input $v \in [c_{\mathrm{l}}, c_{\mathrm{u}}]$ and with $C_\epsilon = \frac{e^\epsilon+1}{e^\epsilon-1}$, the bit flip is defined as

$$Q_{\mathrm{bf}}(Z=z|v) = \begin{cases} \frac{1}{2} - \frac{v - \frac{c_{\mathrm{u}}+c_{\mathrm{l}}}{2}}{(c_{\mathrm{u}}-c_{\mathrm{l}})C_\epsilon} & \text{if } z = z_-, \\ \frac{1}{2} + \frac{v - \frac{c_{\mathrm{u}}+c_{\mathrm{l}}}{2}}{(c_{\mathrm{u}}-c_{\mathrm{l}})C_\epsilon} & \text{if } z = z_+. \end{cases}$$

When the input is close to c_{u}, the output is likely to be z_+; conversely, when the input is close to c_{l}, the output is likely to be z_-.

3.2.2 Quasi-Maximum Likelihood Estimator

Given observations $D_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ generated by distribution F, the likelihood of parameter θ of a model F_θ is evaluated by the log-likelihood function

$$\ell(\theta; D_n) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(\mathbf{x}_i),$$

where f_θ is the density function of F_θ. Roughly speaking, the log-likelihood is the log of the probability that the observations are obtained assuming they are sampled from F_θ. For the likelihood function, QMLE $\hat{\theta}_n$ is defined as $\hat{\theta}_n = \mathrm{argmax}_{\theta\in\Theta} \ell(\theta; D_n)$. Not only D_n but also $\hat{\theta}_n$ itself is a random variable.

In this subsection, we review the consistency and asymptotic normality theorems of QMLEs by White [1982]. To define the log-likelihood function well, we first need to make some assumptions. The first is that the observations are independently generated from a distribution F and that F has a regular Radon–Nikodym density function f. The second condition requires that the model family also has regular density functions.

Assumption 7. *Let ν be an appropriate measure on $\mathcal{X}$. For a constant k, the independent $1 \times k$ random vectors $\mathbf{X}_i, i = 1, \cdots, n$, have common joint distribution function F on $\mathcal{X}$, a measurable Euclidean space, with measurable Radon–Nikodym density $f = dF/d\nu$.*

Assumption 8. *The family of distribution functions $F_\theta(\mathbf{x})$ has Radon–Nikodym densities $f_\theta(\mathbf{x}) = dF_\theta(\mathbf{x})/d\nu$ which are measurable in x for every $\theta \in \Theta$, a compact subset of a d-dimensional Euclidean space, and continuous in θ for every $\mathbf{x} \in \mathcal{X}$.*

To guarantee consistency, we introduce an additional technical assumption.

Assumption 9. *(a) $F \log f$ exists, and $|\log f_\theta(\mathbf{x})| \leq h(\mathbf{x})$ for all $\theta \in \Theta$, where h is integrable with respect to F; (b) $F \log f_\theta$ has a unique maximum at $\theta^* \in \Theta$.*

Under these regularity conditions, the QMLE converges to $\theta^* = \operatorname{argmax}_{\theta\in\Theta} F \log f_\theta$.

Theorem 3 (Theorem 2.2 in [White, 1982])**.** *Given Assumptions 7 to 9, $\hat{\theta}_n \to \theta^*$ as $n \to \infty$ for almost every sequence $\{\mathbf{X}_i\}_{i=1}^n$.*

We also have asymptotic normality under some additional assumptions regarding the existence of scores $\partial \log f_\theta(\mathbf{x})/\partial\theta$ and related quantities.

Assumption 10. *$\partial \log f_\theta(\mathbf{x})/\partial\theta_j, j = 1, ..., d$, are measurable of $\mathbf{x}$ for each $\theta \in \Theta$ and continuously differentiable functions of θ for each $\mathbf{x} \in \mathcal{X}$.*

Assumption 11. *$|\partial \log f_\theta(\mathbf{x})/\partial\theta_{j_1} \cdot \partial \log f_\theta(\mathbf{x})/\partial\theta_{j_2}|$ and $|\partial^2 \log f_\theta(\mathbf{x})/\partial\theta_{j_1}\partial\theta_{j_2}|$, for $j_1, j_2 = 1, \ldots, d$ are dominated by functions integrable with respect to F for all $\mathbf{x}$ in $\mathcal{X}$ and θ in Θ.*

Assumption 12. *(a) θ^* is interior to Θ; (b) $B(\theta) = (F(\partial \log f_\theta/\partial\theta)(\partial \log f_\theta/\partial\theta)^\top$ is nonsingular at $\theta = \theta^*$; (c) θ^* is a regular point of $A(\theta) = F\partial^2 \log f_\theta/\partial\theta^2$.*

The following shows the asymptotic normality.

Theorem 4 (Theorem 3.2 in [White, 1982])**.** *Given Assumptions 7 to 12,*

$$\begin{aligned} &\sqrt{n}(\hat{\theta}_n - \theta^*) \to \mathcal{N}(0, C(\theta^*)) \\ \textit{where} \quad &C(\theta) = A(\theta)^{-1} B(\theta) A(\theta)^{-1}. \end{aligned}$$

When $F_{\theta^*} = F$, $C(\theta^*)$ is called the Fisher information matrix.

3.2.3 Quantile Regression

Linear quantile regression deals with the statistical problem of finding coefficients $\beta \in \mathcal{B} \subset \mathbb{R}^d$ such that, given $\mathbf{x}$, the inner product $\beta^\top \mathbf{x}$ well approximates the α-quantile $\inf\{y | F(Y \leq y|\mathbf{x}) > \alpha\}$ of $Y|\mathbf{x}$. The problem is often formulated as an optimization problem finding $\beta \in \mathcal{B}$ that minimizes the following objective function: Given observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$,

$$\sum_{i=1}^{n} \rho_\alpha(y_i - \beta^\top \mathbf{x}_i) \text{ where } \rho_\alpha(\tau) = \begin{cases} (\alpha - 1)\tau & \text{if } \tau \leq 0, \\ \alpha\tau & \text{if } \tau > 0. \end{cases} \tag{3.1}$$

ρ_α is a convex function, which is called the check loss.

If we assume that objective variable Y is sampled from the asymmetric Laplace distribution defined below, the minimization of (3.1) is equivalent to the likelihood maximization for the parameter of the distributions: With $\sigma > 0$,

$$f_Y(y; \alpha, \mu, \sigma) = \frac{\alpha(1-\alpha)}{\sigma} \exp\left(-\rho_\alpha\left(\frac{y-\mu}{\sigma}\right)\right). \tag{3.2}$$

Hence the log-likelihood function is written as

$$\frac{1}{n}\sum_{i=1}^{n}\log f_Y(y_i;\alpha,\beta^\top\mathbf{x}_i,\sigma)=\log\frac{\alpha(1-\alpha)}{\sigma}-\frac{1}{n\sigma}\sum_{i=1}^{n}\rho_\alpha\left(y_i-\beta^\top\mathbf{x}_i\right). \quad (3.3)$$

Finally, we revisit the classical result of the asymptotic normality of the MLE. Let $\hat{\beta}_n \in \mathcal{B}$ be the MLE that minimizes (3.3), and let β^* be the coefficient such that $F(Y \le y|\mathbf{X}=\mathbf{x}) = F_Y(y;\alpha,{\beta^*}^\top\mathbf{x},\sigma)$ for almost every $\mathbf{x}$ and y with appropriate α and σ. Then, the sequence of MLEs $\{\hat{\beta}_n\}_n$ converges as

$$\sqrt{n}(\hat{\beta}_n-\beta^*)\to\mathcal{N}(0_d,I^{-1}), \quad (3.4)$$

where $\mathcal{N}(0_d, I^{-1})$ is the normal distribution whose mean and covariance are 0_d and I^{-1}, respectively [Davino et al., 2013]. Assuming that $\mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]$ is non-singular, I is the Fisher information matrix defined as

$$I=\frac{\alpha(1-\alpha)}{\sigma^2}\mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]. \quad (3.5)$$

3.3 Proposed Protocol

We provide two protocols for building QMLEs of regression in two different privacy scenarios and give their asymptotic normality theorem. Then, we remark on their advantages, limitations, and possible future works.

3.3.1 Regression with Public X

In this subsection, we consider regression with sensitive objective variable Y and public explanatory variables $\mathbf{X}$. This situation may seem strange, but we will give a practical use case. Consider a situation in which a company is planning to conduct a customer opinion survey on a new product. The company can control its features set $\mathbf{X}$ and gives a new product with certain features $\mathbf{X}=\mathbf{x}$ to each customer. The customer gives an evaluation Y for $\mathbf{X}=\mathbf{x}$. The target of the company is to understand the conditional distribution of Y. In the survey, the company knows the $\mathbf{X}$s and their distribution, and they are public.

The system model is as follows: There are a single curator and n users. The curator selects distribution $F_{\mathbf{X}}$ on $\mathcal{X}\subset\mathbb{R}^k$, a measurable Euclidean space, generates $\mathbf{X}_i$ for each user $i=1,\cdots,n$ following $F_{\mathbf{X}}$, and passes them to each user. Given $\mathbf{X}_i=\mathbf{x}_i$, user i independently generates Y_i following unknown conditional distribution $F(\cdot|\mathbf{x}_i)$ on $\mathcal{Y}\subset\mathbb{R}$, a measurable space, and truncates it into interval $[c_\mathrm{l}, c_\mathrm{u}]$. Let $\bar{Y}_i$ be the truncated version of Y_i:

$$\bar{Y}_i=t(Y_i)\equiv\begin{cases}c_\mathrm{l} & \text{if } Y_i\le c_\mathrm{l},\\ Y_i & \text{if } c_\mathrm{l}<Y_i<c_\mathrm{u},\\ c_\mathrm{u} & \text{if } Y_i\ge c_\mathrm{u}.\end{cases} \quad (3.6)$$

We let $\bar{y}_i$ be a realization of $\bar{Y}_i$. Then, the user perturbs $\bar{y}_i$ by the bit flip. Z_i that is perturbed $\bar{Y}_i$ distributes as

$$p(Z_i=z|\mathbf{X}_i=\mathbf{x})=\int Q_{\mathrm{bf}}(z|t(y))dF(y|\mathbf{x}). \quad (3.7)$$

User i submits z_i which is a realization of Z_i to the curator. The user submission is always only one bit.

The curator considers model family $\{F_\beta(\cdot|\mathbf{x}) : \beta \in \mathcal{B}, \mathbf{x} \in \mathcal{X}\}$ that consists of conditional distributions parameterized by $\mathcal{B}$, a compact subset of a d-dimensional Euclidean space. For each $\beta \in \mathcal{B}$, we define conditional density function $p_\beta(z|\mathbf{x})$ by replacing F by F_β in (3.7). The target of the curator is to find β such that P_β well approximates P. In this subsection, we write P and P_β to designate joint distributions $P(\mathbf{x}, z)$ and $P_\beta(\mathbf{x}, z)$ rather than conditional distributions $P(z|\mathbf{x})$ and $P_\beta(z|\mathbf{x})$.

Given observations $D_n = \{(z_i, \mathbf{x}_i)\}_{i=1}^n$, the log-likelihood function is defined as

$$\begin{aligned}\ell(\beta; D_n) \equiv& \frac{1}{n}\sum_{i=1}^n \log p_\beta(\mathbf{x}_i, z_i) \\ =& \frac{1}{n}\sum_{i=1}^n (z_i \log \Lambda_\epsilon(\beta, \mathbf{x}_i) + (1 - z_i)\log(1 - \Lambda_\epsilon(\beta, \mathbf{x}_i)) + \log F_{\mathbf{X}}(\mathbf{x}_i))\end{aligned}$$

where $\Lambda_\epsilon(\beta, \mathbf{x}) = p_\beta(z = 1|\mathbf{x})$. We define $\hat{\beta}_n = \arg\max_{\beta\in\mathcal{B}} \ell(\beta; D_n)$ and $\beta^* = \arg\max_{\beta\in\mathcal{B}} P \log p_\beta$. The model selection and optimization are performed by the curator, and the users do not have to care about them. The curator can change hyperparameters excepting $c_{\mathrm{u}}, c_{\mathrm{l}}$ and ϵ and can try multiple model families without any additional cost for the users. The pseudo-code is included in Section 3.7.

Now, we analyze the behavior of $\hat{\beta}_n$. To derive the consistency of our QMLE, we replace F and F_θ in Theorem 3 with P and P_β, respectively. We find the conditions under which Assumptions 7 to 9 are satisfied while replacing F and F_θ with P and P_β. To satisfy Assumptions 7 and 8, we introduce the following assumptions.

Assumption 13. *Conditional distribution $F(\cdot|\mathbf{x})$ has a Radon–Nikodym density function $f(y|\mathbf{x}) = dF(y|\mathbf{x})/d\nu$ which is measurable in y for every $\mathbf{x} \in \mathcal{X}$.*

Assumption 14. *$F_{\mathbf{X}}$ has a measurable Radon-Nikodym density $f_{\mathbf{X}} = dF_{\mathbf{X}}/d\mu$ with some appropriate measure μ.*

Assumption 15. *The family of distribution functions $F_\beta(y|\mathbf{x})$ has Radon–Nikodym densities $f_\beta(y|\mathbf{x}) = dF_\beta(y|\mathbf{x})/d\nu$ which are measurable in y for every $\mathbf{x} \in \mathcal{X}$ and $\beta \in \mathcal{B}$, and continuous in β for every $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$.*

These assumptions are satisfied with many distributions e.g., Gaussian and Bernoulli distributions. From these assumptions, it is obvious that $P(\mathbf{x}, z), P_\beta(\mathbf{x}, z)$ are measurable and that the density functions $p(\mathbf{x}, z) = p(z|\mathbf{x})f(\mathbf{x})$ and $p_\beta(\mathbf{x}, z) = p_\beta(z|\mathbf{x})f(\mathbf{x})$ exist.

In order for the QMLE for regression parameter to satisfy Assumption 9, we consider the following two conditions. The first one is the existence of $P \log p$ and integrable function $h(\mathbf{x}, z)$ such that $|\log p_\beta(\mathbf{x}, z)| \le h(\mathbf{x}, z)$ for all $\beta \in \mathcal{B}$. $P \log p$ can be extended as

$$P \log p = F_{\mathbf{X}}(P_{\cdot|\mathbf{X}} \log p(\cdot|\mathbf{X}) + \log f_{\mathbf{X}}(\mathbf{X})).$$

Since $\log p(\cdot|\mathbf{X})$ is always bounded away from $-\infty$ and $+\infty$ by the following lemma, $\log p(\cdot|\mathbf{X})$ is always integrable with respect to P.

Lemma 3. *The value of* $\Lambda_\epsilon(\beta, \mathbf{x})$ *is bounded away from* 0 *and* 1*, for all* $\beta \in \mathcal{B}$ *and* $\mathbf{x} \in \mathcal{X}$*.*

See Section 3.8.1 for the proof. Thus, if $F_{\mathbf{X}} \log f_{\mathbf{X}}$ exists, $P \log p$ also exists. Similarly, the existence of $P \log p_\beta$ depends on the existence of $F_{\mathbf{X}} \log f_{\mathbf{X}}$.

Assumption 16. $F_{\mathbf{X}} \log f_{\mathbf{X}}$ *exists.*

The second condition relates to the uniqueness of the maximum of the log-likelihood function. Because the maxima are not always unique, we adopt the following assumption.

Assumption 17. $P \log p_\beta$ *has a unique maximum.*

We now have consistency.

Theorem 5. *Suppose Assumptions 13 to 17 hold. Then,* $\hat{\beta}_n \to \beta^*$ *as* $n \to \infty$ *surely.*

Next, we derive the asymptotic normality. We find the conditions under which Assumptions 10 to 12 are satisfied. Assumption 10 specifies the continuous differentiability of $\partial \log p_\beta / \partial \beta$. The partial derivative is extended as

$$\frac{\partial}{\partial \beta} \log(p_\beta(\mathbf{x}, z)) = \frac{(2z-1)\Lambda'_\epsilon(\beta, \mathbf{x})}{\Lambda_\epsilon(\beta, \mathbf{x})^z (1 - \Lambda_\epsilon(\beta, \mathbf{x}))^{1-z}}$$

where $\Lambda'_\epsilon(\beta, \mathbf{x}) = \partial \Lambda_\epsilon(\beta, \mathbf{x}) / \partial \beta$. By Lemma 3, the following is sufficient to satisfy the requirement.

Assumption 18. *Each element of* $\Lambda'_\epsilon(\beta, \mathbf{x})$ *is measurable of* $\mathbf{x}$ *for each* $\beta \in \mathcal{B}$ *and continuously differentiable functions of* β *for each* $\mathbf{x} \in \mathcal{X}$*.*

Assumption 11 states that $|\partial^2 \log p_\beta / \partial \beta_{j_1} \partial \beta_{j_2}|$ and $|\partial \log p_\beta / \partial \beta_{j_1} \cdot \partial \log p_\beta / \partial \beta_{j_2}|$ for $j_1, j_2 = 1, \cdots, d$ are bounded by functions integrable with respect to P. To verify this, we extend these values.

$$\begin{aligned}\frac{\partial^2 \log p_\beta(\mathbf{x}, z)}{\partial \beta^2} =& (2z-1) \frac{\Lambda''_\epsilon(\beta, \mathbf{x})}{\Lambda_\epsilon(\beta, \mathbf{x})^z (1 - \Lambda_\epsilon(\beta, \mathbf{x}))^{1-z}} \\ &- \frac{\Lambda'_\epsilon(\beta, \mathbf{x}) \Lambda'_\epsilon(\beta, \mathbf{x})^\top}{\Lambda_\epsilon(\beta, \mathbf{x})^{2z} (1 - \Lambda_\epsilon(\beta, \mathbf{x}))^{2(1-z)}}\end{aligned}$$

where $\Lambda''_\epsilon(\beta, \mathbf{x}) = \partial^2 \Lambda_\epsilon(\beta, \mathbf{x}) / \partial \beta^2$, and

$$\left(\frac{\partial}{\partial \beta} \log p_\beta(\mathbf{x}, z)\right) \left(\frac{\partial}{\partial \beta} \log p_\beta(\mathbf{x}, z)\right)^\top = \frac{\Lambda'_\epsilon(\beta, \mathbf{x}) \Lambda'_\epsilon(\beta, \mathbf{x})^\top}{\Lambda_\epsilon(\beta, \mathbf{x})^{2z} (1 - \Lambda_\epsilon(\beta, \mathbf{x}))^{2(1-z)}}.$$

The denominators are always non-zero by Lemma 3. Thus, the following assumption is sufficient to satisfy the requirement.

Assumption 19. *The absolute values of each element of* $\Lambda'_\epsilon(\beta, \mathbf{x})$ *and* $\Lambda''_\epsilon(\beta, \mathbf{x})$ *are bounded by integrable functions with respect to* P*.*

Assumption 12 consists of three parts. The first part is that β^* is interior to $\mathcal{B}$. We assume this.

Assumption 20. *β^* is interior to $\mathcal{B}$.*

The second part is the non-singularity of $P((\partial \log p_\beta/\partial\beta)(\partial \log p_\beta/\partial\beta)^\top)$ at $\beta = \beta^*$.

$$P\left(\frac{\partial}{\partial\beta}\log p_\beta\right)\left(\frac{\partial}{\partial\beta}\log p_\beta\right)^\top = F_{\mathbf{X}}\left(\frac{p(Z=1|\mathbf{X})}{\Lambda_\epsilon(\beta,\mathbf{X})^2} + \frac{p(Z=0|\mathbf{X})}{(1-\Lambda_\epsilon(\beta,\mathbf{X}))^2}\right)\Lambda'_\epsilon(\beta,\mathbf{X})\Lambda'_\epsilon(\beta,\mathbf{X})^\top.$$

Thus, the following assumption is a sufficient condition of the requirement.

Assumption 21. *$F_{\mathbf{X}}\Lambda'_\epsilon(\beta^*,\mathbf{X})\Lambda'_\epsilon(\beta^*,\mathbf{X})^\top$ is non-singular.*

The third part is non-singularity of $P\partial^2 \log p_\beta/\partial\beta^2$ at $\beta = \beta^*$. We obtain this from Assumption 17. If $P \log p_\beta$ has a second partial derivative along β and β^* is interior to $\mathcal{B}$, then $\partial^2 P \log p_\beta/\partial\beta^2$ must be negative-definite. If not, there exists β' such that $P \log p_{\beta'} = P \log p_{\beta^*}$ and $\beta' \neq \beta^*$. Finally, we obtain asymptotic normality.

Theorem 6. *Suppose Assumptions 13 to 21 hold. Then, $\sqrt{n}(\hat{\beta}_n - \beta^*) \to \mathcal{N}(0_d, C(\beta^*))$ where $C(\beta) = A^{-1}(\beta)B(\beta)A^{-1}(\beta)$ with $A(\beta) = P\partial^2 \log p_\beta/\partial\beta^2$ and $B(\beta) = P(\partial \log p_\beta/\partial\beta)(\partial \log p_\beta/\partial\beta)^\top$.*

3.3.2 Regression with Private X

Next, we consider regression when both objective variables and explanatory variables are sensitive and are submitted with perturbation. The system model is that each user i generates $\mathbf{X}_i$ following unknown distribution $F_{\mathbf{X}}$ and then generates Y_i following unknown conditional distribution $F(\cdot|\mathbf{X}_i)$.

The communication protocol is as follows. User i stochastically perturbs $\mathbf{X}_i$ and Y_i by LDP mechanism Q. We denote the perturbed ones by $\mathbf{Z}^{(\mathbf{X})}$ and $Z^{(Y)}$, respectively. Q consists of $Q_{Z^{(Y)}}$ and $Q_{\mathbf{Z}^{(\mathbf{X})}}$ perturbing Y_i and $\mathbf{X}_i$, respectively. The privatized objective variable $Z^{(Y)}$ is the same as Z in the previous subsection without the privacy budget consumed by the LDP mechanisms. On the other hand, since $\mathbf{Z}^{(\mathbf{X})}$ was not defined in the previous section, we need to define $Q_{\mathbf{Z}^{(\mathbf{X})}}$. We use the bit flip as $Q_{\mathbf{Z}^{(\mathbf{X})}}$ in an element-wise manner. Each element is randomized with privacy budget $\epsilon/(k+1)$. The total consumption of the privacy budget per user does not exceed ϵ by the sequential composition theorem [McSherry, 2009]. We set the domain of $Q_{\mathbf{Z}^{(\mathbf{X})}}$ to $\{-1,+1\}^k$. For each $\mathbf{z}^{(\mathbf{X})} \in \{-1,+1\}^k$,

$$Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})}|\mathbf{x}) = \prod_{j=1}^{k}\left(\frac{1}{2} + \frac{t(x_j)z_j^{(\mathbf{X})}}{2C_{\epsilon/(k+1)}}\right), \tag{3.8}$$

where $t(\cdot)$ is defined in (3.6). The generated privatized variables $(\mathbf{z}_i^{(\mathbf{X})}, z_i^{(Y)})$ are submitted to the curator.

In the communication protocol, each user submits $(k+1)$ bits to the curator, and the curator sends no information to the users. This privacy scenario is nearly the same as the Bhowmich's one [Bhowmick et al., 2018], and our communication protocol is more efficient than theirs. In their protocol, each user receives and

submits d float or double values, either $64d$ bits or $144d$ bits. Thus, our protocol results in communication costs that are roughly 64 or 144 times smaller than their protocol when $k \leq d$.

The curator defines model family $\{F_\beta(y|\mathbf{x}) : \beta \in \mathcal{B}, \mathbf{x} \in \mathcal{X}\}$ and provisional distribution $\hat{F}_{\mathbf{X}}$. Though the true $F_{\mathbf{X}}$ is unknown, the curator must assume some distribution of $\mathbf{X}$ to compute the log-likelihood function, as we will see later. $\hat{F}_{\mathbf{X}}$ is a kind of prior distribution.

Since the discussion of consistency and asymptotic normality has much in common with the previous subsection, here we describe only the differences. See Section 3.8.2 for details. Given observations $D_n = \{(\mathbf{z}_i^{(\mathbf{X})}, z_i^{(Y)})\}_{i=1}^n$, the likelihood function is

$$\begin{aligned}
\ell(\beta; D_n) =& \frac{1}{n}\sum_{i=1}^{n} \big(\log \hat{p}_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}_i^{(\mathbf{X})}) + z_i^{(Y)} \log \Phi(\beta, \mathbf{z}_i^{(X)}) \\
&+ (1 - z_i^{(Y)}) \log(1 - \Phi(\beta, \mathbf{z}_i^{(\mathbf{X})})) \big) \\
\text{where} \quad & \hat{p}_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})}) \equiv \int Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})}|\mathbf{x}) d\hat{F}_{\mathbf{X}}(\mathbf{x}), \\
& \Phi(\beta, \mathbf{z}^{(\mathbf{X})}) \equiv \frac{\hat{F}_{\mathbf{X}}(\Lambda_{\epsilon/(d+1)}(\beta, \mathbf{X}) Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{Z}^{(\mathbf{X})}|\mathbf{X})}{\hat{p}_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})})}.
\end{aligned}$$

QMLE $\hat{\beta}_n$ is defined as $\hat{\beta}_n \equiv \operatorname{argmin}_{\beta \in \mathcal{B}} \ell(\beta; D_n)$.

We can show consistency based on Theorem 3 under the assumption that the curator chooses a regular distribution as $\hat{F}_{\mathbf{X}}$.

Assumption 22. *$\hat{F}_{\mathbf{X}}$ has a measurable Radon-Nikodym density $\hat{f}_{\mathbf{X}} = d\hat{F}_{\mathbf{X}}/d\mu$.*

Theorem 7. *Suppose Assumptions 13 to 15, 17 and 22 hold. Then, $\hat{\beta}_n \to \beta^*$ as $n \to \infty$ for almost every sequence $\{(\mathbf{Z}_i^{(\mathbf{X})}, Z_i^{(Y)})\}_i$.*

For details, see Section 3.8.2. This consistent theorem does not require the existence of $F_{\mathbf{X}} \log f_{\mathbf{X}}$ unlike Theorem 5. We can obtain the existence from Assumption 22 and the properties of $\hat{p}_{\mathbf{Z}^{(\mathbf{X})}}$. The discretization by the bit flip relaxes the integrable condition.

To show asymptotic normality, we adopt several additional assumptions.

Assumption 23. *$\Phi'(\beta, \mathbf{z}^{(\mathbf{X})})$ is continuous differentiable function of β.*

Assumption 24. *Each component of $\Phi''(\beta, \mathbf{z}^{(\mathbf{X})})$ and $(\Phi'(\beta, \mathbf{z}^{(\mathbf{X})}))(\Phi'(\beta, \mathbf{z}^{(\mathbf{X})}))^\top$ is bounded by integrable functions with respect to P.*

Assumption 25. *$\mathbb{E}_{\mathbf{Z}^{(\mathbf{X})}} \left[(\Phi'(\beta, \mathbf{Z}^{(\mathbf{X})})(\Phi'(\beta, \mathbf{Z}^{(\mathbf{X})}))^\top \right]$ is non-singular at $\beta = \beta^*$.*

Assumption 23 is used to prove the requiment corresponding to Assumption 10. The requirement corresponding to Assumption 11 is satisfied with Assumption 24, which requires that the curator should design Φ such that its first and second derivatives almost surely take finite values. The requirement corresponding to Assumption 12 is satisfied with Assumptions 17, 20 and 25. We now have asymptotic normality.

Theorem 8. *Suppose Assumptions 13 to 15, 17, 20 and 22 to 25 hold. Then, $\sqrt{n}(\hat{\beta}_n - \beta^*) \to \mathcal{N}(0_d, C(\beta^*))$ where $C(\beta) = A^{-1}(\beta)B(\beta)A^{-1}(\beta)$ with $A(\beta) = P\partial^2 \log p_\beta / \partial\beta^2$ and $B(\beta) = P(\partial \log p_\beta / \partial\beta)(\partial \log p_\beta / \partial\beta)^\top$.*

3.3.3 Remark and Limitation

The assumptions for proving consistency and asymptotic normality in Theorems 5 to 8 are not relevant to privacy preservation. Even if those assumptions do not hold, users' privacy is still protected as long as the ϵ-LDP mechanisms correctly work. The users who supply data do not need to worry about these assumptions at all.

The requirements of our theorems clarify the properties of the model that the curator should check. The curator is free to choose any linear or non-linear model as long as it satisfies these properties. In addition, those requirements place few restrictions on model selection since the curator can modify the model after data collection.

As we see in Section 3.4, it is not so difficult to craft a model satisfying the requirements. We thus expect that most standard regression models satisfy them.

The first limitation relates to the problem of choosing $\hat{F}$. Although any $\hat{F}$ satisfying Assumptions 16 and 22 can be acceptable, a poor choice of $\hat{F}$ may make it difficult to satisfy the other assumptions. The theorem provides no method for choosing a better $\hat{F}$, which remains an open problem.

The second limitation relates to the true distribution, which is a common problem in most statistical theories. We have no method to evaluate Assumptions 13, 17 and 20. The curator never know the exact value of β^* and $C(\beta^*)$. The curator should carefully consider these assumptions with the help of experts.

The exploration of better mechanisms is our future work. There may exist $Q_{\mathbf{Z}(\mathbf{x})}$ giving us a more sharp covariance matrix. In the context of LDP, vector submission is studied by many researchers e.g., [Duchi et al., 2013, Erlingsson et al., 2014, Bassily and Smith, 2015, Wang et al., 2019].

Better selection of c_{l} and c_{u} is another future work. Whether certain c_{l} and c_{u} are good or bad strongly depends on F, and we have no general strategy to select better c_{l} and c_{u}.

One of the potential applications of our algorithms is bootstrapping. In the above subsections, we described that our algorithms output only one estimator in each protocol. However, without additional privacy loss, the curator can compute many estimators using the subsets of the submitted data. The post-processing invariant enables us to perform such an operation. This is one of the advantages of a non-interactive algorithm.

Another potential application is a misspecification test to determin whether the model family contains the true distribution [White, 1982]. In the LDP setting, since the raw data are distributed, no single entity has knowledge on the statistical properties of the raw data. It is difficult to evaluate whether a model family is appropriate. A curator performs the test as a preliminary experiment. The results of the test would help the curator to quantitatively assess the confidence level of the main survey.

3.4 Example: Quantile Regression

In this section, we show the QMLEs for quantile regression as a concrete example of our QMLEs. One of the main goals of this section is to show that it is possible

to replace some of the assumptions noted in the previous section with a concrete implementation of the model. We note that the notation used in Section 3.4.1 and Section 3.4.2 is the same as that used in Section 3.3.1 and Section 3.3.2. Here, $k = d$.

3.4.1 With Public X

As described in Section 3.2.3, we can formulate the α-quantile regression as a quasi-maximum likelihood estimation problem. For some $\sigma > 0$, we set f_β as

$$f_\beta(y|\mathbf{x}) = \frac{\alpha(1-\alpha)}{\sigma} \exp\left(-\rho_\alpha\left(\frac{y - \beta^\top \mathbf{x}}{\sigma}\right)\right)$$

for each y and $\mathbf{x}$, where ρ_α is defined in (3.1). This construction satisfies Assumption 15: measurable and continuous.

When we choose the product of independent d uniform distributions on interval $[-1, +1]$ as $F_{\mathbf{X}}$, Assumption 16 is satisfied.

Let Ψ_ϵ be the function such that $\Lambda_\epsilon(\beta, \mathbf{x}) = \Psi_\epsilon(\beta^\top \mathbf{x})$. Then, $\Lambda'_\epsilon(\beta, \mathbf{x}) = \Psi'_\epsilon(\beta^\top \mathbf{x})\mathbf{x}$ and $\Lambda''_\epsilon(\beta, \mathbf{x}) = \Psi''_\epsilon(\beta^\top \mathbf{x})\mathbf{x}\mathbf{x}^\top$ where $\Psi'_\epsilon(\theta) = \partial\Psi_\epsilon(\theta)/\partial\theta$ and $\Psi''_\epsilon(\theta) = \partial^2\Psi_\epsilon(\theta)/\partial\theta^2$. It has the following property.

Lemma 4. *$\Psi_\epsilon(\theta)$ is a strictly monotonically increasing function and is bounded away from 0 and 1. $\Psi'_\epsilon(\theta)$ and $\Psi''_\epsilon(\theta)$ exist and for any $\theta \in \mathbb{R}$, and their absolute values are bounded.*

See Section 3.8.3 for the proof. From the second part of Lemma 4, Assumptions 18 and 19 are satisfied. Mover, $F_{\mathbf{X}}(\mathbf{X}\mathbf{X}^\top)$ is a non-singular matrix since

$$F_{\mathbf{X}} X_{j_1} X_{j_2} = \begin{cases} 0 & \text{if } j_1 \neq j_2, \\ \frac{1}{3} & \text{if } j_1 = j_2. \end{cases}$$

Thus, Assumption 21 is satisfied.

As a consequence of Theorems 5 and 8, we have the following corollaries.

Corollary 1. *Suppose Assumptions 13, 17 and 20 hold. Then, $\hat{\beta}_n \to \beta^*$ almost surely and $\sqrt{n}(\hat{\beta}_n - \beta^*) \to \mathcal{N}(0_d, C(\beta^*))$.*

To prove this corollary, we need only three assumptions. The concrete constructions of the model remove some of the assumptions used in Theorems 5 and 8.

Although the accuracy of our QMLEs is not a focus of this chapter, we did conduct a rough comparison of accuracy with existing works. As a result, we found with $\epsilon \downarrow 0$, the Fisher information of our MLEs is $\sigma^2/\alpha(1-\alpha)$ times smaller than the upper bound shown in [Barnes et al., 2020]. For details, see Section 3.9.

3.4.2 With Private X

In this setting, the curator does not know $F_{\mathbf{X}}$. Instead of $F_{\mathbf{X}}$, we adopt the product distribution of d symmetric binary distributions on $\{-1, +1\}$. Then, Assumption 22 is satisfied.

With $\epsilon' = \epsilon/(d+1)$, Φ is extended as

$$\Phi(\beta, \mathbf{z}^{(\mathbf{X})}) = \frac{\sum_{\mathbf{x}\in\{\pm 1\}^d} \Psi_{\epsilon'}(\beta^\top \mathbf{x}) \exp\left(\epsilon' \mathbf{1}\left[z_j^{(\mathbf{X})} = x_j\right]\right)}{p_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})})(e^{\epsilon'}+1)^d 2^d},$$

where $\Psi_{\epsilon'}$ is defined in the previous subsection. Due to the properties of $\Psi_{\epsilon'}$, which we evaluated in the previous subsection, Assumptions 23 and 24 are obviously satisfied. By the monotonicity of $\Psi_{\epsilon'}$, Assumption 25 is also satisfied. Now, as a corollary of Theorems 7 and 8, we obtain the following result.

Corollary 2. *Suppose Assumptions 13, 17 and 20 hold. Then,* $\sqrt{n}(\hat{\beta}_n - \beta^*) \to \mathcal{N}(0_d, C(\beta^*))$ *where* $C(\beta) = A^{-1}(\beta)B(\beta)A^{-1}(\beta)$.

3.5 Numerical Evaluation

In this section, we observe the behavior of our QMLE for real data. We consider the QMLE for quantile regression in the public $\mathbf{X}$ case. Since we do not know the true distribution generating the real data, we cannot perform exact comparisons with the theoretical result, Corollary 3. Here, we observe the empirical covariance of the QMLEs to evaluate the convergence of the distribution of the QMLE. For additional numerical evaluations, see Section 3.10.

We numerically compare the covariance matrices with varying n and ϵ. We use CO and NOx emission data set [Kaya et al., 2019], which consists of $36,733$ records of 11 sensors attached to a turbine of a power plant. Although this data is not sensitive, we chose this data because of its large number of records and its format. We treat the 11th column as y and treat the columns from the first to 9th as $\mathbf{x}$. We set $c_{\mathrm{u}} = 110, c_{\mathrm{l}} = 40, \sigma = 1.0$ and $\alpha = 0.3$. These specific values of hyperparameters do not have a particular meaning. We vary n from $5,000$ to $35,000$ in increments of $5,000$ for $\epsilon \in \{1, 2.5, 5, 10\}$. For each combination of n and ϵ, we sub-sample n records $1,000$ times without replacement from the $36,733$ records. For each sub-data, we perturb ys and compute a QMLE as descrived in Section 3.4.1. With the $1,000$ QMLEs, we obtain the empirical covariance matrix and its Frobenius norm. We implemented the simulations with Python 3.9.2, NumPy 1.19.2, and SciPy 1.6.1. The Python code is contained in the supplementary material.

Figure 3.1 shows the result. The horizontal and vertical axes show n and the value of each Frobenius norm in log-scale, respectively. For each ϵ, with large n, the norm of the covariance matrix is smaller. The decreasing speed is $O(1/n)$, and this result is compatible with the theoretical result. Greater ϵ also gives smaller covariance. In this case, the QMLE is concentrated in one point, and, as n increases, the distribution becomes more concentrated at that point.

3.6 Conclusion

We developed the simple protocols for building QMLEs from distributed data while guaranteeing ϵ-LDP for the users. They address the two different privacy scenarios. In the protocols, users submit only one or a few bits to the curator and do not need to wait for one another. Moreover, the users do not need to perform complex computations such as integration or derivation. Thus, the protocols

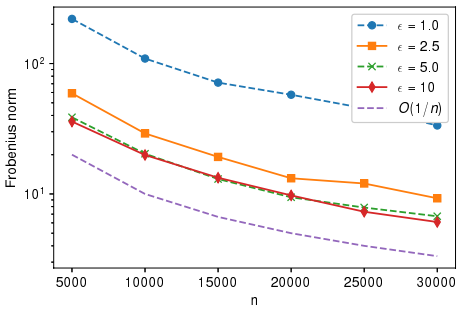


Figure 3.1: Frobenius norm of covariance matrices. The norms decrease in proportion to $1/n$ for each ϵ.

Algorithm 1: Protocol with Public $\mathbf{X}$

Input: Unknown distribution F, privacy parameter ϵ and $\mathcal{B}$
Curator set F_X;
for $i = 1$ *to* n **do**
  Curator generates $\mathbf{x}_i \sim F_X$;
  Send $\mathbf{x}_i$ to user i;
  User i generates $y_i \sim F(\cdot|\mathbf{x}_i)$;
  User i computes $\bar{y}_i$ as (3.6);
  User i generates $z_i \sim Q_{\text{bf}}(\cdot|\bar{y}_i)$;
  Send z_i to curator;
end
Let $D_n = \{(\mathbf{x}_i, z_i)\}_{i=1}^n$;
Curator computes $\ell(\beta; D)$;
Computes $\hat{\beta}_n = \arg\max_{\beta \in \mathcal{B}} \ell(\beta; D)$;
Output: $\hat{\beta}_n$

are highly user-friendly and suitable for low-priced devices. We clarified the sufficient conditions for the QMLEs to be consistent and asymptotically normal, and showed their limitations. We showed that the sufficient conditions are relaxed with a concrete implementation.

3.7 Pseudo-code

Algorithm 1 and Algorithm 2 are the pseudo-codes of the protocols described in Section 3.3.1 and Section 3.3.2, respectively. In the for loops, the processing of each user does not need to be synchronized.

Algorithm 2: Protocol with Private $\mathbf{X}$

Input: Unknown distribution F, F_X, privacy parameter ϵ and $\mathcal{B}$
Curator set $\hat{F}_X$;
for $i = 1$ *to* n **do**
  User i generates $\mathbf{x}_i \sim F_X$;
  User i generates $y_i \sim F(\cdot|\mathbf{x}_i)$;
  User i computes $\bar{y}_i$ as (3.6);
  User i generates $z_i^{(Y)} \sim Q_{\text{bf}}(\cdot|\bar{y}_i)$;
  for $j = 1$ *to* d **do**
    User i compute $\bar{x}_{ij}$ as (3.6);
    Generate $z_{ij}^{(\mathbf{X})} \sim Q_{\text{bf}}(\cdot|\bar{x}_{ij})$;
  end
  Let $\mathbf{z}_i^{(\mathbf{X})} = (z_{ij}^{(\mathbf{X})})_i$;
  Send $(\mathbf{z}_i^{(\mathbf{X})}, z_i^{(Y)})$ to curator;
end
Let $D_n = \{(\mathbf{z}_i^{(\mathbf{X})}, z_i^{(Y)})\}_{i=1}^n$;
Curator computes $\ell(\beta; D)$;
Computes $\hat{\beta}_n = \arg\max_{\beta \in \mathcal{B}} \ell(\beta; D)$;
Output: $\hat{\beta}_n$

3.8 Mathematical Notes

3.8.1 for Section 3.3.1

Proof of Lemma 3

By the definition of $\Lambda_\epsilon(\beta, \mathbf{x})$, it is written as

$$\Lambda_\epsilon(\beta, \mathbf{x}) = p_\beta(Z = 1|\mathbf{X} = \mathbf{x}) = \int Q_{\text{bf}}(1|t(y))dF_\beta(y|\mathbf{x}).$$

From the definition of Q_{bf}, we have

$$\frac{1}{e^\epsilon + 1} \leq Q_{\text{bf}}(1|t(y)) \leq \frac{e^\epsilon}{e^\epsilon + 1}$$

for any $y \in \mathcal{Y}$. Thus, the following relation holds.

$$\Lambda_\epsilon(\beta, \mathbf{x}) \leq \int \frac{e^\epsilon}{e^\epsilon + 1} dF_\beta(y|\mathbf{x}) = \frac{e^\epsilon}{e^\epsilon + 1}.$$

The last equation is by the fact that F_β is a probability distribution. Similarly, we have

$$\Lambda_\epsilon(\beta, \mathbf{x}) \geq \frac{1}{e^\epsilon + 1}.$$

3.8.2 for Section 3.3.2

For each $\mathbf{z}^{(\mathbf{X})} \in \{-1,+1\}^d$, the curator considers the probability distribution of $\mathbf{Z}^{(\mathbf{X})}$ at $\mathbf{z}^{(\mathbf{X})}$ as

$$\hat{p}_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})}) = \int Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})}|\mathbf{x}) d\hat{F}(\mathbf{x}).$$

The joint density is

$$p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)}) = \int Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})}|\mathbf{x}) Q_{Z^{(Y)}}(z^{(Y)}|t(y)) dF_\beta(y|\mathbf{x}) d\hat{F}_X(\mathbf{x}).$$

The conditional distribution of $Z^{(Y)}$ is written as

$$\begin{aligned} p_\beta(\mathbf{z}^{(Y)}|\mathbf{z}^{(\mathbf{X})}) =& \frac{\hat{F}_X(p_\beta(z^{(Y)}|\mathbf{X}) Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})}|\mathbf{X}))}{\hat{p}_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})})} \\ =& \Phi(\beta, \mathbf{z}^{(\mathbf{X})})^{z^{(Y)}} (1 - \Phi(\beta, \mathbf{z}^{(\mathbf{X})}))^{1-z^{(Y)}}. \end{aligned}$$

With Φ, the joint density is written as

$$p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)}) = \Phi(\beta, \mathbf{z}^{(\mathbf{X})})^{z^{(Y)}} (1 - \Phi(\beta, \mathbf{z}^{(\mathbf{X})}))^{1-z^{(Y)}} \hat{p}_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})}).$$

We analyze the sufficient conditions under which Assumptions 2, 3 and 7 are satisfied while replacing F and F_θ in Theorem 3 with P and P_β. We adopt Assumptions 13 to 15 and 22. From these assumptions, it is obvious that $P(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})$ and $P_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})$ are measurable, and that density functions $p(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})$ and $p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})$ exist.

The condition corresponding to Assumption 9 consists of two parts. The first part is the existence of $P \log p$ integrable function $h(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})$ such that $|\log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})| \leq h(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})$ for all β. $P \log p$ is expanded as follows:

$$P \log p = P_{\mathbf{Z}}^{(\mathbf{X})}(P_{\cdot|\mathbf{Z}^{(\mathbf{X})}} \log p(\cdot|\mathbf{Z}^{(\mathbf{X})}) + \log p_{\mathbf{Z}}^{(\mathbf{X})}).$$

To evaluate the bound condition, it is necessary to analyze $p_{\mathbf{Z}^{(\mathbf{X})}}$ and $p(\cdot|\mathbf{Z}^{(\mathbf{X})})$.

Lemma 5. *For any* $\mathbf{z}^{(\mathbf{X})} \in \{-1,+1\}^d$,

$$\left(\frac{1}{e^{\epsilon/(d+1)}+1}\right)^d \leq \hat{p}_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})}) \leq \left(\frac{e^{\epsilon/(d+1)}}{e^{\epsilon/(d+1)}+1}\right)^d.$$

$p_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})})$ *has the same bounds.*

Lemma 6. *For any* $\beta \in \mathcal{B}$ *and* $\mathbf{z}^{(\mathbf{X})} \in \{-1,+1\}^d$,

$$\frac{1}{e^{\epsilon/(d+1)}+1} \leq \Phi(\beta, \mathbf{z}^{(\mathbf{X})}) \leq \frac{e^{\epsilon/(d+1)}}{e^{\epsilon/(d+1)}+1}.$$

With $0 < \epsilon < +\infty$ *and* $1 \leq d < +\infty$, $\Phi(\beta, \mathbf{z}^{(\mathbf{X})})$ *are always bounded away from* 0 *and* 1. *Also,*

$$\frac{1}{e^{\epsilon/(d+1)}+1} \leq 1 - \Phi(\beta, \mathbf{z}^{(\mathbf{X})}) \leq \frac{e^{\epsilon/(d+1)}}{e^{\epsilon/(d+1)}+1}.$$

By the above lemmas, $\log p(\cdot|\mathbf{z}^{(\mathbf{X})})$ and $\log p_{\mathbf{Z}}^{(\mathbf{X})}$ are always bounded away from $\pm\infty$, and $\log p(\cdot|\mathbf{z}^{(\mathbf{X})})$ and $\log p_{\mathbf{Z}}^{(\mathbf{X})}$ are always integrable with respect to P. The existence of integral function h is also obtained.

The second part is the uniqueness of the log-likelihood function. To guarantee that this property holds, we again adopt Assumption 17. Then, we have Theorem 7.

We next analyze the conditions under which Assumptions 10 to 12 are satisfied. The condition corresponding to Assumption 10 is the continuous differentiability of $\partial \log p_\beta / \partial\beta$. The partial derivative is expanded as

$$\begin{aligned}\frac{\partial}{\partial\beta}\log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)}) =& z^{(Y)}\frac{\Phi'(\beta,\mathbf{z}^{(\mathbf{X})})}{\Phi(\beta,\mathbf{z}^{(\mathbf{X})})} - (1-z^{(Y)})\frac{\Phi'(\beta,\mathbf{z}^{(\mathbf{X})})}{1-\Phi(\beta,\mathbf{z}^{(\mathbf{X})})}\\ =& \frac{\Phi'(\beta,\mathbf{z}^{(\mathbf{X})})(z^{(Y)}-\Phi(\beta,\mathbf{z}^{(\mathbf{X})}))}{\Phi(\beta,\mathbf{z}^{(\mathbf{X})})(1-\Phi(\beta,\mathbf{z}^{(\mathbf{X})}))}\end{aligned}$$

where

$$\Phi'(\beta,\mathbf{z}^{(\mathbf{X})}) \equiv \frac{\partial}{\partial\beta}\Phi(\beta,\mathbf{z}^{(\mathbf{X})}).$$

By Lemma 6, $\Phi(\beta,\mathbf{z}^{(\mathbf{X})})$ always takes values greater than 0 and less than 1. So, if Assumption 23 holds, Assumption 10 is satisfied.

The condition corresponding to Assumption 11 is that there exist integrable functions with respect to P that upper bound the absolute values of each component of $\partial^2 \log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})/\partial\beta^2$ and $(\partial \log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})/\partial\beta)(\partial \log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})/\partial\beta)^\top$. The second-order derivative is

$$\begin{aligned}\frac{\partial^2}{\partial\beta^2}\log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)}) =& (2z^{(Y)}-1)\frac{\Phi''(\beta,\mathbf{z}^{(\mathbf{X})})}{\Psi_\epsilon(\beta^\top x)^z(1-\Psi_\epsilon(\beta^\top x))^{1-z^{(Y)}}}\\ &- \frac{\Phi'(\beta,\mathbf{z}^{(\mathbf{X})}))(\Phi'(\beta,\mathbf{z}^{(\mathbf{X})}))^\top}{\Phi(\beta,\mathbf{z}^{(\mathbf{X})})^{2z^{(Y)}}(1-\Phi(\beta,\mathbf{z}^{(\mathbf{X})}))^{2(1-z^{(Y)})}}\end{aligned}$$

where we define $\Phi''(\beta,\mathbf{z}^{(\mathbf{X})}) \equiv \frac{\partial^2}{\partial\beta^2}\Phi(\beta,\mathbf{z}^{(\mathbf{X})})$.
$(\partial \log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})/\partial\beta)(\partial \log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})/\partial\beta)^\top$ is

$$\begin{aligned}&\left(\frac{\partial}{\partial\beta}\log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})\right)\left(\frac{\partial}{\partial\beta}\log p_\beta(\mathbf{z}^{(\mathbf{X})}, z^{(Y)})\right)^\top\\ &= \left(\frac{z-\Phi(\beta,\mathbf{z}^{(\mathbf{X})})}{\Phi(\beta,\mathbf{z}^{(\mathbf{X})})(1-\Phi(\beta,\mathbf{z}^{(\mathbf{X})}))}\right)^2 \Phi'(\beta,\mathbf{z}^{(\mathbf{X})})\Phi'(\beta,\mathbf{z}^{(\mathbf{X})})^\top\end{aligned}$$

By Lemma 6, Assumption 24 is sufficient to make the requirement hold.

The requirement corresponding Assumption 12 consists of three parts. The first part is that β is interior of $\mathcal{B}$. We assume this as Assumption 20. For enough large $\mathcal{B}$, this assumption is not particularly strong. Letting

$$A(\beta) = P\frac{\partial^2}{\partial\beta^2}\log p_\beta \quad \text{and} \quad B(\beta) = P\left(\frac{\partial}{\partial\beta}\log p_\beta\right)\left(\frac{\partial}{\partial\beta}\log p_\beta\right)^\top,$$

the second and third parts are the regularity of $A(\beta^*)$ and $B(\beta^*)$. We have already assumed that $A(\beta^*)$ is regular in Assumption 17. We consider the

regularity of $B(\beta^*)$ here. $B(\beta)$ is

$$B(\beta) = \mathbb{E}_{\mathbf{Z}^{(\mathbf{X})}}\left[\left(\frac{p(1|\mathbf{Z}^{(\mathbf{X})})}{\Phi(\beta, \mathbf{Z}^{(\mathbf{X})})^2} + \frac{p(0|\mathbf{Z}^{(\mathbf{X})})}{(1-\Phi(\beta, \mathbf{Z}^{(\mathbf{X})}))^2}\right)(\Phi'(\beta, \mathbf{Z}^{(\mathbf{X})})(\Phi'(\beta, \mathbf{Z}^{(\mathbf{X})}))^\top\right]$$

Since the scalar part is always finite and positive, Assumption 25 is a sufficient condition of the regularity of $B(\beta^*)$.

Summarizing the above discussions, we obtain Theorem 8.

Proof of Lemma 5

Proof. By definition, for any $\mathbf{z}^{(\mathbf{X})} \in \{-1, +1\}^d$, we have

$$\begin{aligned}\hat{p}_X(\mathbf{z}^{(\mathbf{X})}) = \int Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})}|\mathbf{x}) d\hat{F}(\mathbf{x}) &\leq \int \left(\frac{e^{\epsilon/(d+1)}}{e^{\epsilon/(d+1)}+1}\right)^d d\hat{F}(\mathbf{x}) \\ &= \left(\frac{e^{\epsilon/(d+1)}}{e^{\epsilon/(d+1)}+1}\right)^d.\end{aligned}$$

Similarly, we have

$$\hat{p}_X(\mathbf{z}^{(\mathbf{X})}) \geq \left(\frac{1}{e^{\epsilon/(d+1)}+1}\right)^d.$$

□

Proof of Lemma 6

Proof. By Lemma 3 and (3.8), we have

$$\begin{aligned}\Phi(\beta, \mathbf{z}^{(\mathbf{X})}) = \frac{\hat{F}_X \Lambda(\beta, \mathbf{X}) Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})}|\mathbf{X})}{p_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})})} &\leq \frac{\hat{F}_X \frac{e^{\epsilon/(d+1)}}{e^{\epsilon/(d+1)}+1} Q_{\mathbf{Z}^{(\mathbf{X})}}(\mathbf{z}^{(\mathbf{X})}|\mathbf{X})}{p_{\mathbf{Z}}^{(\mathbf{X})}(\mathbf{z}^{(\mathbf{X})})} \\ &= \frac{e^{\epsilon/(d+1)}}{e^{\epsilon/(d+1)}+1}.\end{aligned}$$

Similarly, we have

$$\Phi(\beta, \mathbf{z}^{(\mathbf{X})}) \geq \frac{1}{e^{\epsilon/(d+1)}+1}.$$

Replacing $\hat{F}_X$ by F_X, we obtain the arguments with regard to $p(z^{(Y)}|\mathbf{z}^{(\mathbf{X})})$. □

3.8.3 for Section 3.4

In this section, we derive $\Psi_\epsilon(\theta)$ used in Section 3.4. As a consequence of the analysis, we obtain Lemma 4. We analyze the function in different three cases. The first case is the case where $c_\mathrm{l} < \theta < c_\mathrm{u}$. For the sake of simplicity of notation, we let $G = \exp\left(-\frac{\alpha-1}{\sigma}(c_\mathrm{l} - \theta)\right)$ and $H = \exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u} - \theta)\right)$. These values appear many times throughout the remainder of this section. First, we

extend the probability $F_\theta(Y_i \leq c_\mathrm{l})$.

$$\begin{aligned}
F_\theta(Y_i \leq c_\mathrm{l}) &= \int_{-\infty}^{c_\mathrm{l}} \frac{\alpha(1-\alpha)}{\sigma} \exp\left(-\rho\left(\frac{y_i-\theta}{\sigma}\right)\right) dy_i \\
&= \int_{-\infty}^{c_\mathrm{l}} \frac{\alpha(1-\alpha)}{\sigma} \exp\left(-\frac{\alpha-1}{\sigma}(y_i-\theta)\right) dy_i \\
&= \left[\frac{\alpha(1-\alpha)}{\sigma}\left(-\frac{\sigma}{\alpha-1}\right)\exp\left(-\frac{\alpha-1}{\sigma}(y_i-\theta)\right)\right]_{-\infty}^{c_\mathrm{l}} \\
&= \alpha \exp\left(-\frac{\alpha-1}{\sigma}(c_\mathrm{l}-\theta)\right) - \alpha \times 0 \\
&= \alpha \exp\left(-\frac{\alpha-1}{\sigma}(c_\mathrm{l}-\theta)\right) = \alpha G.
\end{aligned}$$

Similarly, the probability $F_\theta(Y_i \geq c_\mathrm{u})$ is expanded as:

$$\begin{aligned}
F_\theta(Y_i \geq c_\mathrm{u}) &= \int_{c_\mathrm{u}}^{+\infty} \frac{\alpha(1-\alpha)}{\sigma} \exp\left(-\rho\left(\frac{y_i-\theta}{\sigma}\right)\right) dy_i \\
&= \int_{c_\mathrm{u}}^{+\infty} \frac{\alpha(1-\alpha)}{\sigma} \exp\left(-\frac{\alpha}{\sigma}(y_i-\theta)\right) dy_i \\
&= \left[\frac{\alpha(1-\alpha)}{\sigma}\left(-\frac{\alpha}{\sigma}\right)\exp\left(-\frac{\alpha}{\sigma}(y_i-\theta)\right)\right]_{c_\mathrm{u}}^{+\infty} \\
&= -(1-\alpha) \times 0 + (1-\alpha)\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right) \\
&= (1-\alpha)\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right) = (1-\alpha)H.
\end{aligned}$$

The probability $P_\theta(Z_i = 1)$ is written as follows:

$$\begin{aligned}
P_\theta(Z_i = 1) =& \alpha G\left(\frac{1}{2} - \frac{1}{2C_\epsilon}\right) \\
&+ \frac{\alpha(1-\alpha)}{\sigma}\int_{c_\mathrm{l}}^{\theta}\left(\frac{1}{2} + \frac{y_i - \frac{c_\mathrm{u}+c_\mathrm{l}}{2}}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\exp\left(-(\alpha-1)\frac{y_i-\theta}{\sigma}\right) dy_i \\
&+ \frac{\alpha(1-\alpha)}{\sigma}\int_{\theta}^{c_\mathrm{u}}\left(\frac{1}{2} + \frac{y_i - \frac{c_\mathrm{u}+c_\mathrm{l}}{2}}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\exp\left(-\alpha\frac{y_i-\theta}{\sigma}\right) dy_i \\
&+ (1-\alpha)H\left(\frac{1}{2} + \frac{1}{2C_\epsilon}\right)
\end{aligned} \tag{3.9}$$

Now, we extend each term.

$$
\begin{aligned}
&\frac{\alpha(1-\alpha)}{\sigma}\int_{c_\mathrm{l}}^{\theta}\left(\frac{1}{2}+\frac{y_i-\frac{c_\mathrm{u}+c_\mathrm{l}}{2}}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\exp\left(-(\alpha-1)\frac{y_i-\theta}{\sigma}\right)dy_i\\
=&\frac{\alpha(1-\alpha)}{\sigma}\left(\frac{1}{2}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}+\frac{\theta}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\int_{c_\mathrm{l}}^{\theta}\exp\left(-(\alpha-1)\frac{y_i-\theta}{\sigma}\right)dy_i\\
&+\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\int_{c_\mathrm{l}}^{\theta}(y_i-\theta)\exp\left(-(\alpha-1)\frac{y_i-\theta}{\sigma}\right)dy_i\\
=&\frac{\alpha(1-\alpha)}{\sigma}\left(\frac{1}{2}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}+\frac{\theta}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\frac{\sigma}{1-\alpha}(1-G)\\
&+\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(-\frac{\sigma^2}{(1-\alpha)^2}-\frac{\sigma}{1-\alpha}(c_\mathrm{l}-\theta)G+\frac{\sigma^2}{(1-\alpha)^2}G\right)\\
=&\alpha\left(\frac{1}{2}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}+\frac{\theta}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)(1-G)\\
&+\frac{\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(-\frac{\sigma}{1-\alpha}-(c_\mathrm{l}-\theta)G+\frac{\sigma}{1-\alpha}G\right).
\end{aligned}
\tag{3.10}
$$

Similarly,

$$
\begin{aligned}
&\frac{\alpha(1-\alpha)}{\sigma}\int_{\theta}^{c_\mathrm{u}}\left(\frac{1}{2}+\frac{y_i-\frac{c_\mathrm{u}+c_\mathrm{l}}{2}}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\exp\left(-\alpha\frac{y_i-\theta}{\sigma}\right)dy_i\\
=&-(1-\alpha)\left(\frac{1}{2}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}+\frac{\theta}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)(H-1)\\
&+\frac{\alpha(1-\alpha)}{\sigma C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(-\frac{\sigma}{\alpha}(c_\mathrm{u}-\theta)H-\frac{\sigma^2}{\alpha^2}H+\frac{\sigma^2}{\alpha^2}\right)\\
=&-(1-\alpha)\left(\frac{1}{2}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}+\frac{\theta}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)(H-1)\\
&+\frac{1-\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(-(c_\mathrm{u}-\theta)H-\frac{\sigma}{\alpha}H+\frac{\sigma}{\alpha}\right).
\end{aligned}
\tag{3.11}
$$

Substituting (3.10) and (3.11) into (3.9), we have

$$
\begin{aligned}
\Psi_\epsilon(\theta)=&P_\theta(Z_i=1)\\
=&\frac{\theta}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}+\frac{\alpha}{1-\alpha}\frac{\sigma}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(-\frac{\alpha-1}{\sigma}(c_\mathrm{l}-\theta)\right)\\
&-\frac{1-\alpha}{\alpha}\frac{\sigma}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\\
&+\frac{1}{2}+\left(-\frac{\alpha}{1-\alpha}+\frac{1-\alpha}{\alpha}\right)\frac{\sigma}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}.
\end{aligned}
$$

The first and second derivatives are

$$\begin{aligned}\Psi'_\epsilon(\theta) =& \frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})} - \frac{\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\\ & - \frac{1-\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right),\\ \Psi''_\epsilon(\theta) =& \frac{\alpha(1-\alpha)}{\sigma C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(e^{\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)} - e^{-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)}\right).\end{aligned}$$

By $c_\mathrm{l} < \theta < c_\mathrm{u}$, $\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\beta^\top\mathbf{x})$ and $-\frac{\alpha}{\sigma}(c_\mathrm{u}-\beta^\top\mathbf{x})$ are always negative.

$$|\Psi'_\epsilon(\theta)| < \frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})} \quad \text{and} \quad |\Psi''_\epsilon(\theta)| < \frac{\alpha(1-\alpha)}{\sigma C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}.$$

The second case is the case where $\theta \le c_\mathrm{l}$. $\Psi_\epsilon(\theta)$ is computed as

$$\begin{aligned}&\Psi_\epsilon(\theta)\\ =&\left(-(1-\alpha)\exp\left(-\alpha\frac{c_\mathrm{l}-\theta}{\sigma}\right)+1\right)\left(\frac{1}{2}-\frac{1}{2C_\epsilon}\right)\\ &+\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\frac{\sigma}{\alpha}(c_\mathrm{l}-\theta)+\frac{\sigma^2}{\alpha^2}\right)\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\\ &-\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\frac{\sigma}{\alpha}(c_\mathrm{u}-\theta)+\frac{\sigma^2}{\alpha^2}\right)\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\\ &+\frac{\alpha(1-\alpha)}{\sigma}\left(\frac{1}{2}+\frac{\theta}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\frac{\sigma}{\alpha}(\exp(-\frac{\alpha}{\sigma}(c_\mathrm{l}-\theta))\\ &-\exp(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)))\\ &+(1-\alpha)\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\left(\frac{1}{2}+\frac{1}{2C_\epsilon}\right)\\ =&\left(-(1-\alpha)\exp\left(-\alpha\frac{c_\mathrm{l}-\theta}{\sigma}\right)+1\right)\left(\frac{1}{2}-\frac{1}{2C_\epsilon}\right)\\ &+\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\frac{\sigma}{\alpha}c_\mathrm{l}+\frac{\sigma^2}{\alpha^2}\right)\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\\ &-\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\frac{\sigma}{\alpha}c_\mathrm{u}+\frac{\sigma^2}{\alpha^2}\right)\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\\ &+\frac{\alpha(1-\alpha)}{\sigma}\left(\frac{1}{2}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\frac{\sigma}{\alpha}(\exp(-\frac{\alpha}{\sigma}(c_\mathrm{l}-\theta))\\ &-\exp(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)))\\ &+(1-\alpha)\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\left(\frac{1}{2}+\frac{1}{2C_\epsilon}\right)\\ =&\frac{1}{2}-\frac{1}{2C_\epsilon}+\frac{(1-\alpha)\sigma}{\alpha}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\\ &-\frac{(1-\alpha)\sigma}{\alpha}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right).\end{aligned}$$

Its first and second derivatives are

$$\Psi'_\epsilon(\theta) = \frac{1-\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{l}-\theta)\right) - \exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\right),$$
$$\Psi''_\epsilon(\theta) = \frac{\alpha(1-\alpha)}{\sigma C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{l}-\theta)\right) - \exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\right)$$

Since $\theta \le c_\mathrm{l}$ and $c_\mathrm{u} > c_\mathrm{l}$, $\Psi'_\epsilon(\theta)$ is positive, and $\Psi''_\epsilon(\theta)$ is positive. Moreover, we have

$$|\Psi'_\epsilon(\theta)| < \frac{1-\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})} \quad \text{and} \quad |\Psi''_\epsilon(\theta)| < \frac{\alpha(1-\alpha)}{\sigma C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}.$$

The last case is the case where $\theta \ge c_\mathrm{u}$.

$$\begin{aligned}
&\Psi_\epsilon(\theta)\\
=&\alpha\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\left(\frac{1}{2}-\frac{1}{2C_\epsilon}\right)\\
&-\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\frac{\sigma}{1-\alpha}(c_\mathrm{l}-\theta)-\frac{\sigma^2}{(1-\alpha)^2}\right)\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\\
&+\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\frac{\sigma}{1-\alpha}(c_\mathrm{u}-\theta)-\frac{\sigma^2}{(1-\alpha)^2}\right)\exp\left(-\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\\
&+\frac{\alpha(1-\alpha)}{\sigma}\left(\frac{1}{2}+\frac{\theta}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\frac{\sigma}{1-\alpha}(-\exp(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta))\\
&+\exp(\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)))\\
&+\left(1-\alpha\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\right)\left(\frac{1}{2}+\frac{1}{2C_\epsilon}\right)\\
=&\alpha\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\left(\frac{1}{2}-\frac{1}{2C_\epsilon}\right)\\
&-\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\frac{\sigma}{1-\alpha}c_\mathrm{l}-\frac{\sigma^2}{(1-\alpha)^2}\right)\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\\
&+\frac{\alpha(1-\alpha)}{\sigma}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\frac{\sigma}{1-\alpha}c_\mathrm{u}-\frac{\sigma^2}{(1-\alpha)^2}\right)\exp\left(-\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\\
&+\frac{\alpha(1-\alpha)}{\sigma}\left(\frac{1}{2}-\frac{c_\mathrm{u}+c_\mathrm{l}}{2C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\right)\frac{\sigma}{1-\alpha}(-\exp(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta))\\
&+\exp(\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)))\\
&+\left(1-\alpha\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\right)\left(\frac{1}{2}+\frac{1}{2C_\epsilon}\right)\\
=&\frac{1}{2}+\frac{1}{2C_\epsilon}+\frac{\alpha\sigma}{1-\alpha}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)\\
&-\frac{\alpha\sigma}{1-\alpha}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(-\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)\right).
\end{aligned}$$

Its first and second derivatives are

$$\Psi'_\epsilon(\theta) = \frac{\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(-\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)+\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\right),$$
$$\Psi''_\epsilon(\theta) = \frac{\alpha(1-\alpha)}{\sigma C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-\theta)\right)-\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{u}-\theta)\right)\right).$$

Since $\theta \geq c_\mathrm{u}$ and $c_\mathrm{u} > c_\mathrm{l}$, $\Psi'_\epsilon(\theta)$ is positive, and $\Psi''_\epsilon(\theta)$ is negative. Moreover, since $(1-\alpha)(c_\mathrm{l}-\theta)/\sigma < (1-\alpha)(c_\mathrm{u}-\theta)/\sigma \leq 0$, we have

$$|\Psi'_\epsilon(\theta)| < \frac{\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})} \quad \text{and} \quad |\Psi''_\epsilon(\theta)| < \frac{\alpha(1-\alpha)}{\sigma C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}.$$

We also analyze their behavior on the boundaries. $\Psi_\epsilon(\theta)$ is continuous at $\theta = c_\mathrm{l}$ and c_u if and only if $\lim_{\theta\downarrow c_\mathrm{u}}\Psi_\epsilon(\theta) = \lim_{\theta\uparrow c_\mathrm{u}}\Psi_\epsilon(\theta)$ and $\lim_{\theta\downarrow c_\mathrm{l}}\Psi_\epsilon(\theta) = \lim_{\theta\uparrow c_\mathrm{l}}\Psi_\epsilon(\theta)$. As we see below, these equations hold.

$$\begin{aligned}
&\lim_{\theta\downarrow c_\mathrm{u}}\Psi_\epsilon(\theta) = \lim_{\theta\uparrow c_\mathrm{u}}\Psi_\epsilon(\theta)\\
=&\frac{1}{2}+\frac{1}{2C_\epsilon}+\frac{\alpha\sigma}{1-\alpha}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-c_\mathrm{u})\right)-\frac{\alpha\sigma}{1-\alpha}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})},\\
&\lim_{\theta\downarrow c_\mathrm{l}}\Psi_\epsilon(\theta) = \lim_{\theta\uparrow c_\mathrm{l}}\Psi_\epsilon(\theta)\\
=&\frac{1}{2}-\frac{1}{2C_\epsilon}+\frac{(1-\alpha)\sigma}{\alpha}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}-\frac{(1-\alpha)\sigma}{\alpha}\frac{1}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-c_\mathrm{l})\right).
\end{aligned}$$

We next evaluate the existence of first and second derivatives at $\theta = c_\mathrm{l}$ and c_u.

$$\lim_{\theta\downarrow c_\mathrm{u}}\Psi'_\epsilon(\theta) = \lim_{\theta\uparrow c_\mathrm{u}}\Psi'_\epsilon(\theta) = \frac{\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(-\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-c_\mathrm{u})\right)+1\right),$$
$$\lim_{\theta\downarrow c_\mathrm{l}}\Psi'_\epsilon(\theta) = \lim_{\theta\uparrow c_\mathrm{l}}\Psi'_\epsilon(\theta) = \frac{1-\alpha}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(1-\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-c_\mathrm{l})\right)\right).$$

$$\lim_{\theta\downarrow c_\mathrm{u}}\Psi''_\epsilon(\theta) = \lim_{\theta\uparrow c_\mathrm{u}}\Psi''_\epsilon(\theta) = \frac{\alpha(1-\alpha)}{C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(\exp\left(\frac{1-\alpha}{\sigma}(c_\mathrm{l}-c_\mathrm{u})\right)-1\right),$$
$$\lim_{\theta\downarrow c_\mathrm{l}}\Psi''_\epsilon(\theta) = \lim_{\theta\uparrow c_\mathrm{l}}\Psi''_\epsilon(\theta) = \frac{\alpha(1-\alpha)}{\sigma C_\epsilon(c_\mathrm{u}-c_\mathrm{l})}\left(1-\exp\left(-\frac{\alpha}{\sigma}(c_\mathrm{u}-c_\mathrm{l})\right)\right).$$

3.9 Comparison with Non-private Estimator

For comparison with existing work, we also consider the correct model case.

Assumption 26. *Given $\mathbf{x} \in \mathbb{R}^d$, Y is a random variable sampled from the asymmetric Laplace distribution $f(\cdot;\alpha,\beta^\top\mathbf{x},\sigma)$, which is defined in (3.2). For each $i \in [n]$, y_i is a realization of random variable Y_i that is a copy of Y.*

Under this condition, Corollary 1 is more specified.

Corollary 3. *Suppose Assumptions 13, 16, 17 and 26 hold. The MLE $\hat{\beta}_n$ is distributed asymptotically normally as $\sqrt{n}(\hat{\beta}_n-\beta^*) \to \mathcal{N}(0_d, I_{\beta^*}^{-1})$ where $I_{\beta^*} = F_X\frac{\Psi'_\epsilon(\beta^\top\mathbf{X})^2}{\Psi_\epsilon(\beta^\top\mathbf{X})(1-\Psi_\epsilon(\beta^\top\mathbf{X}))}\mathbf{X}\mathbf{X}^\top$.*

To obtain an intuitive understanding of the result, we roughly compare the Fisher information matrix derived in Corollary 3 and the non-private Fisher matrix (3.5), and analyze some extreme cases. First, we consider the concentrated case in which the scale parameter σ is extremely small. For a σ sufficiently small that $\sigma \ll |(1-\alpha)(c_{\mathrm{l}} - {\beta^*}^\top \mathbf{x})|$ and $\sigma \ll |\alpha(c_{\mathrm{u}} - {\beta^*}^\top \mathbf{x})|$ for most $\mathbf{x}$,

$$\Psi({\beta^*}^\top \mathbf{x}) \approx \frac{1}{2} + \left(-\frac{\alpha}{1-\alpha} + \frac{1-\alpha}{\alpha}\right)\frac{\sigma}{2C_\epsilon} + \frac{{\beta^*}^\top \mathbf{x}}{2C_\epsilon} \quad \text{and} \quad \Psi'({\beta^*}^\top \mathbf{x}) \approx \frac{1}{2C_\epsilon}.$$

Thus,

$$\frac{\Psi'({\beta^*}^\top \mathbf{x})^2}{\Psi({\beta^*}^\top \mathbf{x})(1-\Psi({\beta^*}^\top \mathbf{x}))} \approx \frac{1}{C_\epsilon^2 - \left(\frac{1-2\alpha}{\alpha(1-\alpha)}\sigma + {\beta^*}^\top \mathbf{x}\right)^2} \geq \frac{1}{C_\epsilon^2 - \left(\frac{1-2\alpha}{\alpha(1-\alpha)}\sigma + c_{\mathrm{l}}\right)^2}.$$

In comparing this with (3.4), we can see that the Fisher information matrix of our LDP estimator is $\Omega\left(\epsilon^2 \frac{\sigma^2}{\alpha(1-\alpha)}\right)$ times smaller than that of the non-private estimator as $\epsilon \downarrow 0$. This lower bound agrees with the complexity of ϵ but is $\sigma^2/\alpha(1-\alpha)$ times lower. Since we assumed that σ is small, this gap can be large. Although our MLE tends to lose more information regarding the structure of f_{β^*} than an optimal MLE, it experiences minimum information loss due to perturbation for privacy.

We omit the comparison of the MLE of the regression coefficient with the private $\mathbf{X}$. The Fisher information matrix strongly depends on the structure of the distribution of $\mathbf{X}$. We have no informative comparison in this case.

3.10 Additional Numerical Evaluation

In this section, we perform some additional numerical evaluations with the real data, which is the same data used in Section 3.5.

We implemented our simulation in Python and used the data in "https://archive.ics.uci.edu/ml/datasets/Gas+Turbine+CO+and+NOx+Emission+Data+Set".

3.10.1 Evaluation of Private X

Here, we observe the behavior of our QMLE for the private $\mathbf{X}$ scenario, which is described in Section 3.4.2.

Due to implementation needs, we have made some modifications to the description in the main part. First, we made some changes to $\Phi(\beta, \mathbf{z}^{(\mathbf{X})})$. Theoretically, Φ and $1-\Phi$ never take negative values. However, we found that the value of Φ can exceed 1 by a small amount due to rounding error. Then, $1-\Phi$ is negative, and the computation corrupts since the log function is inputted a negative value. To avoid this undesirable situation, we multiplied Φ by $e^{-0.000001}$.

Second, we changed the domain of $\hat{F}_{\mathbf{X}}$ because no element of each $\mathbf{x}_i$ is in the interval $[-1, 1]$. In the simulation, each user truncates the components of $\mathbf{X}_i$ into the intervals $[5, 10]$, $[1000, 1030]$, $[70, 100]$, $[4, 6]$, $[20, 30]$, $[1000, 1100]$, $[530, 570]$, $[130, 170]$, and $[10, 15]$. We recommend that the curators should set the intervals with the help of experts when they use our algorithm in reality.

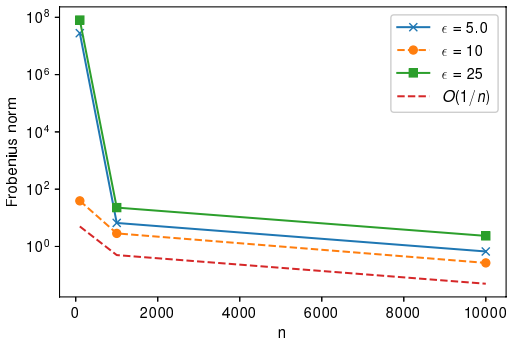


Figure 3.2: Frobenius norm of covariance matrices in private $\mathbf{X}$ scenario. The norm decrease in proportion to $1/n$ for each ϵ.

We observe the covariance matrices for $n \in \{100, 1000, 10000\}$ and $\epsilon \in \{5.0, 10, 25\}$ with $\alpha = 0.3$ and $\sigma = 1.0$. For each combination of n and ϵ, we sub-sample n records $1,000$ times without replacement from the $36,733$ records. For each sub-data, we simulate the protocol described in Section 3.4.2 and obtain a QMLE. Then, we compute the Frobenius norm of covariance matrices of the $1,000$ QMLEs,

Figure 3.2 shows the result. The horizontal and vertical axes show n and the value of each Frobenius norm in log-scale, respectively. For each ϵ, with large n, the norm of the covariance matrix is smaller. The decreasing speed is $O(1/n)$, These properties are similar to those in the public $\mathbf{X}$ scenario, which is described in Section 3.5.

3.10.2 Evaluation of Effect of Truncation

In this subsection, we evaluate the effect of the truncation in the public $\mathbf{X}$ scenario.

With $\epsilon = 2.5$ and $n = 10,000$, we try intervals $[50, 100], [40, 110], [30, 120]$ and $[20, 130]$ for the truncation. The other setting is the same as Section 3.5.

Figure 3.3 shows the result. A shorter interval makes the estimators more concentrated. We remark that the concentration does not necessarily imply a good approximation of the true distribution. In general, there is a trade-off between bias and variance.

3.10.3 Comparison with Non-private Estimator

In this subsection, we evaluate the difference between the centers of the distributions of our QMLEs and the non-private QMLEs which is described in Section 3.2.3. Our theoretical result does not say that those QMLEs converge to the same point. Thus, we consider it with numerical simulations.

First, we observe the behavior of the non-private QMLE. Figure 3.4 shows the Frobenius norm of covariance matrices. It is seen that the non-private QMLEs converge to one point. We treat the average vector of the non-private QMLEs with $n = 30,000$ as the grand truth in the main observation as described below. We remark that the "grand truth" can be biased.

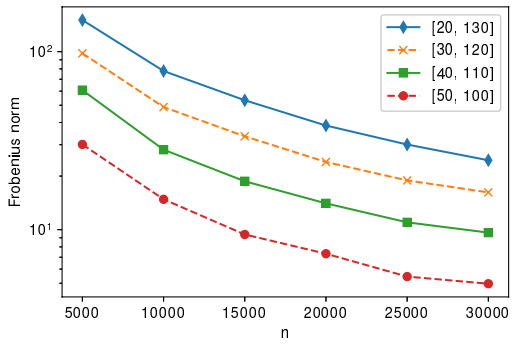


Figure 3.3: Frobenius norm of covariance matrices with various $[c_{\mathrm{l}}, c_{\mathrm{u}}]$s. A smaller interval makes the norm smaller.

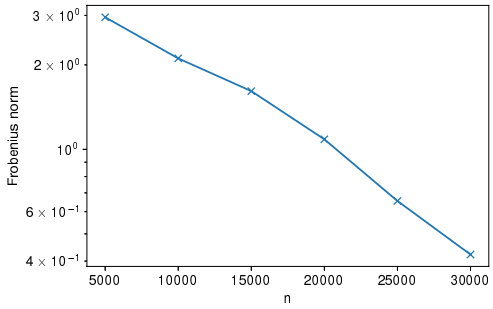


Figure 3.4: Frobenius norm of covariance matrices of non-private QMLE. It seems that the non-private QMLEs converge to one point.

We use the same simulation result used in Section 3.5. We compute the difference of the average vector of our QMLE and the grand truth and observe the norm for each n and ϵ.

Figure 3.5 shows the main result. The horizontal and vertical axes show n and the value of the norm of the covariance matrices, respectively. The bias is not zero for all ϵ. Smaller ϵ tends to give smaller bias. It is seen that n does not affect the bias. This result implies that the non-private QMLE and our QMLE can converge to different points.

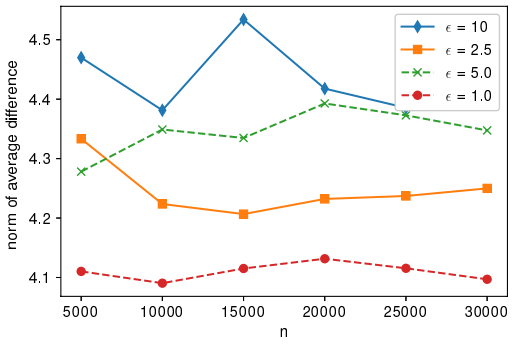


Figure 3.5: Norm of difference between the centers of non-private QMLEs and our QMLEs with various n and ϵ. The difference does not depend on n.

Chapter 4

Local Privacy in the Presence of Unexpected Values

4.1 Introduction

In Chapter 3, we use the technique of truncation to handle extremely large or small values. In this chapter, we discuss the issue of out-of-bounds values more generally.

The privacy guarantee associated with LDP can fail when a system faces an unexpected value such as an out-of-domain entry or a non-response since the definition of LDP is based on a privacy mechanism with a known domain. A system is deployed as a set of computer programs; when a program encounters an unexpected value, it behaves in an unexpected way, raises an exception or fails to produce any output. Observing such abnormal system behavior, the curator can conclude that the user supplied an unexpected value or did not input anything. Since a non-response can be correlated with a sensitive attribute, this is a clear privacy violation in the sense of LDP.

Although the curator can prevent some types of unexpected values from being input into an LDP mechanism by forcing users to provide input values through a particular user interface, it is extremely difficult to exclude all undesirable inputs. If, for example, the response form is a list of choices, the user would be unable to enter values that are not included in the list. However, although such a simple arrangement would seem to solve all the issues associated with unexpected values, it is not wholly satisfactory, as a user may not finish answering a question within the allotted time, or the user interface may raise an exception before passing an answer to the LDP mechanism, in which case the LDP mechanism receives an error report rather than the original values. Recent computer programs are complex, highly modular, and created by many programmers. No one programmer is able to predict every possible error in advance. Thus, we should assume that there always exists the possibility of an input that is unanticipated by the programmers of an LDP mechanism. This issue is also raised in the context of anonymization, which leads to a serious vul-

nerability in privacy protection [Ciglic et al., 2016]. In this chapter, we discard the assumption that users input only valid values to an LDP mechanism and consider a more-realistic system model.

We propose a new system model that involves unexpected inputs from users. Figure 4.1 shows a comparison of an ordinary LDP model and our proposed model. In our model, we introduce an agent that models a device such as a smartphone or PC used by each user. In terms of privacy protection, each user trusts her agent but does not trust the curator. The agent acts as an intermediary between each user and the curator. The most important task of the agent is to perturb the user input for privacy. We call the channel between a user and the agent a *pre-agent mechanism* and call the channel between the agent and the curator a *post-agent mechanism*. A pre-agent mechanism is a simplified model of the complex process involving user decision-making and a user interface for surveys; a post-agent mechanism is a model of the physical communication channel. The pre-agent mechanism maps user inputs to the expected domain or a special character $\perp$, whereas the post-agent mechanism maps an output of the LDP mechanism to the identical domain or $\perp$.[1] To fit the new system model, we modify the definition of LDP.

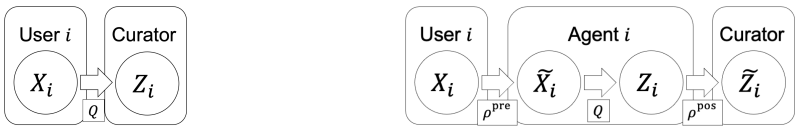
We show that a perturbation mechanism satisfying the standard LDP can violate privacy in our system model and then derive a sufficient condition for a perturbation mechanism to guarantee local privacy for users. We strongly recommend that curators attach to the perturbation mechanism an exception handler such that the curator cannot determine whether the system has encountered an exception.

The extended system model raises issues not only with regard to privacy analysis but also with respect to utility analysis. When we include unexpected values in the system model, the existing analyses [Duchi et al., 2013, Duchi et al., 2018, Duchi and Rogers, 2019a, Kairouz et al., 2014, Ye and Barg, 2018] for LDP estimation problems under the known-domain assumption are not applicable to our problem. Although there exist studies on algorithms for estimating standard statistics in the presence of missing values [Sun et al., 2018, Sun et al., 2020], we analyze the degree to which unexpected values decrease utility independent of any specific algorithm.

We provide a framework to analyze a lower bound for minimax risk (a popular measure of the difficulty of an estimation problem) of a locally private estimation problem in the presence of unexpected values. In the proposed framework, we separately analyze the three mechanisms. The framework requires us to evaluate the amount by which total variation or KL divergence between two distributions decreases for each mechanism. Since the decrease for an LDP mechanism has already been derived [Duchi et al., 2013], we additionally analyze the decrease attributable to pre- and post-agent mechanisms. In particular, we consider some concrete mechanisms and derive upper bounds of the decrease in total variation produced by these mechanisms.

We confirm that the lower bounds derived using our framework are achievable in two concrete examples. To show achievability, we design perturbation mechanisms with safe exception handlers and evaluate their risks. In the first example, we find that unexpected values do not necessarily harm the estima-

[1] Although Murakami and Kawamoto [2019] proposed a pre-processor mapping of high sensitive values to some semantic tags $\perp_1, \perp_2, \cdots$, their objective is to introduce intermediately sensitive data, which is different from ours.



(a) Ordinary system model. (b) Our system model.

Figure 4.1: Comparison of system models related to user i. The random variables $\widetilde{X}_i$ and $\widetilde{Z}_i$ can take a special value $\perp$ which represents exceptions. In (a), a record is perturbed only once. In (b), a record is perturbed three times.

tion. The degree of difficulty of an estimation problem is highly dependent on the definition of the evaluation measure. Differences in the evaluation measure alone can lead to either negligible or catastrophic errors due to unexpected values. In the second example, we find that a pre-agent mechanism is more critical than a post-agent mechanism even when they are essentially the same erasure mechanism, which replaces an inputted value by $\perp$ in probability. The results suggest that the curator should devote more effort to reducing the erasure rate of the pre-agent mechanism than that of the post-agent mechanism. In the case in which the pre- or post-agent mechanism uniformly erases its input at some rate γ, the minimax risk is proportional to $(1-\gamma)^{-2}$ or $(1-\gamma)^{-1}$, respectively.

The remainder of this chapter is organized as follows. In Section 4.2, we introduce the background knowledge necessary for a proper understanding of the chapter. In Section 4.3, we define our system model involving unexpected value for a curator and point out that a perturbation mechanism without an exception handler does not protect user privacy. In Section 4.4, we derive an abstract lower bound of risk of a locally private estimation problem with unexpected values and some information bounds with regard to some ρ^{pre} and ρ^{pos}. In Section 4.5, we consider two estimation problems and derive lower and upper bounds for the problems. In Section 4.6, we offer our conclusion.

4.2 Background

To clarify the distribution P of the random variable X, we denote the expectation of $f(X)$ as $P(f(X))$. We will refer to the j-th element of a vector v as $[v]_j$.

4.2.1 Local Differential Privacy

In this chapter, we conduct privacy analyses based on local differential privacy with a non-interactive model, which is the simplest of the local differential privacy models. Although the non-interactive model does not cover several important algorithms, it is sufficient for addressing the issue that we discuss in this chapter.

We begin by describing the system model. There exist n users and a single curator. Each user possesses a record x_i and submits it with perturbation by a mechanism Q to the curator. The Q stochastically maps record domain $\mathcal{X}$ to some set $\mathcal{Z}$. We denote the perturbed x_i by z_i. The local differential privacy is a property of each Q.

Definition 2 (Standard local differential privacy)**.** *Given $\epsilon > 0$, we say that a perturbation mechanism Q is ϵ-locally differentially private or ϵ-LDP if*

$$\forall x, x' \in \mathcal{X}, S \in \sigma(\mathcal{Z}), \frac{Q(S|x)}{Q(S|x')} \leq e^{\epsilon},$$

where $\sigma(\mathcal{Z})$ is a sigma algebra on $\mathcal{Z}$.

This definition says that if it is difficult to learn the input from output of Q, the perturbation mechanism Q protects privacy and that a smaller ϵ means a safer perturbation mechanism.

One of the most important properties of LDP mechanisms is the *post-processing invariance* that any operation for an output of an LDP mechanism cannot undermine protection.

Proposition 1 (Post-processing invariance [Dwork and Roth, 2014])**.** *For any deterministic or stochastic function f whose domain is $\mathcal{Z}$, if Q is ϵ-LDP, the composition $f \circ Q$ is ϵ-LDP.*

Since the proposition does not care about the codomain of f, even when the codomain includes $\perp$, the proposition holds. This plays an important role in our privacy analysis.

Minimax risk analyses for locally differentially private estimation problems have been studied. We introduce one of them. The minimax risk of a ϵ-LDP estimation problem is defined as follows:

$$R_n(\theta(P), w \circ d, \epsilon) \equiv \inf_{\hat{\theta}_n, Q} \sup_{\theta \in \Theta} Q(P_\theta(w(d(\hat{\theta}_n(Z_1, \cdots, Z_n), \theta))))$$

where $\inf_Q$ is taken over the set of all ϵ-LDP mechanisms. Duchi et al. modified Theorem 1 for local privacy estimations and obtained the following proposition.

Proposition 2 (Proposition 1 of [Duchi et al., 2018])**.** *Suppose we are given n i.i.d. observations from an ϵ-locally differential private channel for some $\epsilon \in [0, 23/35]$. Then for any pair of distributions (P_0, P_1) that is 2δ-separated with respect to θ, the ϵ-LDP minimax risk has a lower bound*

$$R_n(\theta(P), w \circ d, \epsilon) \geq \frac{w(\delta)}{2} \left(1 - \sqrt{4\epsilon^2 n \left\|P_0 - P_1\right\|_{TV}^2}\right). \tag{4.1}$$

Comparing the lower bound (4.1) with the classical lower bound (2.4), $\|P_0^n - P_1^n\|_{\mathrm{TV}}$ is replaced by $\sqrt{4\epsilon^2 n \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2}$. The main part of the proof of this proposition is the following inequality:

$$\|M_0^n - M_1^n\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}(M_1^n \| M_2^n) \leq 4\epsilon^2 n \left\|P_0 - P_1\right\|_{\mathrm{TV}}^2 \tag{4.2}$$

where M_0 and M_1 are the marginal distributions of each Z_i when $P = P_0$ and $P = P_1$, respectively. The left inequality is Pinsker's inequality Tsybakov [2008], and the right inequality is their contribution. Roughly speaking, this inequality represents how much smaller the total variation between M_0 and M_1 is than the total variation between P_0 and P_1.

4.3 Local Privacy with Unexpected Values

We redefine local differential privacy in the presence of unexpected values. We consider a different system model in this section and give some interpretations of the new system model below. There are n users with sensitive records and a single curator who seeks to determine certain statistics pertaining to the records. We denote X_i as the record possessed by the i-th user. Each X_i is independently generated from an unknown distribution P on an unknown domain $\overline{\mathcal{X}}$. We denote a realization of X_i by x_i. Moreover, there is an agent that mediates communication between each user and the curator. User i passes her record to her agent through a channel ρ^{pre} which stochastically or deterministically maps $\overline{\mathcal{X}}$ to $\widetilde{\mathcal{X}}$. The codomain $\widetilde{\mathcal{X}}$ is the union of set $\{\perp\}$ and set $\mathcal{X}$ that is the expected domain; $\widetilde{\mathcal{X}} = \{\perp\} \cup \mathcal{X}$. The $\perp$ is a special character that represents an error or missing, and the curator knows $\mathcal{X}$. The agent perturbs the received record by Q and submits it to the curator through a channel ρ^{pos}. We denote the outputs of Q and ρ^{pos} by Z_i and $\widetilde{Z}_i$, respectively. They are random variables on $\mathcal{Z}$ and $\widetilde{\mathcal{Z}}$, where $\widetilde{\mathcal{Z}} = \mathcal{Z} \cup \{\perp\}$. Using $\widetilde{Z}_1, \cdots, \widetilde{Z}_n$, the curator estimates certain statistics. In the system model, we redefine LDP.

Definition 3 (Extended LDP)**.** *Given some positive value ϵ and channels ρ^{pos} and ρ^{pre}, we say that perturbation mechanism Q is $(\epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}})$-LDP if we have*

$$\forall x, x' \in \overline{\mathcal{X}}, \widetilde{z} \in \widetilde{\mathcal{Z}}, \quad \frac{\sum_{\widetilde{x},z} \rho^{\mathrm{pos}}(\widetilde{z}|z) Q(z|\widetilde{x}) \rho^{\mathrm{pre}}(\widetilde{x}|x)}{\sum_{\widetilde{x},z} \rho^{\mathrm{pos}}(\widetilde{z}|z) Q(z|\widetilde{x}) \rho^{\mathrm{pre}}(\widetilde{x}|x')} \leq e^{\epsilon}.$$

The system model includes that of the standard LDP. In fact, when ρ^{pre} and ρ^{pos} are the identity functions, the two models are identical. In this chapter, since we mainly consider the case in which the random variables are discrete, we do not strictly distinguish between probability mass functions and probability density functions for simplicity of notation.

We next describe the reason that we consider the two domains $\overline{\mathcal{X}}$ and $\mathcal{X}$. $\mathcal{X}$ is the set of the values that the curator expects as inputs and is known to the curator. $\overline{\mathcal{X}}$ is the set of the values that the users can input and is unknown to the curator. In general, the two sets do not agree. We assume that $\mathcal{X} \subset \overline{\mathcal{X}}$. The relative complement $\overline{\mathcal{X}} \backslash \mathcal{X}$ consists of the elements representing the user inputs that the programmer of the perturbation mechanism cannot expect in advance. Suppose, for example, a curator creates the question, what is your blood type? The curator would expect the answer to be among the options A, B, AB, and O in $\mathcal{X}$. However, in reality, some people have rare blood types. In this example, the set of the rare blood types would be $\overline{\mathcal{X}} \backslash \mathcal{X}$.

We present a more detailed description for ρ^{pre}, ρ^{pos}, and the agents since they do not appear in the standard system model for LDP. An agent is a model of a device such as a smartphone or PC. The agent translates the user's thoughts into an electronic record and passes it to the curator. The curator receive only $\tilde{z}_i$ and cannot see z_i, $\widetilde{x}_i$, or x_i. Unexpected values can appear when the data provider passes a value to her agent or when the agent passes a value to the curator.

To model the mechanisms that produce unexpected values, we use the notion of pre-agent and post-agent mechanisms, ρ^{pre} and ρ^{pos}. A ρ^{pre} is a model of the user interface and the decision-making of the user. In this chapter, we treat ρ^{pre} as a conditional probability function. The probability is conditioned by the true

user input x_i and determines the probability of $\widetilde{x}_i$. Its support should be $\widetilde{\mathcal{X}}$. One example of a phenomenon that would be modeled by a ρ^{pre} is the projection of $\overline{\mathcal{X}}$ to $\mathcal{X}$ by a user interface. When a curator designs an input form as a choice, the user selects a response from among the available choices. Another example is an error or exception of the interface. A modern computer program is highly modular, and some of those modules differ from device to device. It would be unrealistic for a curator to verify all of these behaviors. To the curator, these errors may seem random. A third example is a non-response by users, where the user either refuses to answer the question or mistypes her answer. Non-responses are often discussed in the context of missing-data analysis and are modeled by probabilistic models [Little and Rubin, 2019]. We can interpret non-response as a stochastic replacement of a user's input x with a special symbol $\perp$ as we do in Section 4.4.

A ρ^{pos} is a model of the communication channel between user devices and the curator. We define ρ^{pos} as a conditional probability function whose condition is one of $\mathcal{Z}$ and whose support is $\widetilde{\mathcal{Z}}$. Examples of phenomena that are modeled by a ρ^{pos} would included physical noise and packet loss. Regarding the channel, some bits of a submission may be inverted by physical noise, or they may be lost in part due to packet loss which is a common phenomenon that occurs when communications are concentrated on a single server or router.

It should be noted that a naive application of a mechanism of the standard ϵ-LDP is not necessary to protect privacy in the presence of unexpected values. We here consider the simple pre-agent channel ρ^{pre} that replaces $x \in \overline{\mathcal{X}} \backslash \mathcal{X}$ by $\perp$:

$$\rho^{\text{pre}}(x|x) = 1 \text{ for } x \in \mathcal{X}, \quad \text{and} \quad \rho^{\text{pre}}(\perp|x) = 1 \text{ for } x \in \overline{\mathcal{X}} \backslash \mathcal{X}.$$

Let Q be a ϵ-LDP mechanism, which is a probability distribution on $\mathcal{Z}$ conditioned by $\mathcal{X}$. We assume that Q outputs $\perp$ when and only when Q receives $\perp$. In such cases, there is no finite ϵ such that

$$\frac{Q(\perp|\perp)}{Q(\perp|x)} \leq e^{\epsilon} \quad \text{for} \quad x \in \mathcal{X}. \tag{4.3}$$

This implies that the naive approach fails to protect privacy in the sense of $(\epsilon, \rho^{\text{pre}}, \rho^{\text{pos}})$-LDP. Clearly, the developer of a perturbation mechanism must implement a safe exception handler $Q(\cdot|\perp)$.

This phenomenon is in contrast to a planned partial record deletion, which promotes protection. For example, Bassily et al. provided a minimax optimal algorithm discarding $k-1$ elements of k-ary vector [Bassily and Smith, 2015]. As another example, in the context of the design of (non-locally) differentially private algorithms, planned partial record deletions are called sub-sampling and are used widely [Balle et al., 2018, Dwork and Roth, 2014, Wang et al., 2019]. These techniques are helpful for achieving a better trade-off between privacy and utility. On the other hand, an unexpected data deletion can break privacy preservation completely as we illustrated in the previous paragraph.

Unlike pre-agent channels, any post-agent channel does not harm privacy at all. We formally state this property in the following proposition:

Proposition 3. *Given positive value ϵ and pre-agent channel ρ^{pre}, if a stochastic function Q is $(\epsilon, \rho^{\text{pre}}, 1)$-LDP, where the 1 is the identity function, the Q is $(\epsilon, \rho^{\text{pre}}, \rho^{\text{pos}})$-LDP for any ρ^{pos}.*

This property is immediately obtained from the post-processing invariant, which is described in Section 4.2.1. Moreover, as long as z_i is fixed, no matter how many times z_i is transmitted, no problem with privacy arises. We need not seriously care about the post-agent channels.

Since we do not know the ρ^{pos} and ρ^{pre}, we are unable to precisely evaluate the integration appearing in Definition 3. We now offer a proposition that gives a sufficient condition for Q satisfying Definition 3.

Proposition 4. *Q is $(\epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}})$-LDP for any ρ^{pre} and ρ^{pos} if Q satisfies*

$$\sup_{S\subset\mathcal{Z}} \sup_{\widetilde{x},\widetilde{x}'\in\widetilde{\mathcal{X}}} \frac{Q(S|\widetilde{x})}{Q(S|\widetilde{x}')} \leq e^{\epsilon}. \tag{4.4}$$

Proof. First, we show that Q is $(\epsilon,\ \rho^{\mathrm{pre}},\ 1)$-LDP if Q satisfies (4.4). For any $S \subset \mathcal{Z}$, we have

$$\begin{aligned}\frac{\sum_{\widetilde{x}} Q(S|\widetilde{x})\rho^{\mathrm{pre}}(\widetilde{x}|x)}{\sum_{\widetilde{x}} Q(S|\widetilde{x})\rho^{\mathrm{pre}}(\widetilde{x}|x')} \leq& \frac{\sum_{\widetilde{x}} \sup_{\widetilde{x}'} Q(S|\widetilde{x}')\rho^{\mathrm{pre}}(\widetilde{x}|x)}{\sum_{\widetilde{x}} \inf_{\widetilde{x}'} Q(S|\widetilde{x}')\rho^{\mathrm{pre}}(\widetilde{x}|x')} \\ =& \frac{\sup_{\widetilde{x}'} Q(S|\widetilde{x}')}{\inf_{\widetilde{x}'} Q(S|\widetilde{x}')} = \sup_{\widetilde{x},\widetilde{x}'} \frac{Q(S|\widetilde{x})}{Q(S|\widetilde{x}')} \leq e^{\epsilon}.\end{aligned}$$

Then, using Proposition 3, we have finished the proof. □

This proposition shows the condition under which a perturbation mechanism is an LDP regardless pre- and post-agent channels.

We offer a few remarks to help you achieve more reliable privacy protection in reality. First, the single special character $\perp$ may not be sufficient to represent all exceptions. In the real world, there are many types of exceptions, including, for example, overflow, type mismatch, and time out. If a curator wants to use our system model directly in privacy analysis for a survey, a function must be deployed mapping any value in $\overline{\mathcal{X}}\backslash\mathcal{X}$ to $\perp$ as a part of a pre-agent mechanism. Although it is generally difficult to implement an algorithm for determining whether an input belongs to the expected domain, careful selection of the domain in which such a determination is feasible is important to enabling a real deployment of a locally private survey based on the proposed model.

Second, since our system model does not include interactivity, it is non-trivial whether Proposition 4 holds for interactive algorithms such as SGD. An analysis that include interactivity remains an open problem.

Third, in practice, it is not easy for users to check whether the exception handlers are deployed correctly. It is unrealistic to believe that we can eliminate the possibility that a curator is an adversary and embed a backdoor in the exception handler and the perturbation mechanism. Technologies of formal verification can be helpful in handling this issue [Tschantz et al., 2011, Zhang and Kifer, 2017]. Strictly speaking, we need an official third party to verify the exception handler and perturbation mechanism.

4.4 Lower Bound of Estimation Problem

In the previous section, we discussed privacy analysis. We now analyze the minimax risk of a locally private estimation problem in the presence of unexpected values.

First, we redefine minimax risk for our problem. We additionally consider ρ^{pre} and ρ^{pos}. We embed the new building block into the definition of minimax risk.

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \equiv \inf_{\hat{\theta}_n, Q} \sup_{\theta \in \Theta} \widetilde{M}_{\theta,Q}(w(d(\hat{\theta}_n(\widetilde{Z}_1, ..., \widetilde{Z}_n), \theta))).$$

where $\widetilde{M}_{\theta,Q}$ is the marginal distributions of $Z_1, \cdots, Z_n$ given Q and P_θ.

4.4.1 Abstract Framework

As an analog of Theorem 1, we have an abstract lower bound. To derive the lower bound, we focus on the marginal distributions of $\widetilde{Z}_i$. We denote the marginal distributions of $\widetilde{Z}_i$ by $\widetilde{M}_0$ and $\widetilde{M}_1$ when each X_i follows P_0 and P_1, respectively. We also denote the marginal distributions of $\widetilde{X}_i$ and Z_i by $\widetilde{P}_j$ and M_j for $j = 0, 1$.

Theorem 9. *Let $D_{kl}(\cdot\|\cdot)$ be KL divergence. For a pair of P_0 and P_1, if we have*

$$\left\|\widetilde{P}_0 - \widetilde{P}_1\right\|_{TV} \leq \beta_1 \|P_0 - P_1\|_{TV}, \quad D_{kl}(M_0\|M_1) \leq \beta_2 \left\|\widetilde{P}_0 - \widetilde{P}_1\right\|_{TV}^2,$$
$$\text{and} \quad D_{kl}(\widetilde{M}_0\|\widetilde{M}_1) \leq \beta_3 D_{kl}(M_0\|M_1),$$

then we have

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \geq \frac{w(\delta)}{2}\left(1 - \sqrt{n\beta_1^2\beta_2\beta_3 \|P_0 - P_1\|_{TV}^2}\right). \quad (4.5)$$

Proof. Using the same procedure that yielded inequality (2.4), we obtain

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \geq \frac{w(\delta)}{2}\left(1 - \left\|\widetilde{M}_0^n - \widetilde{M}_1^n\right\|_{\mathrm{TV}}\right).$$

This relation implies that we have a lower bound of R_n if we have an upper bound of $\|\widetilde{M}_0^n - \widetilde{M}_1^n\|_{\mathrm{TV}}$. By Pinsker's inequality, we have $\|\widetilde{M}_0^n - \widetilde{M}_1^n\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{kl}}(\widetilde{M}_0^n\|\widetilde{M}_1^n)$. Since $\widetilde{M}_0^n$ and $\widetilde{M}_1^n$ are products distributions, we have $D_{\mathrm{kl}}(\widetilde{M}_0^n\|\widetilde{M}_1^n) = nD_{\mathrm{kl}}(\widetilde{M}_0\|\widetilde{M}_1)$. From the assumption that $D_{\mathrm{kl}}(\widetilde{M}_0\|\widetilde{M}_1) \leq \beta_3 D_{\mathrm{kl}}(M_0\|M_1)$, we have $\|\widetilde{M}_0^n - \widetilde{M}_1^n\|_{\mathrm{TV}}^2 \leq n\beta_3 D_{\mathrm{kl}}(M_0\|M_1)$. Moreover, from the assumptions that $D_{\mathrm{kl}}(M_0\|M_1) \leq \beta_2\|\widetilde{P}_0 - \widetilde{P}_1\|_{\mathrm{TV}}^2$ and that $\|\widetilde{P}_0 - \widetilde{P}_1\|_{\mathrm{TV}} \leq \beta_1 \|P_0 - P_1\|_{\mathrm{TV}}$, we have

$$\left\|\widetilde{M}_0^n - \widetilde{M}_1^n\right\|_{\mathrm{TV}}^2 \leq n\beta_1^2\beta_2\beta_3 \|P_0 - P_1\|_{\mathrm{TV}}^2.$$

We have finished the proof. □

Theorem 9 implies that we obtain a minimax lower bound when we have the three coefficients β_1, β_2, and β_3 and that we can analyze the three mechanisms separately. Given this property, we can consider a number of specific examples of mechanisms independently and then combine them freely. Since we have already obtained β_2 by (4.2), our interest is on the β_1 and β_3. In the following subsections, we analyze coefficients β_1 and β_3 with concrete situations.

4.4.2 Simple erasure of out-of-domain values

Simple erasure models a user interface that raises an exception for an unexpected value. In a realistic problem, $\perp$ may correspond to the "other" option.

Formally, this is defined as follows:

$$\rho^{\mathrm{pre}}(x|x) = 1 \text{ for } x \in \mathcal{X}, \quad \text{and} \quad \rho^{\mathrm{pre}}(\perp|x) = 1 \text{ for } x \in \overline{\mathcal{X}}\backslash\mathcal{X}. \tag{4.6}$$

With the ρ^{pos}, total variation $\|\widetilde{P}_0 - \widetilde{P}_1\|_{\mathrm{TV}}$ is evaluated as

$$\begin{aligned}\min\left\{\sup_{S\subset\mathcal{X}}|P_0(S) - P_1(S)|, |P_0(\widetilde{\mathcal{X}}\backslash\mathcal{X}) - P_1(\widetilde{\mathcal{X}}\backslash\mathcal{X})|\right\} &\leq \left\|\widetilde{P}_0 - \widetilde{P}_1\right\|_{\mathrm{TV}} \\ &\leq \sup_{S\subset\mathcal{X}}|P_0(S) - P_1(S)| + |P_0(\widetilde{\mathcal{X}}\backslash\mathcal{X}) - P_1(\widetilde{\mathcal{X}}\backslash\mathcal{X})|.\end{aligned}$$

This equation implies that the total variation $\|\widetilde{P}_0 - \widetilde{P}_1\|_{\mathrm{TV}}$ is highly dependent on the selection of P_0 and P_1. Even if $\|P_0 - P_1\|_{\mathrm{TV}}$ is not small, $\|\widetilde{P}_0 - \widetilde{P}_1\|_{\mathrm{TV}}$ can be small. An adversarial example can be used to clarify. Let $\mathcal{X} = \{1, 2\}$, and let $\overline{\mathcal{X}} = \{1, 2, \cdots, k\}$ where k is a natural number greater than 3. We consider the following P_0 and P_1.

$$P_0(x) = \begin{cases} 0 & \text{if } x \neq k-1, \\ 1 & \text{if } x = k-1, \end{cases} \quad \text{and} \quad P_1(x) = \begin{cases} 0 & \text{if } x \neq k, \\ 1 & \text{if } x = k. \end{cases} \tag{4.7}$$

For the P_0 and P_1, we have $\|P_0 - P_1\|_{\mathrm{TV}} = 1$ and $\|\widetilde{P}_0 - \widetilde{P}_1\|_{\mathrm{TV}} = 0$. This is not necessary to imply that the minimax risk does not converge to 0. The total variation is not the only factor deciding the RHS of (4.5), which is a minimax lower bound. Semi-distance d also affects the lower bound. To see the effect of semi-distance on the lower bound, we examine two extreme cases below.

Optimistic case.

We first consider the optimistic case in which the curator loses no utility. Let $\Theta = \{\theta \in \mathbb{R}^k : \sum_{j=1}^k [\theta]_j = 1\}$. In this case, we define semi-distance d as

$$d(\theta, \theta') \equiv |[\theta]_1 - [\theta']_1| = |P_\theta(1) - P_{\theta'}(1)|. \tag{4.8}$$

The curator can use this semi-distance when he assumes that the support set is binary. The semi-distance always takes a positive value for two different Bernoulli distributions. Regarding the semi-distance (4.8), we have the following proposition.

Proposition 5. *On semi-distance d defined in* (4.8)*, for each $0 \leq \delta \leq 1$, there are $P_0, P_1 \in \mathcal{P}$ such that*

$$d(\theta(P_0), \theta(P_1)) = \delta \quad \textit{and} \quad \left\|\widetilde{P}_0 - \widetilde{P}_1\right\|_{TV} \geq \delta$$

where $\theta(P)$ is the parameter of distribution P. Moreover, when the equation holds, there exists a pair (P_0, P_1) such that

$$\left\|\widetilde{P}_0 - \widetilde{P}_1\right\|_{TV} = \|P_0 - P_1\|_{TV} = \delta. \tag{4.9}$$

This proposition asserts that we can set $\beta_1 = 1$ in this case and that we can always select a pair (P_0, P_1) such that the lower bound converges to 0 with $n \to \infty$.

Pessimistic case.

We next consider the pessimistic case. Unlike the optimistic case, we can select a pair such that the lower bound does not converge to 0 even with $n \to \infty$. We define semi-distance d as

$$d(\theta, \theta') = \left| \frac{1}{k} \sum_{j=1}^{k} j[\theta]_j - \frac{1}{k} \sum_{j=1}^{k} j[\theta']_j \right|. \tag{4.10}$$

This semi-distance corresponds to the comparison of the expectations.

Proposition 6. *With semi-distance d defined in* (4.10)*, for each $0 < \delta \leq 1$, there exist $P_0, P_1 \in \mathcal{P}$ such that*

$$d(\theta(P_0), \theta(P_1)) = \delta \quad \textit{and} \quad \left\| \widetilde{P}_0 - \widetilde{P}_1 \right\|_{TV} = 0.$$

Moreover, we have

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \geq \frac{w(\delta)}{2} > 0$$

for any w such that $w(\delta) \neq 0$.

Since the δ is independent of n, this proposition implies that there is no estimator that makes the risk 0 in its worst case.

From these examples, we can see that careful selection of the statistics or semi-distance to be studied may ignore the information loss due to simple deletion, and that improper selection may make estimation completely impossible.

4.4.3 Stochastic erasure as ρ^{pre} and ρ^{pos}

A stochastic-erasure channel replaces its input with $\perp$ with probability. It is a model of missingness due to physical noise and due to the non-response of a user who refuses to respond to a statistical survey. Especially, we consider the case in that the replacement occurs independently of the inputs.

Here, we consider the case in which $\overline{\mathcal{X}} = \mathcal{X}$. Let γ and λ be the erasure rates, where $0 < \gamma < 1$ and $0 < \lambda < 1$. Channels ρ^{pre} and ρ^{pos} are defined as follows:

$$\rho^{\mathrm{pre}}(\widetilde{x}|x) = \begin{cases} 1-\gamma & \text{if } \widetilde{x} = x, \\ \gamma & \text{if } \widetilde{x} = \perp, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad \rho^{\mathrm{pos}}(\widetilde{z}|z) = \begin{cases} 1-\lambda & \text{if } \widetilde{z} = z, \\ \lambda & \text{if } \widetilde{z} = \perp, \\ 0 & \text{otherwise.} \end{cases} \tag{4.11}$$

From Remark 3.2 of [Raginsky, 2016], we see that the coefficients β_1 and β_3 appearing Theorem 9 are $1-\gamma$ and $1-\lambda$, respectively:

$$\left\| \widetilde{P}_0 - \widetilde{P}_1 \right\|_{\mathrm{TV}} \leq (1-\gamma) \left\| P_0 - P_1 \right\|_{\mathrm{TV}}, \tag{4.12}$$

$$\text{and} \quad D_{\mathrm{kl}}(\widetilde{M_0} \| \widetilde{M_1}) \leq (1-\lambda) D_{\mathrm{kl}}(M_0 \| M_1). \tag{4.13}$$

4.5 Examples

In this section, we derive lower and upper bounds for two concrete estimation problems. The objective is to offer a practical algorithm to estimate θ and to evaluate the tightness of the lower bound derived using Theorem 9.

4.5.1 First Example

We use discrete random variable on $\overline{\mathcal{X}} = \{1, 2, \cdots, k\}$, where k is natural number greater than 2, and distribution family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ where $\Theta = \{\theta \in [0,1]^k : \sum_{j=1}^k [\theta]_j = 1\}$. Each distribution P_θ is defined as $P_\theta(j) = [\theta]_j$ for each $j = 1, \cdots, k$. Let d be the semi-distance defined in (4.8), and let $w(t) = t^2$. The curator assumes $\mathcal{X} = \{1, 2\}$. Consider the case in which pre-agent channel ρ^{pre} and post-agent channel ρ^{pos} are simple deletion and the identical function, respectively.

As an instantiation of Theorem 9, we obtain the following lower bound.

Proposition 7. *In the estimation problem described in the previous paragraph, we have*

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \geq \frac{1}{64\epsilon^2 n}.$$

Proof. From (4.9) and (4.2), we can set $\beta_1 = 1$, $\beta_2 = 4\epsilon^2$, $\beta_3 = 1$, and $\|P_0 - P_1\|_{\mathrm{TV}} = \delta$ for any $\delta > 0$. Substituting these values for the βs in Theorem 9, we immediately obtain the following inequality:

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \geq \frac{\delta^2}{2}\left(1 - 2\epsilon\sqrt{n}\delta\right).$$

With $\delta = 1/(4\epsilon\sqrt{n})$, we have

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \geq \frac{1}{2}\left(\frac{1}{4\epsilon\sqrt{n}}\right)^2\left(1 - \frac{1}{2}\right) = \frac{1}{64\epsilon^2 n}.$$

□

Next, we derive an upper bound by providing and analyzing a concrete algorithm. Consider the perturbation mechanism

$$Q(z|\perp) = \begin{cases} \frac{1}{e^\epsilon+1} & \text{if } z = 1 \\ \frac{e^\epsilon}{e^\epsilon+1} & \text{if } z = 2 \end{cases}, \quad Q(z|\widetilde{x}) = \begin{cases} \frac{e^\epsilon}{e^\epsilon+1} & \text{if } z = x, \\ \frac{1}{e^\epsilon+1} & \text{if } z \neq x, \end{cases} \quad \text{for } \widetilde{x}, z = 1, 2. \tag{4.14}$$

From Proposition 4, we can immediately establish that this perturbation mechanism is $(\rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}, \epsilon)$-LDP. The definition of $Q(z|\widetilde{x} = \perp)$ is a safe exception handling. The perturbation mechanism here is an extension of the randomized response, which is a standard ϵ-LDP method [Warner, 1965]. Notably, the curator can design this mechanism even if he does not know the domain $\overline{\mathcal{X}}$.

Let C_1 be a random variable representing the count of 1 that the curator observes. That is, $C_1 \equiv \sum_{i=1}^n \mathbb{1}\left(\widetilde{Z}_i = 1\right)$. We refer to the realization as c_1. Given c_1, the curator constructs an estimate of $[\theta]_1$ as follows:

$$[\hat{\theta}(z_1, ..., z_n)]_1 = \frac{1}{e^\epsilon - 1}\left((e^\epsilon + 1)\frac{c_1}{n} - 1\right). \tag{4.15}$$

Proposition 8. *In the problem discussed in this section, for perturbation mechanism Q defined in* (4.14) *and estimator $\hat{\theta}$ defined in* (4.15)*, we have the following inequality:*

$$\mathbb{E}\left[([\hat{\theta}(Z_1,...,Z_n)]_1 - [\theta]_1)^2\right] \leq \frac{e^\epsilon(e^\epsilon+1)}{(e^\epsilon-1)^2 n}.$$

Moreover, for $\epsilon \downarrow 0$, we have $\mathbb{E}\left[([\hat{\theta}(Z_1,...,Z_n)]_1 - [\theta]_1)^2\right] \in O\left(\frac{1}{\epsilon^2 n}\right)$.

The proof appears in Section 4.7. This proposition implies that the lower bound shown in Proposition 7 is achievable at most constant factor and that the algorithm to build the estimator is reasonable. Since the upper and lower bounds do not contain k, we can say that the deletion does not make the problem more difficult. When no agent perturbs the record, this conclusion is trivial; however, since the agent must perturb a record to hide even $\perp$ in this case, this conclusion is not trivial.

4.5.2 Second Example

We consider the case in which ρ^{pre} and ρ^{pos} are the stochastic erasure channels (4.11). Let $\overline{\mathcal{X}} = \mathcal{X} = \{1,2\}$, $\mathcal{P} = \{P_\theta : \theta \in [0,1], P_\theta(1) = \theta\}$, and d is (4.8).

As an instantiation of Theorem 9, we obtain the following lower bound since we know that $\beta_1 = (1-\gamma)^2$, $\beta_2 = 4\epsilon^2$, and $\beta_3 = (1-\lambda)$ in this case.

Proposition 9. *In the proposition described in the previous paragraph, we have*

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \geq \frac{1}{64(1-\gamma)^2\epsilon^2(1-\lambda)n}.$$

Next, we find an upper bound of the minimax risk by constructing a concrete perturbation mechanism and estimator. Consider the following perturbation mechanism Q:

$$Q(z|\perp) = \begin{cases} \frac{1}{e^\epsilon+1} & \text{if } z = 1 \\ \frac{e^\epsilon}{e^\epsilon+1} & \text{if } z = 2, \end{cases} \quad Q(z|\widetilde{x}) = \begin{cases} \frac{e^\epsilon}{e^\epsilon+1} & \text{if } z = \widetilde{x} \\ \frac{1}{e^\epsilon+1} & \text{if } z \neq \widetilde{x} \end{cases} \quad \text{for } \widetilde{x}, z = 1,2. \tag{4.16}$$

We use the following estimator:

$$\hat{\theta}(z_1,\cdots,z_n) \equiv \frac{1}{(e^\epsilon-1)(1-\gamma)}\left(\frac{e^\epsilon+1}{(1-\lambda)n}c_1 - 1\right). \tag{4.17}$$

Proposition 10. *In the problem described in this subsection, with perturbation mechanism* (4.16) *and estimator* (4.17)*, we have*

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \leq \frac{(e^\epsilon+1)e^\epsilon}{(e^\epsilon-1)^2(1-\gamma)^2(1-\lambda)n}.$$

Moreover, for $\epsilon \downarrow 0$, we have

$$R_n(\theta(P), w \circ d, \epsilon, \rho^{\mathrm{pre}}, \rho^{\mathrm{pos}}) \in O\left(\frac{1}{\epsilon^2(1-\gamma)^2(1-\lambda)n}\right).$$

The proof appears in Section 4.8. Two comments regarding Propositions 9 and 10 would seem in order. First, the lower bound derived in Proposition 9 is achievable at most constant factor. This can be seen in the comparison of Propositions 9 and 10. Second, ρ^{pre} would make the accuracy of the estimation problem worse than ρ^{pos} even if ρ^{pre} and ρ^{pos} were essentially the same. In this problem, both ρ^{pre} and ρ^{pos} the identical conditional distribution essentially; however, their impacts on the lower bound differ: the lower bound is proportional to $(1-\lambda)^{-1}$ and is proportional to $(1-\gamma)^{-2}$.

4.6 Conclusion

In this chapter, we asserted that the standard LDP definition is not sufficient to evaluate privacy in the real world, and proposed the modified LDP that can evaluate privacy in the presence of unexpected values. Our privacy analysis implies that we can design a locally private mechanism if the expected domain $\mathcal{X}$ is clearly defined and we can decide that each value $x \in \overline{\mathcal{X}}$ is $x \in \mathcal{X}$ or $x \notin \mathcal{X}$; otherwise, no statistical survey should not proceed. Moreover, we established the framework to analyze the minimax risk for the problem and confirmed that lower bounds are achievable at most constant factor in the two concrete examples. It is our belief that the issues raised and the approach proposed in this chapter will significantly enhance the ability to conduct locally private surveys.

4.7 Proof of Proposition 8

Proof. For each $i = 1, \cdots, n$, we define new random variable W_i: We define new random variable W_i for each $i = 1, ..., n$:

$$W_i \equiv \frac{1}{(e^\epsilon - 1)n}\left((e^\epsilon + 1)\mathbb{1}\left(\widetilde{Z}_i = 1\right) - 1\right).$$

Then, $\hat{\theta}_n(Z_1, ..., Z_n) = \sum_{i=1}^n W_i$. Since

$$\begin{aligned}
&\mathbb{E}\left[W_i\right] \\
=&\frac{1}{(e^\epsilon - 1)n}\left((e^\epsilon + 1)\Pr(\widetilde{Z}_i = 1) - 1\right) \\
=&\frac{1}{(e^\epsilon - 1)n}\left((e^\epsilon + 1)\left(\frac{e^\epsilon}{e^\epsilon + 1}[\theta]_1 + \frac{1}{e^\epsilon + 1}[\theta]_2 + \frac{1}{e^\epsilon + 1}([\theta]_3 + \cdots + [\theta]_k)\right) - 1\right) \\
=&\frac{1}{(e^\epsilon - 1)n}\left((e^\epsilon + 1)\left(\frac{e^\epsilon}{e^\epsilon + 1}[\theta]_1 + \frac{1}{e^\epsilon + 1}[\theta]_2 + \frac{1}{e^\epsilon + 1}(1 - [\theta]_1 - [\theta]_2)\right) - 1\right) \\
=&\frac{1}{n}[\theta]_1,
\end{aligned}$$

The expectation $\mathbb{E}\left[[\hat{\theta}(Z_1, ..., Z_n)]_1\right]$ is $[\theta]_1$, that is, the estimator is unbiased. We used the fact that $[\theta]_3 + \cdots [\theta]_k = 1 - [\theta]_1 - [\theta]_2$. From the unbiasedness,

we have $\mathbb{E}\left[([\hat{\theta}(Z_1,...,Z_n)]_1-[\theta]_1)^2\right]=\mathbb{E}\left[[\hat{\theta}(Z_1,...,Z_n)]_1^2\right]-[\theta]_1^2$. Since

$$\begin{aligned}
&\mathbb{E}\left[W_i^2\right]\\
=&\frac{1}{(e^\epsilon-1)^2}\frac{(e^\epsilon+1)^2}{(1-\lambda)^2n^2}\frac{1-\lambda}{e^\epsilon+1}((e^\epsilon-1)[\theta]_1+1)\left(1-\frac{1-\lambda}{e^\epsilon+1}((e^\epsilon-1)[\theta]_1+1)\right)\\
&+\frac{1}{n^2}[\theta]_1^2\\
=&\frac{1}{(e^\epsilon-1)^2n^2}((e^\epsilon-1)[\theta]_1+1)\left(\frac{e^\epsilon+1}{1-\lambda}-((e^\epsilon-1)[\theta]_1+1)\right)+\frac{1}{n^2}[\theta]_1^2\\
\leq&\frac{e^\epsilon(e^\epsilon+1)}{(e^\epsilon-1)^2(1-\lambda)n^2}+\frac{1}{n^2}[\theta]_1^2
\end{aligned}$$

and $\mathbb{E}\left[W_iW_{i'}\right]=\frac{1}{n^2}[\theta]_1^2$ for $i\neq i'$, we have

$$\begin{aligned}
\mathbb{E}\left[([\hat{\theta}(Z_1,...,Z_n)]_1-[\theta]_1)^2\right]=&\mathbb{E}\left[([\hat{\theta}(Z_1,...,Z_n)]_1)^2\right]-[\theta]_1^2\\
=&\mathbb{E}\left[\left(\sum_{i=1}^n W_i\right)^2\right]-[\theta]_1^2\\
=&\mathbb{E}\left[\sum_{i=1}^n W_i^2\right]+\mathbb{E}\left[\sum_{i\neq i'}^n W_iW_{i'}\right]-[\theta]_1^2\\
\leq&\frac{e^\epsilon(e^\epsilon+1)}{(e^\epsilon-1)^2(1-\lambda)n}+[\theta]_1^2-[\theta]_1^2\\
=&\frac{e^\epsilon(e^\epsilon+1)}{(e^\epsilon-1)^2(1-\lambda)n}.
\end{aligned}$$

This inequality is the first half of the proposition.

Next, we show the second half. Since, by Taylor expansion, $e^\epsilon\approx 1+\epsilon$ for enough small ϵ, we immediately obtain the second half. □

4.8 Proof of Proposition 10

Proof. For each $i=1,\cdots,n$, we define new random variable W_i:

$$W_i\equiv\frac{1}{(e^\epsilon-1)(1-\gamma)n}\left(\frac{e^\epsilon+1}{(1-\lambda)}\mathbb{1}\left(\widetilde{Z}_i=1\right)-1\right).$$

We denote a realization of W_i by w_i. These random variables satisfies $\hat{\theta}(z_1,\cdots,z_n)=\sum_{i=1}^n w_i$. Since

$$\begin{aligned}
\mathbb{E}\left[W_i\right]=&\frac{1}{(e^\epsilon-1)(1-\gamma)n}\left(\frac{e^\epsilon+1}{1-\lambda}\Pr(\widetilde{Z}_i=1)-1\right)\\
=&\frac{1}{(e^\epsilon-1)(1-\gamma)n}\left(\frac{e^\epsilon+1}{1-\lambda}(1-\lambda)\left(\frac{e^\epsilon-1}{e^\epsilon+1}\theta(1-\gamma)+\frac{1}{e^\epsilon+1}\right)-1\right)=\frac{1}{n}\theta,
\end{aligned}$$

we have

$$\mathbb{E}\left[\hat{\theta}(Z_1,\cdots,Z_n)\right]=\theta.$$

Thus,

$$\begin{aligned}\mathbb{E}\left[(\hat{\theta}(Z_1,\cdots,Z_n)-\theta)^2\right] =& \mathbb{E}\left[\left(\sum_{i=1}^{n} W_i\right)^2\right] - \theta^2 \\ =& \sum_{i=1}^{n}\mathbb{E}\left[W_i^2\right] + \sum_{i'\neq i}\mathbb{E}\left[W_i W_{i'}\right] - \theta^2.\end{aligned}$$

For each $i = 1, \cdots, n$,

$$\begin{aligned}\mathbb{E}\left[W_i^2\right] =& \frac{1}{(e^\epsilon-1)^2(1-\gamma)^2 n^2}\left(\frac{e^\epsilon+1}{1-\lambda}\right)^2(1-\lambda)\left(\frac{e^\epsilon-1}{e^\epsilon+1}\theta(1-\gamma)+\frac{1}{e^\epsilon+1}\right) \\ & \times\left(1-(1-\lambda)\left(\frac{e^\epsilon-1}{e^\epsilon+1}\theta(1-\gamma)+\frac{1}{e^\epsilon+1}\right)\right)+\frac{1}{n^2}\theta^2 \\ \leq& \frac{1}{(e^\epsilon-1)^2(1-\gamma)^2 n^2}\frac{(e^\epsilon+1)^2}{1-\lambda}\left(\frac{e^\epsilon-1}{e^\epsilon+1}+\frac{1}{e^\epsilon+1}\right)\times 1+\frac{1}{n^2}\theta^2 \\ =& \frac{(e^\epsilon+1)e^\epsilon}{(e^\epsilon-1)^2(1-\gamma)^2(1-\lambda)n^2}+\frac{1}{n^2}\theta^2\end{aligned}$$

and, for each $i' \neq i$, $\mathbb{E}\left[W_i W_{i'}\right] = \frac{1}{n^2}\theta^2$. Thus, we have

$$\mathbb{E}\left[(\hat{\theta}(Z_1,\cdots,Z_n)-\theta)^2\right] \leq \frac{(e^\epsilon+1)e^\epsilon}{(e^\epsilon-1)^2(1-\gamma)^2(1-\lambda)n}.$$

□

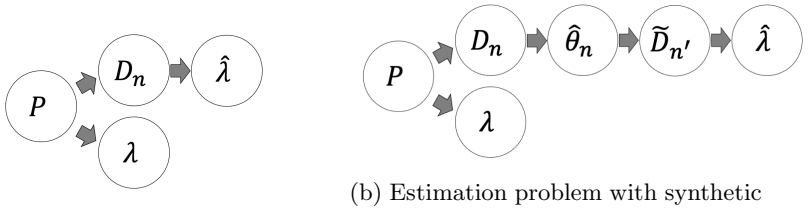
Chapter 5

Inconsistency Due to Synthetic-data Use

5.1 Introduction

Publishing synthetic data instead of raw microdata is one way to balance privacy protection and data use [Rubin, 1993, Elliot and Domingo Ferrer, 2018]. Microdata consisting of individuals' records can be used in a variety of ways because such data provide users with much more flexibility for data analysis than summary statistics. On the other hand, because microdata may contain sensitive information about real individuals, there are strict restrictions on the usage of microdata for surveys and research. A practical way to make microdata available to a large number of people while protecting the privacy of the provider is to release synthetic data that retain the statistical properties of the original data. For example, the US [Kinney et al., 2011, 2014] and Germany [Drechsler et al., 2007, 2008] have been publishing synthetic data of official microdata. With the development of research on generative models [Neunhoeffer et al., 2021] and the availability of R packages, such as synthpop [Nowok, 2016, Nowok et al., 2016], for creating synthetic data, the creation of synthetic data has been easier every year, gaining importance in society.

Privacy can be protected by releasing synthetic data instead of raw data, but how about the utility of synthetic data in statistical analysis? When synthetic data have high similarity to the raw data based on some similarity metric, do we always obtain similar results for a specific task with the synthetic and raw data? This issue has been studied with respect to both general and specific utilities [Snoke et al., 2018]. The general utility is a measure of similarity between the distribution of raw data and that of synthetic data. Examples include propensity score, KL divergence, and Wasserstein distance. When a data owner selects a parameter of the generative model for synthetic data, the general utility may be related to the objective function for the parameter selection. It is reasonable to rely on such an indicator when a specific application is not known in advance. The general utility is controversial and has been studied in various ways [Woo et al., 2009]. As an alternative, a specific utility is a score assigned to the solution of a specific task using synthetic data. This type of utility is of more interest to data users and external analysts with specific goals than the



(a) Ordinary estimation problem

(b) Estimation problem with synthetic data

Figure 5.1: Comparison of Bayesian networks of two estimation problems. In both problems, $\lambda(P)$ is the target statistic that an analyst wants to know. (A) Network of an ordinary estimation problem. Estimation $\hat{\lambda}(D_n)$ is directly obtained from raw data D_n. (B) Network of an estimation problem with synthetic data (problem discussed in this chapter). Estimation $\hat{\lambda}(\widetilde{D}_{n'})$ is obtained from synthetic data $\widetilde{D}_{n'}$, and the analyst cannot access the raw data D_n.

general utility. Specific utilities have been evaluated in some situations, mainly through experimental evaluations [Reiter, 2005a, Raghunathan et al., 2003].

However, it is difficult to establish a general principle for predicting various specific utilities by extracting the results of numerical and empirical evaluations. We aim to establish a general theory for analyzing the relationship between specific accuracy and synthetic data. In this chapter, we answer the following key question: When does a serious failure of a statistical estimation occur for synthetic data users who cannot access the raw data or the population?

To analyze this problem quantitatively, we formulate the estimation problem with synthetic data as follows. Nature chooses a distribution P from a family of distributions $\mathcal{P}$. From that distribution, a data owner independently generates data D_n of size n. In particular, the data owner selects a distribution P_θ from the distribution family $\{P_\theta : \theta \in \Theta\}$ that approximates D_n well, and generates and publishes synthetic data $\tilde{D}_{n'}$ using P_θ. An external analyst uses $\tilde{D}_{n'}$ to obtain an estimate $\hat{\lambda}(\tilde{D}_{n'})$ for the target statistic $\lambda(P)$. The external analyst does not know the true data distribution P. Figure 5.1 shows a comparison between the Bayesian networks of the problem in this chapter and an ordinary estimation problem. In our model shown in Figure 5.1, the general utility is the similarity between P and P_θ, and the specific utility is a measure of how close the statistic $\lambda(P)$ is to the estimator $\hat{\lambda}(\tilde{D}_{n'})$.

We show that there exists an estimation problem in which an external analyst cannot construct any consistent estimator of the target statistic $\lambda(P)$. To identify such a problem, we take an information-theoretic approach, minimax risk analysis. Minimax risk is a measurement of the difficulty of an estimation problem and is defined as the estimation error of the optimal estimator in its worst case. If the minimax risk does not converge to 0, no estimator can be consistent in its worst case. To obtain a lower bound, we reduce the estimation problem to a binary testing problem in which an analyst determines which of P_0 and P_1 is selected to generate synthetic data $\tilde{D}_{n'}$. In the analysis, we consider two kinds of distances simultaneously. The first distance is $\rho(\lambda(P_0), \lambda(P_1))$. If this (sub-)distance is positive, then $\lambda(P_0)$ and $\lambda(P_1)$ have significantly different values. The second one is the distance between the distributions of $\tilde{D}_{n'}$ when $P = P_0$ and $P = P_1$. If this distance is positive, then the analyst can distin-

guish P_0 and P_1 from the observation of the synthetic data. As a consequence of the analysis, we find a problem in which the first distance is positive and the second distance is zero. We show that there is no consistent estimator in such a problem.

To concretely illustrate the problem, we provide a concrete pair of a model family and a target statistic. Exponential families constitute a broad class of stochastic models and include the Gaussian and Poisson distributions. We consider the case in which the data owner chooses $\hat{\theta}_n$ as a maximum likelihood estimator (MLE). When an exponential family satisfies some regularity conditions, the MLE converges in distribution to a Gaussian random variable. Using this property, we determine a condition such that an external analyst cannot build a consistent estimator. To visualize the inconsistency, we perform numerical experiments with artificial data. Since exponential families are widely used, this example is a warning for many data owners and analysts.

The remainder of this chapter is organized as follows. In Section 5.2, we formulate the focal problem. In Section 5.3, we show a sufficient condition such that there exists no consistent estimator. In Section 5.4, we provide a concrete example satisfying the sufficient condition. In Section 5.5, we perform a visualization for the inconsistency. In Section 5.6, we discuss some open problems and our future work. We introduce related work in Section 5.7 and draw conclusions in Section 5.8.

5.2 Analytic Target

In this section, we quantitatively define our problem, which is an estimation problem with synthetic data, and the goal of our analysis, the minimax risk of the problem.

First, we define the flow of data generation and estimator construction with stakeholders. From a family $\mathcal{P}$ of distributions, nature chooses a distribution P. The data D_n of size n consist of independent samples from the identical distribution P. The data holder selects a model P_θ from the model family $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$ that is most likely to be the distribution generating D_n and generates and publishes synthetic data $\tilde{D}_{n'}$, whose size is n', from P_θ. The generated synthetic data $\tilde{D}_{n'}$ are published. An external analyst uses $\tilde{D}_{n'}$ to compute an estimation $\hat{\lambda}(\tilde{D}_{n'})$ of the target statistic $\lambda(P)$. The objective of the analyst is to obtain estimation $\hat{\lambda}(\tilde{D}_{n'})$ minimizing $\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P))$, where ρ is a semi-distance. The difference from the traditional estimation problem introduced in Section 2.2 of this problem is that the estimation is performed using the synthetic data $\tilde{D}_{n'}$ instead of using the raw data D_n directly.

Next, we define the risk and minimax risk for this problem. The maximum risk of an estimator $\hat{\lambda}$ is defined as

$$\sup_{P \in \mathcal{P}} P(w(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P))))),$$

where w and ρ are the same ones defined in Section 2.2. The minimax risk of this problem is defined as

$$\mathcal{R}^*_{n,n'}(\mathcal{P}, \lambda, \mathcal{P}_\Theta, w \circ \rho) \equiv \inf_{\hat{\lambda}} \sup_{P \in \mathcal{P}} P(w(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P))))).$$

The infimum is taken over the set of all estimators.

We here emphasize some difference between this problem and the classical estimation problem described in Section 2.2. In this problem, the analyst uses synthetic data instead of raw data. The distributions of the synthetic and raw data are not identical. We consider the situation that the synthetic data come from a misspecified model. Moreover, the objective is different. In this problem, the analyst wants to estimate the target statistic $\lambda(P)$ instead of $\theta(P)$ which is the model parameter for the synthetic data generation.

On the problem formulation, we can restate our research question as follows: Does there exist a combination of λ and $\mathcal{P}_\Theta$ such that the minimax risk does not converge to 0?

5.3 Minimax lower bound analysis

In this section, we provide a framework for analyzing the minimax lower bound of this problem. Then, on the basis of the framework, we show when serious estimation errors occur. As in the standard minimax risk analysis described in Section 2.3, we analyze the minimax lower bound by reducing the estimation problem to a hypothesis testing problem.

We first derive a lower bound, which is characterized by the probability of estimation $\hat{\lambda}(\tilde{D}_{n'})$ being far from the true value of the target statistic $\lambda(P)$, of the risk. For any positive real value δ such that $w(\delta) > 0$, we have

$$P(w(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P)))) \geq w(\delta) P(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P)) \geq \delta).$$

Next, we focus on the probability $P(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P)) \geq \delta)$, which appears in the right-hand side of the above inequality. It is clear that

$$\inf_{\hat{\lambda}} \sup_{P \in \mathcal{P}} P(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P))) \geq \delta) \geq \inf_{\hat{\lambda}} \max_{j \in \{0,1\}} P_j(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P_j)) \geq \delta). \quad (5.1)$$

We call the hypotheses P_0, P_1 and call a test any measurable function $\psi : \mathcal{X}^{n'} \to \{0, 1\}$. We select the hypotheses P_0 and P_1 such that

$$\rho(\lambda(P_0), \lambda(P_1)) \geq 2\delta.$$

Lemma 7. *For any estimator $\hat{\lambda}$, we have*

$$P_j(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P_j)) \geq \delta) \geq P_j(\psi^*(\tilde{D}_{n'}; \hat{\lambda}) \neq j), \quad j = 0, 1, \quad (5.2)$$

where $\psi^* : \mathcal{X}^{n'} \to \{0, 1\}$ is the minimum distance test defined by

$$\psi^*(\tilde{d}_{n'}; \hat{\lambda}) = \arg\min_{j \in \{0,1\}} \rho(\hat{\lambda}(\tilde{d}_{n'}), \lambda(P_j)) \quad \textit{for each } \tilde{D}_{n'} = \tilde{d}_{n'} \in \mathcal{X}^{n'}.$$

Proof. For notational simplicity, we write $\psi^*(\tilde{d}_{n'}; \hat{\lambda})$ as just ψ^* in this proof. It is sufficient to show

$$\psi^* \neq j \implies \rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P_j)) \geq \delta \quad \text{for each } j = 0, 1.$$

From the triangle inequality, we have

$$\rho(\lambda(P_j), \lambda(P_{1-j})) \leq \rho(\lambda(P_j), \hat{\lambda}(\tilde{D}_{n'})) + \rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P_{1-j})).$$

By the definition of ψ^*, we have $\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P_j)) \geq \rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P_{1-j}))$ with $\psi^* \neq j$. Thus,

$$\rho(\lambda(P_j), \lambda(P_{1-j})) \leq 2\rho(\lambda(P_j), \hat{\lambda}(\tilde{D}_{n'})).$$

Finally,

$$\psi^* \neq j \implies 2\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P_j)) \geq 2\delta \quad \text{for each } j = 0, 1.$$

□

Equation (2.3) is shown in a similar manner. Combining (5.1) and (5.2), we obtain

$$\inf_{\hat{\lambda}_n} \sup_{P \in \mathcal{P}} P(\rho(\hat{\lambda}(\tilde{D}_{n'}), \lambda(P)) \geq \delta) \geq \tilde{p}_{\mathrm{e}} \quad \text{where} \quad \tilde{p}_{\mathrm{e}} \equiv \inf_{\psi} \max_{j=0,1} P_j(\psi(\tilde{D}_{n'}) \neq j).$$

This inequality implies that the minimax risk in this problem is lower bounded using the error probability of the testing problem as in the standard estimation problem described in Chapter 2.

We next lower bound the error probability $\tilde{p}_{\mathrm{e}}$ by modifying Lemma 2. The key tool is the marginal distribution of $\tilde{D}_{n'}$. Let $\mathcal{A}'$ be a σ-algebra on $\mathcal{X}^{n'}$. We define $\tilde{P}_0^{(n,n')}$ and $\tilde{P}_1^{(n,n')}$ that are the marginal distributions of $\tilde{D}_{n'}$ conditioned on $j = 0, 1$ as follows.

$$\tilde{P}_j^{(n,n')}(S) = \int \mathbb{1}(\hat{\lambda}(\tilde{d}_{n'}) \in S) dP_{\hat{\theta}_n(d_n)}(\tilde{d}_{n'}) dP_j(d_n) \text{ for } S \in \mathcal{A}', j = 0, 1$$

where d_n and $\tilde{d}_{n'}$ are realizations of the raw and synthetic data, respectively. The $\mathcal{A}$ must be a σ-algebra such that $\tilde{P}_j^{(n,n')}$ is measurable. We obtain the following lemma replacing P_0^n and P_1^n in Lemma 2 with $\tilde{P}_0^{(n,n')}$ and $\tilde{P}_1^{(n,n')}$.

Lemma 8. *If* $\left\| \tilde{P}_0^{(n,n')} - \tilde{P}_1^{(n,n')} \right\|_{TV} \leq \alpha < 1$, *then*

$$\tilde{p}_{\mathrm{e}} \geq \frac{1-\alpha}{2}.$$

Proof. We first confirm $P_0(\psi(\tilde{D}_n) \neq 0) = \tilde{P}_0^{n,n'}(\psi(\tilde{D}_n) \neq 0)$ and $P_1(\psi(\tilde{D}_{n'}) \neq 1) = \tilde{P}_1(\psi(\tilde{D}_{n'}) \neq 1)$. Thus, we have the following inequality:

$$\begin{aligned}
\tilde{p}_{\mathrm{e}} = \inf_{\psi} \max_{j=0,1} P_j(\psi(\tilde{D}_{n'}) \neq j) &\geq \frac{1}{2} \inf_{\psi} (P_0(\psi(\tilde{D}_{n'}) \neq 0) + P_1(\psi(\tilde{D}_{n'}) \neq 1)) \\
&= \frac{1}{2}(P_0(\psi^*(\tilde{D}_{n'}) \neq 0) + P_1(\psi^*(\tilde{D}_{n'}) \neq 1)) \\
&= \frac{1}{2}(\tilde{P}_0(\psi^*(\tilde{D}_{n'}) \neq 0) + \tilde{P}_1(\psi^*(\tilde{D}_{n'}) \neq 1)),
\end{aligned}$$

where ψ^* is the maximum likelihood test

$$\psi^*(\tilde{d}_{n'}) = \begin{cases} 0, & \text{if } \tilde{p}_0(\tilde{d}_{n'}) \geq \tilde{p}_1(\tilde{d}_{n'}), \\ 1, & \text{otherwise} \end{cases} \quad \text{for each } \tilde{D}_{n'} = \tilde{d}_{n'} \in \mathcal{X}^{n'},$$

in which $\tilde{p}_0$ and $\tilde{p}_1$ are the densities of $\tilde{P}_0^{(n,n')}$ and $\tilde{P}_1^{(n,n')}$ with respect to ν, a measure dominating both of $\tilde{P}_0^{(n,n')}$ and $\tilde{P}_1^{(n,n')}$. From Lemma 2.1 of [Tsybakov, 2009], we have

$$\begin{aligned}\frac{1}{2}(\tilde{P}_0(\psi^*(\tilde{D}_n) \neq 0) + \tilde{P}_1(\psi^*(\tilde{D}_n) \neq 1)) =& \frac{1}{2}\int \min(d\tilde{P}_0^{(n,n')}, d\tilde{P}_1^{(n,n')}) \\ =& \frac{1}{2}\left(1 - \left\|\tilde{P}_0^{(n,n')} - \tilde{P}_1^{(n,n')}\right\|_{\mathrm{TV}}\right) \\ \geq& \frac{1-\alpha}{2}.\end{aligned}$$

□

Moreover, by data processing inequality, we can obtain a lower bound using the distributions of the model parameters instead of the distribution of the synthetic data. Letting J be the random variable selecting j, we have Markov chain $J \to D_n \to \hat{\theta}_n(D_n) \to \tilde{D}_{n'}$. From the Markov chain, we have the data processing inequality

$$\left\|\tilde{P}_0^{(n,n')} - \tilde{P}_1^{(n,n')}\right\|_{\mathrm{TV}} \leq \left\|\hat{P}_0^{(n)} - \hat{P}_1^{(n)}\right\|_{\mathrm{TV}} \quad \text{for any} \quad n, n' \in \mathbb{N}, \tag{5.3}$$

$$\text{where} \quad \hat{P}_j^{(n)}(S) = \int \mathbb{1}(\hat{\theta}_n(d_n) \in S) dP_j(d_n) \quad \text{for } S \in \mathcal{A}'', j = 0, 1.$$

Here $\mathcal{A}''$ is an appropriate σ-algebra on Θ. $\hat{P}_0^{(n)}, \hat{P}_1^{(n)}$ are the marginal distributions of $\hat{\theta}_n(D_n)$ conditioned on $j = 0, 1$, respectively. Equation (5.3) is immediately obtained from the fact that the total variation is an f-divergence and a basic property of f-divergences. If we select the α which appears in Lemma 8 such that $\left\|\hat{P}_0^{(n)} - \hat{P}_1^{(n)}\right\|_{\mathrm{TV}} < \alpha$, then α automatically satisfies $\left\|\tilde{P}_0^{(n,n')} - \tilde{P}_1^{(n,n')}\right\|_{\mathrm{TV}} < \alpha$. This gives us the following lemma.

Lemma 9. *Let $\hat{P}_0^{(n)}$ and $\hat{P}_1^{(n)}$ be the marginal distributions of $\hat{\theta}_n(D_n)$ conditioned on $j = 0, 1$, respectively. If $\left\|\hat{P}_0^{(n)} - \hat{P}_1^{(n)}\right\|_{TV} \leq \alpha < 1$, then we have $\tilde{p}_{\mathrm{e}} \geq (1-\alpha)/2$.*

This lemma implies that, if $\left\|\hat{P}_0^{(n)} - \hat{P}_1^{(n)}\right\|_{\mathrm{TV}} = 0$, then $\tilde{p}_{\mathrm{e}} \geq 1/2$. When a direct analysis of the distributions of $\tilde{D}_{n'}$ is difficult, this lemma allows us to analyze the distributions of $\hat{\theta}_n(D_n)$.

Combining the results so far, we have obtained a lower bound of our problem as follows.

Lemma 10. *Let $\hat{P}_0^{(n)}(D_n)$ and $\hat{P}_1^{(n)}$ be the marginal distributions of $\hat{\theta}_n$ conditioned on $j = 0, 1$, respectively. If $\left\|\hat{P}_0^{(n)} - \hat{P}_1^{(n)}\right\|_{TV} \leq \alpha < 1$, then we have*

$$\mathcal{R}^*_{n,n'}(\mathcal{P}, \lambda, \mathcal{P}_\Theta, w \circ \rho) \geq w(\delta)\frac{1-\alpha}{2}.$$

□

Using Lemma 10, we immediately obtain the following theorem.

Theorem 10. *Suppose that the weighting function w is strictly increasing. If there exist P_0 and $P_1 \in \mathcal{P}$ such that*

$$\left(\left\|\tilde{P}_0^{(n,n')} - \tilde{P}_1^{(n,n')}\right\|_{TV} \leq c' < 1 \quad or \quad \left\|\hat{P}_0^{(n)} - \hat{P}_1^{(n)}\right\|_{TV} \leq c' < 1\right)$$
$$and \quad \rho(\lambda(P_0), \lambda(P_1)) = c > 0 \tag{5.4}$$

for any natural numbers n, n' and some universal constants c', c, minimax risk $\mathcal{R}^*_{n,n'}(\mathcal{P}, \lambda, \mathcal{P}_\Theta, w \circ \rho)$ does not converge to 0 as $n, n' \uparrow \infty$.

We can make some remarks about this theorem. First, if there exist only P_0, P_1 such that $\left\|\hat{P}_0^{(n)} - \hat{P}_1^{(n)}\right\|_{\mathrm{TV}} \leq c' \implies \rho(\lambda(P_0), \lambda(P_1)) = 0$, then c is always 0. In this case, we only have the trivial lower bound, $\mathcal{R}^*_{n,n'} \geq 0$. We can also note that this theorem directly answers the question posed in Section 5.1. Under condition (5.4), any estimator will be inconsistent in its worst case. Thus, (5.4) is the condition causing a serious failure of estimation, which is what we were looking for.

Corollary 4. *Suppose that there exists a constant $C > 0$ such that $\rho(\lambda(P), \lambda(P')) < C$ for any $P, P' \in \mathcal{P}$. If there exist P_0 and P_1 satisfying (5.4), then, for any estimator $\hat{\lambda}$, there exists P such that $\hat{\lambda}(\tilde{D}_{n'})$ does not converge to $\lambda(P)$ in probability as $n, n' \to +\infty$.*

Proof. We here show the contraposition of the corollary.

We assume that there exists a consistent estimator $\hat{\lambda}$. For any $\epsilon > 0$, we have

$$\begin{aligned}
&P(w(\rho(\hat{\lambda}(\widetilde{D}_{n'}), \lambda(P)))) \\
=&P(\mathbb{1}(\rho(\hat{\lambda}(\widetilde{D}_{n'}), \lambda(P)) > \epsilon) w(\rho(\hat{\lambda}(\widetilde{D}_{n'}), \lambda(P)))) \\
&+ P(\mathbb{1}(\rho(\hat{\lambda}(\widetilde{D}_{n'}), \lambda(P)) \leq \epsilon) w(\rho(\hat{\lambda}(\widetilde{D}_{n'}), \lambda(P)))) \\
\leq& w(C) \cdot P(\rho(\hat{\lambda}(\widetilde{D}_{n'}), \lambda(P)) > \epsilon) + \epsilon \cdot P(\rho(\hat{\lambda}(\widetilde{D}_{n'}), \lambda(P)) \leq \epsilon).
\end{aligned}$$

By the definition of convergence in probability, we can take an arbitrary small ϵ and can make $P(\rho > \epsilon)$ arbitrarily small by choosing n sufficiently large. Thus, if there exists a consistent estimator, the minimax risk converges to 0 as $n \to \infty$. □

Our next interest is whether a pair P_0, P_1 satisfying condition (5.4) can exist for some practical problems, and if so, which ones? If no problem satisfying the condition existed, then the condition would be completely meaningless.

5.4 An Example of Inconsistency

We find that one situation in which P_0, P_1 satisfy (5.4) with $c' = 0$ in Theorem 10 is when $\hat{\theta}_n$ is the QMLE for the parameter of a certain subfamily of some *exponential family* and λ is a function satisfying certain conditions.

First, we define $\mathcal{P}_\Theta$. A probabilistic model which is a member of an exponential family of interest has a density function of the following form:

$$p_\theta(\mathbf{x}) = h(\mathbf{x}) \exp\left(\mathbf{b}(\theta)^\top \mathbf{T}(\mathbf{x}) - S(\mathbf{b}(\theta))\right), \tag{5.5}$$

where $\mathbf{b} : \Theta \to \mathbb{R}^{k'}, \mathbf{T} : \mathcal{X} \to \mathbb{R}^{k'}, S : \Theta \to \mathbb{R}$ for a natural number k'. $\mathbf{b}(\theta)$ and $\mathbf{T}$ are called the natural parameter and the sufficient statistic, respectively. Setting these functions concretely, we can represent many distributions. For the sake of discussion, we consider certain subfamilies of exponential families that satisfy some assumptions. The first assumption is that the matrix

$$\frac{\partial \mathbf{b}(\theta)}{\partial \theta} \equiv \begin{pmatrix} \frac{\partial [\mathbf{b}(\theta)]_1}{\partial [\theta]_1} & \cdots & \frac{\partial [\mathbf{b}(\theta)]_{k'}}{\partial [\theta]_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial [\mathbf{b}(\theta)]_1}{\partial [\theta]_k} & \cdots & \frac{\partial [\mathbf{b}(\theta)]_{k'}}{\partial [\theta]_k} \end{pmatrix}$$

is full rank for any $\theta \in \Theta$. That is, there is a one-to-one correspondence between $\mathbf{b}(\theta)$ and θ. For this assumption, we take $k' = k$. The second assumption is that $S(\mathbf{b}(\theta))$ is injective. That is, $\theta \neq \theta' \implies S(\mathbf{b}(\theta)) \neq S(\mathbf{b}(\theta'))$.

Given realization $d_n = (\mathbf{x}_1, ..., \mathbf{x}_n)$ of data and the $\mathcal{P}_\Theta$ defined immediately above, we consider QMLE to select a good p_θ from $\mathcal{P}_\Theta$. The log-likelihood function is written as follows:

$$L_n(\theta | d_n) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta | \mathbf{x}_i) \quad \text{where} \quad \ell(\theta | \mathbf{x}) = \log p_\theta(\mathbf{x}).$$

As we saw in Section 2.4, a QMLE sequence $\{\hat{\theta}_n\}_n$ has asymptotic normality under some regularity conditions. The following lemma gives a sufficient condition on data distributions P_0 and P_1 for the two QMLE sequences for P_0 and P_1 converging in distribution to the same normal random variable .

Lemma 11. *Suppose that Assumptions 1 to 6 hold. If P_0 and P_1 satisfy*

$$P_0[\mathbf{T}(\mathbf{X})] = P_1[\mathbf{T}(\mathbf{X})] \quad \textit{and} \quad P_0[\mathbf{T}(\mathbf{X})\mathbf{T}(\mathbf{X})^\top] = P_1[\mathbf{T}(\mathbf{X})\mathbf{T}(\mathbf{X})^\top], \tag{5.6}$$

then the sequences of $\hat{\theta}_n$ for P_0 and P_1 converge in distribution to the same normal random variable.

Proof. For a data distribution P, the QMLE sequence $n \to \infty$ converges to θ_P that satisfies the following equation under the regularity conditions:

$$\frac{\partial P[\ell(\theta | \mathbf{X})]}{\partial \theta}\bigg|_{\theta=\theta_P} = P\left[\frac{\partial \ell(\theta | \mathbf{X})}{\partial \theta}\right]\bigg|_{\theta=\theta_P} = 0. \tag{5.7}$$

Since the partial derivative with respect to θ of the log-likelihood function is

$$\frac{\partial \ell(\theta | \mathbf{x})}{\partial \theta} = \frac{\partial \mathbf{b}}{\partial \theta}\mathbf{T}(\mathbf{x}) - \frac{\partial \mathbf{b}}{\partial \theta} S'(\mathbf{b}(\theta)) \quad \text{where} \quad S'(\mathbf{b}(\theta)) \equiv \frac{\partial S(\mathbf{b}(\theta))}{\partial \mathbf{b}(\theta)}, \tag{5.8}$$

(5.7) is evaluated as follows:

$$\frac{\partial \mathbf{b}}{\partial \theta} P[\mathbf{T}(\mathbf{X})] - \frac{\partial \mathbf{b}}{\partial \theta} S'(\mathbf{b}(\theta_P)) = 0.$$

By the full-rank assumption of $\partial \mathbf{b}/\partial \theta$, the above equality is equivalent to

$$P[\mathbf{T}(\mathbf{X})] = S'(\mathbf{b}(\theta_P)).$$

In this form, the true data distribution P appears only on the left-hand side, and the parameter θ appears only on the right-hand side. From this property, we have that if data distributions P_0, P_1 satisfy

$$P_0[\mathbf{T}(\mathbf{X})] = P_1[\mathbf{T}(\mathbf{X})], \tag{5.9}$$

then the QMLEs for these distributions converge in probability to the same point for $n \to \infty$: $\theta_{P_0} = \theta_{P_1}$. This condition is for the distributions of two QMLE sequences for two data distributions converging to normal distributions having identical centers. We still do not know the condition for the covariance matrices of the normal distributions being identical.

Next, we discuss the covariance matrix of the normal random variable that the QMLE sequence converges to. As we saw in Section 2.4, we need to analyze the following two matrices in order to evaluate the covariance matrix:

$$A(\theta_P) = P\frac{\partial^2 \ell(\theta|\mathbf{X})}{\partial\theta^2}\bigg|_{\theta=\theta_P} \quad \text{and} \quad B(\theta_P) = P\left(\frac{\partial \ell(\theta|\mathbf{X})}{\partial\theta}\right)\left(\frac{\partial \ell(\theta|\mathbf{X})}{\partial\theta}\right)^\top\bigg|_{\theta=\theta_P}. \tag{5.10}$$

To analyze $A(\theta)$, we consider the second derivative of the log-likelihood function. By the definition of the log-likelihood function ℓ, its second partial derivative with respect to θ is evaluated as follows:

$$\begin{aligned}
\frac{\partial^2 \ell(\theta|\mathbf{x})}{\partial\theta^2} =& \left(\frac{\partial}{\partial\theta}\frac{\partial \mathbf{b}}{\partial\theta}^\top\right)\mathbf{T}(\mathbf{x}) - \left(\frac{\partial}{\partial\theta}\frac{\partial \mathbf{b}}{\partial\theta}^\top\right)S'(\mathbf{b}(\theta)) - \frac{\partial \mathbf{b}}{\partial\theta}S''(\mathbf{b}(\theta))\frac{\partial \mathbf{b}}{\partial\theta}^\top \\
=& \left(\frac{\partial}{\partial\theta}\frac{\partial \mathbf{b}}{\partial\theta}^\top\right)(\mathbf{T}(\mathbf{x}) - S'(\mathbf{b}(\theta))) - \frac{\partial \mathbf{b}}{\partial\theta}S''(\mathbf{b}(\theta))\frac{\partial \mathbf{b}}{\partial\theta}^\top \\
\text{where} \quad & S''(\mathbf{b}(\theta)) \equiv \frac{\partial^2 S(\mathbf{b}(\theta))}{\partial\theta^2}.
\end{aligned}$$

Thus, since $P[\mathbf{T}(\mathbf{x})] - S'(\mathbf{b}(\theta_P)) = 0$,

$$A(\theta_P) = -\frac{\partial \mathbf{b}}{\partial\theta}S''(\mathbf{b}(\theta))\frac{\partial \mathbf{b}}{\partial\theta}^\top.$$

The right-hand side does not contain $\mathbf{x}$. This implies that $\theta_{P_0} = \theta_{P_1} \implies A(\theta_{P_0}) = A(\theta_{P_1})$.

We proceed to the analysis of $B(\theta_P)$. From the definition of the log-likelihood function ℓ and (5.8), we have

$$\begin{aligned}
& P\left(\frac{\partial \ell(\theta|\mathbf{X})}{\partial\theta}\right)\left(\frac{\partial \ell(\theta|\mathbf{X})}{\partial\theta}\right)^\top \\
=& P\frac{\partial \mathbf{b}}{\partial\theta}\left(\mathbf{T}(\mathbf{X}) - S'(\mathbf{b}(\theta))\right)\left(\mathbf{T}(\mathbf{X}) - S'(\mathbf{b}(\theta))\right)^\top\frac{\partial \mathbf{b}}{\partial\theta}^\top \\
=& \frac{\partial \mathbf{b}}{\partial\theta}\Big(P[\mathbf{T}(\mathbf{X})\mathbf{T}(\mathbf{X})^\top] - P[T(\mathbf{X})]S'(\mathbf{b}(\theta))^\top \\
& \qquad - S'(\mathbf{b}(\theta))P[\mathbf{T}(\mathbf{X})]^\top - S'(\mathbf{b}(\theta))S'(\mathbf{b}(\theta))^\top\Big)\frac{\partial \mathbf{b}}{\partial\theta}^\top.
\end{aligned}$$

From this equation, we can see that, if a pair of data distributions P_0, P_1 and a parameter θ satisfy $P_0[\mathbf{T}(\mathbf{X})] = P_1[\mathbf{T}(\mathbf{X})] = S'(\mathbf{b}(\theta))$, then we have

$$P_0[\mathbf{T}(\mathbf{X})\mathbf{T}(\mathbf{X})^\top] = P_1[\mathbf{T}(\mathbf{X})\mathbf{T}(\mathbf{X})^\top] \implies B(\theta_{P_0}) = B(\theta_{P_1}).$$

□

Lemma 11 does not always imply convergence in total variation. Convergence in distribution is weaker than convergence in total variation. Lemma 11 is not sufficient to obtain (5.4). If the CDFs of two random variables converge to the same function at all points, we say that the two random variables converge in distribution. This argument does not require that the density functions agree. Even if the density functions of two random variables do not agree at some finite number of points, their CDFs can agree. In such a case, the total variation between the random variables is not necessarily zero.

To deal with the issue of convergence, we consider a discretizing operation that maps the estimator to a finite set. This operation models the fact that modern computers cannot handle real numbers and instead approximate them with finite numbers of digits as type float or double. This assumption eliminates the need for us to consider pathological exceptions. For a positive real number Δ, a natural number L, and a point $\theta_0 \in \Theta$,

$$\Theta' \equiv \{\theta_0 + 2\Delta\mathbf{v} : \mathbf{v} = (l_1, l_2, ..., l_j) \quad \text{for } l_1, ..., l_k = 0, ..., L-1, \infty\}.$$

After optimization, estimation $\hat{\theta}$ is projected to Θ' by function π defined by

$$[\pi(\hat{\theta})]_j = \begin{cases} [\theta_0]_j & \text{if } [\hat{\theta}]_j \leq [\theta_0]_j + \Delta, \\ [\theta_0]_j + 2\Delta \left\lceil \frac{[\hat{\theta}]_j - ([\theta_0]_j + \Delta)}{2\Delta} \right\rceil & \text{if } [\theta_0]_j + \Delta < [\hat{\theta}]_j \leq [\theta_0]_j + (2L-1)\Delta, \\ +\infty & \text{if } [\hat{\theta}]_j > [\theta_0]_j + (2L-1)\Delta, \end{cases}$$

where $\lceil \cdot \rceil$ is the ceiling function defined as $\lceil \cdot \rceil : \mathbb{R} \to \mathbb{N}; x \mapsto \min\{n \in \mathbb{N} : x \leq n\}$. In the remainder of this section, let $\hat{P}^{(n)}$ be the distribution of the discretized $\hat{\theta}_n$ for P.

Though the discretization can result in an error in the estimation of $\lambda(P)$, we can ignore this error since we can take Δ as small as we want. In fact, the last digit of the double type is extremely small.

With the discretization and Lemma 11, we obtain convergence in total variation.

Lemma 12. *For any $P_0, P_1 \in \mathcal{P}$ satisfying Assumptions 1 to 6 and condition (5.6) and any positive real number $\delta > 0$, there exists a natural number N such that, for any $n > N$,*

$$\left\| \hat{P}_0^{(n)} - \hat{P}_1^{(n)} \right\|_{TV} \leq \delta.$$

Proof. Given CDF F of $\hat{\theta}_n(D_n)$, the probability function of $\pi(\hat{\theta}_n(D_n))$ is recursively evaluated as

$$\begin{aligned} &P(\pi(\hat{\theta}_n(D_n)) = \theta') \\ =&F([\hat{\theta}_n(D_n)]_j \leq [\theta']_j + \Delta \text{ for } j = 1, ..., d) - P(\pi(\hat{\theta}_n(D_n) \in \text{pre}(\theta')) \end{aligned}$$

where pre : $\Theta' \to$ (subsets of Θ') is defined as

$$\mathrm{pre}(\theta') = \{\theta'' \in \Theta' : [\theta'']_j \leq [\theta']_j \text{ for all } j = 1, ..., k \wedge \theta'' \neq \theta'\}.$$

Since Θ' is a finite set and

$$P(\pi(\hat{\theta}_n(D_n)) \in \mathrm{pre}(\theta')) = \sum_{\theta'' \in \mathrm{pre}(\theta')} P(\pi(\hat{\theta}_n(D_n)) = \theta''),$$

the probability function is evaluated in a finite number of computations of the CDF values. Thus, if two sequences of $\hat{\theta}_n(D_n)$ converge to each other in distribution, the two sequences of $\pi(\hat{\theta}_n(D_n))$ converge in total variation. □

We will also give an example of target statistic λ such that $\rho(\lambda(\hat{P}_0), \lambda(\hat{P}_1)) > 0$ for some P_0, P_1 satisfying (5.4). λ using higher than second moments of $\mathbf{T}(\mathbf{X})$ satisfy the condition. The skewness of $\mathbf{T}(\mathbf{X})$ is a good example of a statistic using such a higher moment. We conclude that there exists a situation satisfying (5.4).

Modifying Corollary 4 using Lemma 12, we obtain the following theorem.

Theorem 11. *Let $w : [0, \infty) \to [0, \infty)$ be a strictly monotone function satisfying (2.1). Let $\lambda : \mathcal{P} \to \mathbb{R}$ be the skewness of $\mathbf{T}(\mathbf{X})$. Suppose that there exist data generating mechanisms $P_0, P_1 \in \mathcal{P}$ satisfying Assumptions 1 to 6 and condition (5.6). Moreover, the skewnesses of $\mathbf{T}(\mathbf{X})$ for P_0 and P_1 take different value. Then, there exists P such that $\hat{\lambda}(\tilde{D}_{n'})$ does not converge to $\lambda(P)$ in probability as $n, n' \to +\infty$.* □

In this theorem, the abstract condition (5.4) is replaced with the regularity conditions and concrete conditions (5.6). We can find a P as mentioned in the theorem. At least one of P_0 and P_1 is such a P. The mixture distributions of P_0 and P_1 can also be the P.

Finally, we offer a few remarks on Theorem 11. Roughly speaking, Theorem 11 says that we can construct no consistent estimator of some statistics which depend on the third or higher moments of $\mathbf{T}(\mathbf{X})$ from the synthetic data. Although this result might not seem surprising, we emphasize the following three points. The first is that we have proven theoretically our intuition. Even if the conclusion is not surprising, the method of proof is nontrivial. The second point is that we do not deny the possibility of estimating second-order statistics. Although the fitting result is given as a solution for the first-order equivalence, we might be able to estimate second-order statistics. The third point is that, Theorem 11 does not make any prediction about when the regularity conditions do not hold. Thus, the general results Theorem 10 and Corollary 4 can potentially lead to a surprising result. We discuss the third point in Section 5.6 as an open problem.

5.5 Numerical Experiments

To illustrate that we cannot distinguish two distributions visually, we perform numerical experiments with artificial data sets consisting of scalar records that are generated from P_0 and P_1. We use two model families satisfying (5.4), fit the data sets to models in model families whose parameters are scalars, and observe

empirical $\hat{P}_0^{(n)}$ and $\hat{P}_1^{(n)}$, which are the distributions of the obtained parameter for each data set. Then, we can see that $\hat{P}_0^{(n)}$ and $\hat{P}_1^{(n)}$ converge toward each other with increasing n. We implemented the experiment with Python code.

5.5.1 Fitting to a Gaussian Model

We first describe the artificial data generation. We utilize the half-normal distribution, whose density function is defined as

$$p(x;\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \quad \text{for} \quad x > 0,$$

where $\sigma > 0$ is the parameter and π is the mathematical constant. The mean of the distribution is $\mu = \sigma\sqrt{2/\pi}$. Then, we define two data-generating distributions P_{r} and P_{l} whose densities are p_{r} and p_{l} defined as

$$p_{\mathrm{r}}(x;\sigma) = p(x+\mu;\sigma) \quad x > -\mu \quad \text{and} \quad p_{\mathrm{l}}(x;\sigma) = p(-x+\mu;\sigma) \quad x < \mu. \tag{5.11}$$

The r and l subscripts indicate right and left. P_{r} and P_{l} have long tails on the right and left, respectively. Both have mean 0 and variance $\sigma^2(1-2/\pi)$. Despite having identical means and variances, their skewnesses differ. The skewnesses of random variables following P_{r} and P_{l} are $\sqrt{2}(4-\pi)/(\pi-2)^{3/2}$ and $-\sqrt{2}(4-\pi)/(\pi-2)^{3/2}$, respectively.

We consider a Gaussian family $\mathcal{P}_\Theta$ with fixed variance $\tau > 0$. The density of a Gaussian distribution is

$$\frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{(y-\theta)^2}{2\tau^2}\right) = \frac{\exp(-y^2/(2\tau^2))}{\tau\sqrt{2\pi}} \exp\left(\frac{2y\theta-\theta^2}{2\tau^2}\right).$$

The family is an instance of an exponential family with

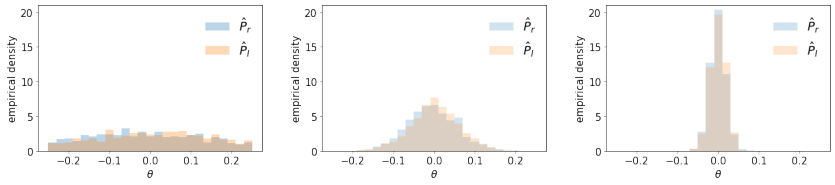
$$h(y) = \frac{\exp(-x^2/(2\tau^2))}{\sigma\sqrt{2\pi}},\ b(\theta) = \frac{2\theta}{2\tau^2},\ T(x) = x,\ S(b(\theta)) = \left(\frac{\theta}{\tau^2}\right)^2 \frac{\tau^2}{2}.$$

We can see that $P_{\mathrm{r}}[T(x)] = P_{\mathrm{l}}[T(x)]$ and $P_{\mathrm{r}}[T(x)^2] = P_{\mathrm{l}}[T(x)^2]$. Thus, since the random variables following P_{r} and P_{l} have different skewnesses from each other, the sufficient condition (5.4) for serious failure is satisfied with $P_0 := P_{\mathrm{r}}, P_1 := P_{\mathrm{l}}, \lambda :=$ skewness, and $\rho = |\cdot|$.

We describe the generation of $\hat{\theta}_n$ as follows. Given data $d_n = (x_1, ..., x_n)$, QMLE $\hat{\theta}_n$ is the $\theta \in \Theta$ that minimizes the log-likelihood function defined as

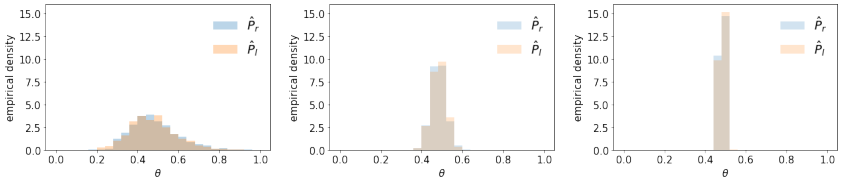
$$L_n(\theta|d_n) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\theta)^2.$$

With $n = 10, 100, 1\,000$, we generate D_n using P_{r} and P_{l} and compute $\hat{\theta}_n$ for each data set. For each n and distribution, we repeat observation 1 000 times and plot the empirical densities of $\hat{\theta}_n(D_n)$. Figure 5.2 shows the fitting results. Each bar shows the number calculated by dividing the number of points in an interval by the width of the interval. Blue and orange bars correspond to the empirical densities of $\hat{P}_{\mathrm{r}}^{(n)}$ and $\hat{P}_{\mathrm{l}}^{(n)}$, respectively. With greater n, the area of blue and orange is less, and the area of brown is greater. This means that it is more difficult to distinguish those empirical distributions with greater n.



(a) $n = 10$ (b) $n = 100$ (c) $n = 1\,000$

Figure 5.2: Empirical densities of $\hat{\theta}_n$ for fitting to a Gaussian model.



(a) $n = 10$ (b) $n = 100$ (c) $n = 1\,000$

Figure 5.3: Empirical densities of $\hat{\theta}_n$ for fitting to a Laplace model.

5.5.2 Fitting to a Laplace Model

In this subsection, we continue to use P_{r} and P_{l} as the data generators.

We employ a Laplace-model family $\mathcal{P}_\Theta$ with a fixed mean $\mu > 0$, where $\Theta = (0, +\infty)$ is the parameter set. The density of a Laplace model P_θ with mean 0 is

$$f_\theta(x) = \frac{1}{2\theta} \exp\left(-\frac{|x|}{\theta}\right) = \exp\left(-\frac{|x|}{\theta} - \log(2\theta)\right).$$

The density family is an instance of an exponential family with

$$h(x) = 1,\ b(\theta) = -1/\theta,\ T(x) = |x|,\ S(b(\theta)) = \log(2\theta).$$

We can see that $P_{\mathrm{r}}[T(x)] = P_{\mathrm{l}}[T(x)]$ and $P_{\mathrm{r}}[T(x)^2] = P_{\mathrm{l}}[T(x)^2]$. Thus, since the random variables following P_{r} and P_{l} have different skewnesses from each other, the sufficient condition (5.4) for serious failure is satisfied with $P_0 := P_{\mathrm{r}}, P_1 := P_{\mathrm{l}}, \lambda :=$ skewness, and $\rho = |\cdot|$.

We describe the generation of $\hat{\theta}_n$ as follows. Given data $d_n = (x_1, ..., x_n)$, QMLE $\hat{\theta}_n$ is the $\theta \in \Theta$ that minimizes the log-likelihood function defined as

$$L_n(\theta | d_n) = \frac{1}{n} \sum_{i=1}^{n} \left(-\frac{|x_i|}{\theta} - \log(2\theta)\right).$$

With $n = 10, 100, 1\,000$, we generate D_n using P_{r} and P_{l} and compute $\hat{\theta}_n$ for each data set. For each n and distribution, we repeat observation 1 000 times and plot the empirical densities of $\hat{\theta}_n$. In Figure 5.3, we can see a similar trend to Figure 5.2.

5.6 Discussion and Future Work

In this study, we show a sufficient condition for analysts being unable to build a consistent estimator and provide a realistic example of the sufficient condition holding. However, we do not recommend how the data holder and analysts should avoid such an unfavorable situation. For the purpose of avoiding such situations, it is important to investigate not only sufficient conditions but also necessary conditions, which remain an open problem.

A general strategy for finding P_0 and P_1 satisfying (5.4) is also an open problem. In Section 5.4, we provided the strategy to find such P_0 and P_1 in a certain fitting problem. The strategy relies on the asymptotic normality theorem and the likelihood equations, which are the equations to find QMLEs. Since these tools are not always available, this strategy is not always useable.

In this chapter, we focused on the case where all estimators are inconsistent and have not yet investigated the case of all estimators being inefficient. We expect that the key technique to analyze the inefficient case is strong data processing inequalities, which evaluate how data processing decreases information. The idea is used for the analysis of minimax risk in the context of machine learning, e.g., [Zhang et al., 2013, Braverman et al., 2016, Duchi and Rogers, 2019b]. Equation (5.3) is simply a data processing inequality, but it may be possible to evaluate it more precisely with a strong data processing inequality.

Our theory does not support data syntheses by non-parametric methods. Non-parametric methods are popular for synthetic data generation, for example, bagging [Drechsler and Reiter, 2011], classification and regression trees [Reiter, 2005b], random forest [Caiola and Reiter, 2010], support vector machine [Drechsler, 2010], and genetic algorithm [Chen et al., 2016]. Supporting these non-parametric methods is our future work.

A generative adversarial network (GAN) [Goodfellow et al., 2014] is a popular method to generate synthetic data. Table data can be generated by GAN [Zhao et al., 2021], and GAN can be used for microdata synthesis. An objective function used for GAN can have multiple minima, and its estimated parameter is often not the global minimum but a local minimum due to the difficulty of finding the global minimum. The GAN does not satisfy Assumption 3, which corresponds to the assumption of a unique minimum of the objective function. Thus, our theory in Section 5.4 does not support GAN use. Development of a theory also covering GAN is our future work.

Techniques for differential privacy [Dwork et al., 2006] can enhance privacy preservation in the publication of synthetic data, and our theory is potentially applicable to the case where differentially private synthetic data are used. Especially, we can immediately provide a sufficient condition for the nonexistence of a consistent estimator when QMLE $\hat{\theta}_n$ or $\tilde{D}_{n'}$ is stochastically perturbed for differential privacy. Such a privacy-preserving strategy is called output perturbation and is popular [Dwork and Roth, 2014]. However, analyzing other preserving strategies, such as objective perturbation and gradient perturbation, is nontrivial.

Inconsistency in statistical estimation is one thing, and privacy protection is another. The inability to estimate a statistic does not necessarily mean that information about the raw data cannot be obtained by reverse-engineering the published statistic. The inability to estimate a statistic does not in itself help to protect privacy, and data holders should keep this in mind.

5.7 Related Work

We describe differences between the present study and studies on the issue of misspecified models, which is a classical topic in the statistics community [White, 1981, 1982]. Such studies have some overlap with ours. Essentially, inconsistency, which we study in this chapter, comes from misspecification of the model family. Thus, misspecification is a necessary condition for inconsistency, but misspecification alone is not sufficient to cause inconsistency.

Our study can be regarded as an exploration of a kind of insufficient statistics. In the sense of classical statistics and information theory, a sufficient statistic $\mathbf{T}(\mathbf{X})$ relative to distribution family $\{f_\theta\}$ is a function such that $I(\theta; \mathbf{T}(\mathbf{X})) = I(\theta; \mathbf{X})$, where I is mutual information [Cover and Thomas, 2006]. From the perspective of sufficient statistics, we can say that condition (5.4) is a sufficient condition for $\tilde{D}_{n'}$ not to be a sufficient statistic of $\lambda(P)$. In this chapter, we explored a more complicated situation than the classical situation and provided a concrete analysis of the problem of practical estimation.

5.8 Conclusion

In this chapter, we found that synthetic data use must result in inconsistent estimators for some statistics and clarified the sufficient condition for such inconsistency. We showed the sufficient condition in a practical problem concretely and visualized the inconsistency with artificial data.

Chapter 6

Conclusion

In this thesis, we considered methods to avoid privacy composition in differential privacy and analyzed their performance. As a possible solution to avoid privacy composition, we consider LDP data. Although LDP helps a data curator avoid privacy composition, LDP raises difficulty in that the curator does not know the domain of the raw data. Due to this difficulty, data providers can input undesirable values, which the curator does not expect. Those undesirable values lead to security holes of privacy protection based on LDP, decreasing the data utility.

First, we developed the simple protocols for building QMLEs from distributed data while guaranteeing ϵ-LDP for the users. The protocols generate perturbed data that satisfies LDP even when the original data contain extremely large values. Moreover, the protocols are provider-friendly; in the protocols, providers submit only one or a few bits to a curator and do not need to wait for one another, and they do not need to perform complex computations such as integration or derivation. We clarified the sufficient conditions for the QMLEs to be consistent and asymptotically normal and showed their limitations. We showed that the sufficient conditions are relaxed with a concrete implementation. Our analysis helps curators understand data without direct observation of the raw data.

Second, we analyzed the utility of LDP data in the presence of general undesirable values. The undesirable values are risks to privacy protection. Curators and users should always prepare some exception handler against undesirable values even when they think that there are no undesirable values in a survey. Moreover, curators should keep in mind that existing performance analysis may be too optimistic since undesirable values can decrease the performance of an estimator.

As another possible solution to avoid privacy composition, we considered the utilization of synthetic data and found some negative results. We found that synthetic data utilization must result in inconsistent estimators for some statistics and clarified the sufficient condition for such inconsistency. This result suggests that a data curator should make effort to avoid the sufficient condition.

Although it is not so difficult to avoid privacy composition, the solutions can easily degrade the utility of perturbed data. Most components of this thesis focused on pointing out the degeneration. We provided only a limited number of concrete algorithms to balance privacy and data utility. To provide concrete

methods to better trade-off between privacy and data utility is our future work.

Chapter 7

Acknowledge

I am grateful to my supervisor, Professor K. Minami for his generous support. I also wish to thank my co-supervisor, Professor H. Hino for his technical advice. Professor S. Mano and Dr. T. Murakami gave me many variable comments on this thesis.

Bibliography

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014.

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*, CCS '16, pages 308–318, New York, NY, USA, 2016. Association for Computing Machinery.

John M. Abowd. The u.s. census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2867, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355520. doi: 10.1145/3219819.3226070. URL https://doi.org/10.1145/3219819.3226070.

Frank D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, page 19–30, New York, NY, USA, 2009. Association for Computing Machinery.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1376–1385, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/kairouz15.html.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus

Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14 (1–2):1–210, jun 2021. ISSN 1935-8237. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, page 1054–1067, New York, NY, USA, 2014. Association for Computing Machinery.

Apple Differential Privacy Team. Learning with privacy at scale, 2017. URL https://machinelearning.apple.com/2017/12/06/learning-with-privacy-at-scale.html.

Hajime Ono, Kazuhiro Minami, and Hideitsu Hino. One-bit submission for locally private quasi-mle: Its asymptotic normality and limitation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2762–2783. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/ono22a.html.

Marcel Neunhoeffer, Steven Wu, and Cynthia Dwork. Private post-GAN boosting. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6isfR3JCbi.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000. URL https://EconPapers.repec.org/RePEc:cup:cbooks:9780521784504.

Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2009.

Krishna B Athreya and Soumendra N Lahiri. *Measure theory and probability theory*. Springer Science & Business Media, 2006.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912526.

Bolin Ding, Harsha Nori, Paul Li, and Joshua Allen. Comparing population means under local differential privacy: With significance and power. In *AAAI*, pages 26–33, 2018.

Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1626–1635. PMLR, 10–15 Jul 2018.

Abhishek Bhowmick, John Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection against reconstruction and its applications in private federated learning. *arXiv preprint arXiv:1812.00984*, 2018.

Adam Smith, Abhradeep Thakurta, and Jalaj Upadhyay. Is interaction necessary for distributed private learning? In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 58–77, 2017. doi: 10.1109/SP.2017.35.

Andrew McGregor, Ilya Mironov, Toniann Pitassi, Omer Reingold, Kunal Talwar, and Salil Vadhan. The limits of two-party differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 81–90, 2010. doi: 10.1109/FOCS.2010.14.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, page 127–135, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746632. URL https://doi.org/10.1145/2746539.2746632.

Di Wang, Marco Gaboardi, and Jinhui Xu. Empirical risk minimization in non-interactive local differential privacy revisited. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 973–982, Red Hook, NY, USA, 2018. Curran Associates Inc.

Cristina Davino, Marilena Furno, and Domenico Vistocco. *Quantile regression: theory and applications*. Wiley series in probability and statistics. Wiley, Hoboken, NJ, 2013.

Kai Zheng, Wenlong Mou, and Liwei Wang. Collect at once, use effectively: Making non-interactive locally private learning possible. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4130–4139. PMLR, 06–11 Aug 2017.

Di Wang, Adam Smith, and Jinhui Xu. Noninteractive locally private learning of linear models via polynomial approximations. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 898–903. PMLR, 22–24 Mar 2019.

Di Wang, Huangyu Zhang, Marco Gaboardi, and Jinhui Xu. Estimating smooth GLM in non-interactive local differential privacy model with public unlabeled data. In Vitaly Feldman, Katrina Ligett, and Sivan Sabato, editors, *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*,

volume 132 of *Proceedings of Machine Learning Research*, pages 1207–1213. PMLR, 16–19 Mar 2021.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, page 813–822, New York, NY, USA, 2011. Association for Computing Machinery.

Kamalika Chaudhuri and Daniel Hsu. Convergence rates for differentially private statistical estimation. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1715–1722, Madison, WI, USA, 2012. Omnipress.

Marco Avella-Medina. Privacy-preserving parametric inference: A case for robust statistics. *Journal of the American Statistical Association*, 0(0):1–15, 2020.

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

N. Wang, X. Xiao, Y. Yang, J. Zhao, S. C. Hui, H. Shin, J. Shin, and G. Yu. Collecting and analyzing multidimensional data with local differential privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649, 2019.

L. P. Barnes, W. N. Chen, and A. Özgür. Fisher information under local differential privacy. *IEEE Journal on Selected Areas in Information Theory*, 1(3): 645–659, 2020.

Heysem Kaya, Pinar Tüfekci, and Erdinç Uzun. Predicting co and nox emissions from gas turbines: novel data and a benchmark pems. *Turkish Journal of Electrical Engineering & Computer Sciences*, 27(6):4783–4796, 2019.

Margareta Ciglic, Johann Eder, and Christian Koncilia. Anonymization of data sets with null values. In *Special Issue on Database- and Expert-Systems Applications on Transactions on Large-Scale Data- and Knowledge-Centered Systems XXIV - Volume 9510*, page 193–220, Berlin, Heidelberg, 2016. Springer-Verlag. ISBN 9783662492130. doi: 10.1007/978-3-662-49214-7_7. URL https://doi.org/10.1007/978-3-662-49214-7_7.

Takao Murakami and Yusuke Kawamoto. Utility-optimized local differential privacy mechanisms for distribution estimation. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 1877–1894, USA, 2019. USENIX Association. ISBN 9781939133069.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018. doi: 10.1080/01621459.2017.1389735.

John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of Machine Learning Research*, volume 99, pages 1161–1191, Phoenix, USA, 25–28 Jun 2019a. PMLR.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf.

Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018. doi: 10.1109/TIT.2018.2809790.

H. Sun, B. Dong, H. Wang, T. Yu, and Z. Qin. Truth inference on sparse crowdsourcing data with local differential privacy. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 488–497, 2018.

Lin Sun, Xiaojun Ye, Jun Zhao, Chenhui Lu, and Mengmeng Yang. [this paper will appears in dasfaa2020 proceedings] bisample: Bidirectional sampling for handling missing data with local differential privacy. *arXiv preprint arXiv:2002.05624*, 2020.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data, Third Edition*. John Wiley & Sons, April 2019. doi: 10.1002/9781119482260.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/3b5020bb891119b9f5130f1fea9bd773-Paper.pdf.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled renyi differential privacy and analytical moments accountant. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1226–1235. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/wang19b.html.

Michael Carl Tschantz, Dilsun Kaynar, and Anupam Datta. Formal verification of differential privacy for interactive systems (extended abstract). *Electronic Notes in Theoretical Computer Science*, 276:61–79, 2011. ISSN 1571-0661. doi: https://doi.org/10.1016/j.entcs.2011.09.015. URL https://www.sciencedirect.com/science/article/pii/S157106611100106X. Twenty-seventh Conference on the Mathematical Foundations of Programming Semantics (MFPS XXVII).

Danfeng Zhang and Daniel Kifer. Lightdp: Towards automating differential privacy proofs. In *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages*, POPL 2017, page 888–901, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346603. doi: 10.1145/3009837.3009884. URL https://doi.org/10.1145/3009837.3009884.

Maxim Raginsky. Strong data processing inequalities and ϕ -sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016. doi: 10.1109/TIT.2016.2549542.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309): 63–69, 1965. doi: 10.1080/01621459.1965.10480775. PMID: 12261830.

Donald B Rubin. Statistical disclosure limitation. *Journal of official Statistics*, 9(2):461–468, 1993.

M. J. Elliot and J Domingo Ferrer. The future of statistical disclosure control. Paper published as part of The National Statistician's Quality Review., 2018. URL https://gss.civilservice.gov.uk/wp-content/uploads/2018/12/12-12-18_FINAL_Mark_Elliot_Josep_Domingo-Ferrer.pdf.

Satkartar K. Kinney, Jerome P. Reiter, Arnold P. Reznek, Javier Miranda, Ron S. Jarmin, and John M. Abowd. Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review*, 79(3):362–384, 2011. doi: https://doi.org/10.1111/j.1751-5823.2011.00153.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2011.00153.x.

Satkartar K. Kinney, Jerome P. Reiter, and Javier Miranda. Synlbd 2.0: Improving the synthetic longitudinal business database. *Statistical Journal of the IAOS*, 30:129–135, 2014. doi: 10.3233/SJI-140808.

Jörg Drechsler, Stefan Bender, and Susanne Rässler. Comparing fully and partially synthetic data sets for statistical disclosure control in the german iab establishment panel : supporting paper für die work session on data confdentiality 2007 in manchester. In *EUNECE / Programmes*, page 6, 2007. URL https://fis.uni-bamberg.de/handle/uniba/19712.

Jörg Drechsler, Agnes Dundler, Stefan Bender, Susanne Rässler, and Thomas Zwick. A new approach for disclosure control in the iab establishment panel—multiple imputation for a better data access. *AStA Advances in Statistical Analysis*, 92(4):439–458, 2008. doi: 10.1007/s10182-008-0090-1. URL https://doi.org/10.1007/s10182-008-0090-1.

Beata Nowok. synthpop: An r package for generating synthetic versions of sensitive microdata for statistical disclosure control. Technical report, Technical report, Administrative Data Research Centre, University of Edinburgh, 2016. URL https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_24_bnowok_synthpop.pdf.

Beata Nowok, Gillian M. Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. *Journal of Statistical Software*, 74(11):1–26, 2016. doi: 10.18637/jss.v074.i11. URL https://www.jstatsoft.org/index.php/jss/article/view/v074i11.

Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(3):663–688,

2018. doi: https://doi.org/10.1111/rssa.12358. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12358.

Mi-Ja Woo, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), Apr. 2009. doi: 10.29012/jpc.v1i1.568. URL https://journalprivacyconfidentiality.org/index.php/jpc/article/view/568.

Jerome P. Reiter. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):185–205, 2005a. doi: https://doi.org/10.1111/j.1467-985X.2004.00343.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-985X.2004.00343.x.

Trivellore E Raghunathan, Jerome P Reiter, and Donald B Rubin. Multiple imputation for statistical disclosure limitation. *Journal of official statistics*, 19(1):1, 2003.

Yuchen Zhang, John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 2328–2336, Red Hook, NY, USA, 2013. Curran Associates Inc.

Mark Braverman, Ankit Garg, Tengyu Ma, Huy L. Nguyen, and David P. Woodruff. *Communication Lower Bounds for Statistical Estimation Problems via a Distributed Data Processing Inequality*, page 1011–1020. Association for Computing Machinery, New York, NY, USA, 2016. ISBN 9781450341325. URL https://doi.org/10.1145/2897518.2897582.

John Duchi and Ryan Rogers. Lower bounds for locally private estimation via communication complexity. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1161–1191. PMLR, 25–28 Jun 2019b. URL https://proceedings.mlr.press/v99/duchi19a.html.

Jörg Drechsler and Jerome P. Reiter. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55(12):3232–3243, 2011. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2011.06.006. URL https://www.sciencedirect.com/science/article/pii/S0167947311002076.

Jerome P Reiter. Using cart to generate partially synthetic public use microdata. *Journal of official statistics*, 21(3):441, 2005b.

Gregory Caiola and Jerome P. Reiter. Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy*, 3(1):27–42, apr 2010. ISSN 1888-5063.

Jörg Drechsler. Using support vector machines for generating synthetic datasets. In Josep Domingo-Ferrer and Emmanouil Magkos, editors, *Privacy in Statistical Databases*, pages 148–161, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15838-4.

Yingrui Chen, Mark Elliot, and Joseph Sakshaug. A genetic algorithm approach to synthetic data production. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*, PrAISe '16, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343046. doi: 10.1145/2970030.2970034. URL https://doi.org/10.1145/2970030.2970034.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. Ctab-gan: Effective table data synthesizing. In Vineeth N. Balasubramanian and Ivor Tsang, editors, *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pages 97–112. PMLR, 17–19 Nov 2021. URL https://proceedings.mlr.press/v157/zhao21a.html.

Halbert White. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433, 1981. doi: 10.1080/01621459.1981.10477663. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1981.10477663.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.