Tuning-free Estimation and Inference of Cumulative Distribution Function under Local Differential Privacy

Yi Liu [1] **Qirui Hu** [2] **Linglong Kong** [1]

Abstract

We introduce a novel algorithm for estimating Cumulative Distribution Function (CDF) values under Local Differential Privacy (LDP) by exploiting an unexpected connection between LDP and the current status problem, a classical survival data problem in statistics. This connection leads to the development of tools for constrained isotonic estimation based on binary queries. Through mathematical proofs and extensive numerical testing, we demonstrate that our method achieves uniform and L_2 error bounds when estimating the entire CDF curve. By employing increasingly dense grids, the error bound can be improved, exhibiting an asymptotic normal distribution of the proposed estimator. Theoretically, we show that the error bound smoothly changes as the number of grids increases relative to the sample size n. Computationally, we demonstrate that our constrained isotonic estimator can be efficiently computed deterministically, eliminating the need for hyperparameters or random optimization.

1. Introduction

The last few years have seen unprecedented advancements in data science, transforming how we interact with technology and data. While this proliferation of advanced data-driven technologies has significantly altered the landscape of information processing and analysis, the ability to exploit seemingly harmless data has been increasingly demonstrated. Studies on mobile device sensor data, for instance, have revealed that driving patterns and locations can be identified (Hua et al., 2016), and spoken words can be reconstructed from accelerometer data (Zhang et al., 2015; Anand et al., 2019). Alarmingly, even a user's internet activity can be inferred from a wireless charger (Liu et al., 2022).

These developments, while technologically impressive, bring forth privacy concerns that extend beyond conventional threats such as data breaches or hacker attacks. The vulnerability of individual data, far from being inconsequential, has become increasingly apparent. One significant source of privacy risk originates from the legitimate uses of published data, models, and their outputs. Incidents like the Netflix challenge and the AOL search log database (Narayanan & Shmatikov, 2006; 2008; Barbaro & T. Zeller, 2006), where users were re-identified from anonymized datasets, exemplify this type of threat. Additionally, recent findings reveal that engineered prompts can specifically extract training data from large language models (Nasr et al., 2023), introducing a novel aspect to these privacy concerns.

Concurrently, the more direct threat posed by malpractices of data curators remains a pressing issue. A notable instance of this occurred in March 2023, when an error in ChatGPT inadvertently allowed users to view other users' personal information, leading to a temporary service halt. Another significant breach at a healthcare provider in Washington DC compromised sensitive data of federal legislators, underscoring the long-standing data security challenges in the US healthcare industry (Lee, 2022). These incidents underscore the continuous challenges faced by data curators, as discussed in (Ayyagari, 2012; Quach et al., 2022).

The pressing need to shield users from emerging privacy threats has catalyzed the development of privacy-preserving statistical techniques. Differential Privacy (Dwork et al., 2006), a prominent approach in this domain, is renowned for providing robust privacy safeguards against the first type of threat, enabling meaningful data analysis while maintaining user privacy. However, it falls short in scenarios where the data curator is compromised. In contrast, Local Differential Privacy (Duchi et al., 2013) addresses this very concern. By elevating privacy standards, LDP operates under the assumption that users cannot trust the curator, thus offering a more stringent level of privacy protection.

These frameworks, LDP and DP, have reshaped the land-

[*]Equal contribution [1]Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada [2]Center for Statistical Science, Department of Industrial Engineering, Tsinghua University, Beijing, China. Correspondence to: Linglong Kong <lkong@ualberta.ca>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

scape of statistics and machine learning. They pose a challenge to the development of new estimations and inferences that adhere to these heightened privacy standards. While some basic problems can be effectively addressed using generic mechanisms like the Laplace mechanism, particularly in scenarios with trusted curators, more complex issues require intricate solutions. For example, the problem of sample quantile estimation under DP was initially tackled in (Smith, 2011), subsequently extended to multiple quantile estimation in (Gillenwater et al., 2021), and to regression in (Chen & Chua, 2023). However, the LDP setting, where curators are not trusted, presents greater challenges. It wasn't until 2022 that an optimal algorithm for mean estimation under LDP was proposed (Asi et al., 2022). For quantile estimation in this setting, while (Duchi et al., 2013) initially suggested a median estimation algorithm, a more versatile approach applicable to all quantiles, along with a method for calculating confidence intervals using self-normalization techniques, was later introduced in (Liu et al., 2023).

This paper aims to extend these efforts in the LDP context, moving from single quantile estimation to multiple and potentially all quantiles. Such advancements are poised to eventually lead to a comprehensive estimation of the CDF, marking a significant stride in privacy-preserving data analysis.

1.1. Related work

Estimation of CDF is a fundamental task in statistical analysis. The most commonly used approach is based on the empirical cumulative distribution function (ECDF) of observed dataset and one builds the corresponding asymptotic properties via emprical process theory in different model setting, for example, see (Komlós et al., 1975; Lahiri et al., 1999; Dehling & Philipp, 2002; Shorack & Wellner, 2009). To overcome the discontinuity of ECDF, another method utilizes kerner regression, see (Liu & Yang, 2008), or other nonparametrc smoothing technique, see (Xue & Wang, 2010).

Under the context of DP, the problem is more complicated. There have been several attempts to address these issues, which can be traced back to the development of the Frequency Oracle (FO) (Erlingsson et al., 2014; Bassily & Smith, 2015; Acharya et al., 2019). These mechanisms were primarily designed for discrete domains. By applying discretization to continuous domains, these methods can be used for estimating continuous distributions, but at the cost of losing some information about the continuous structure. Later, Li et al. (Li et al., 2020) improved the FO algorithm by computing an MLE using the Expectation-Maximization algorithm through a square wave mechanism. However, such approach still relies on discretization (bucketing), and the quality of the output is highly sensitive to the stopping criteria hyperparameter. The finite number of bucketing not only prevents the estimator from being asymptotically consistent but also introduces additional challenges due to the extra parameter.

1.2. Outline

We begin by presenting a brief review of the relevant definitions and background related to CDF estimation and LDP followed by our data collection procedure, which employs a series of random-response randomizers to transform sensitive individual information into a private view of binary variables. We highlight that this data collection process results in a private view that resembles the structure of the current status problem, a well-studied issue in survival analysis. Subsequently, we construct an estimator based on the private view obtained in the previous step. Interestingly, the LDP treatment and statistical analysis technique can be disentangled by considering an alternative view of the collected data, where the randomized response can be treated as a truthful response originating from an alternative variable. We then refine the naive MLE method by imposing monotonic and bound constraints on the estimator and demonstrate that such an estimator can be computed in a fast, deterministic, and hyperparameter-free manner. Following this, we investigate the asymptotic properties of the proposed estimator, providing a comprehensive analysis of its performance under various conditions. Under different sampling strategies, introduced below, we first establish L_2 and uniform consistency up to order $\mathcal{O}_p(n^{-1/3})$ and $\mathcal{O}_p(n^{-1/3}\log n)$ respectively. Then, we derive the pointwise weak convergence results of the proposed estimator. In addition, the theoretical justifications show that the convergence rate varies continuously between these two samplings from $\mathcal{O}_p(n^{-1/3})$ to $\mathcal{O}_p(n^{-1/2})$. Especially, for the estimation on finite grids, we obtain the asymptotic normality for whole design points with a diagonal asymptotic covariance matrix, which can be applied for constructing confidence intervals and hypothesis testing. Lastly, we demonstrate the effectiveness of our proposed protocol through numerical experiments, showcasing its practical utility and accuracy in CDF estimation under the LDP framework.

1.3. Notations

In this paper, we employ the following notations. $\mathbf{1}_{\{\cdot\}}$ is the indicator function and $[a]$ denotes the largest integer that does not exceed a. $\mathcal{O}_p$ (or o_p) denotes a sequence of random variables of a certain order in probability. For instance, $o_p(n^{-1/2})$ means a smaller order than $n^{-1/2}$. For sequences a_n and b_n, denote $a_n \asymp b_n$ if there exist positive constants c and C such that $cb_n \le a_n \le Cb_n$. The symbol $\xrightarrow{d}$ means weak convergence or converge in distribution.

2. Preliminaries

2.1. Central Differential Privacy

The foundational concept of Central Differential Privacy (CDP) is based on the observation that the overall pool of knowledge remains predominantly unchanged when a single user's data is excluded from the dataset. This exclusion of an individual's data guarantees that their personal privacy is no longer compromised by the dataset. A key characteristic of CDP is that the distribution of outputs should be similar when comparing two datasets that differ only by the data of one individual. Different interpretations of 'similarity' between distributions give rise to various forms of Differential Privacy. To maintain the focus of our paper and avoid ambiguity in defining distribution similarity, we adopt the most widely recognized standard of pure DP, defined as follows:

Definition 2.1. (Dwork et al., 2006) A randomized algorithm $\mathcal{A}$, taking a dataset consisting of individuals as its input, is (ϵ, δ)-differentially private if, for any pair of datasets S and S' that differ in the record of a single individual and any event E, satisfies the below condition:

$$\mathbb{P}[\mathcal{A}(S) \in E] \leq e^{\epsilon}\mathbb{P}\left[\mathcal{A}\left(S'\right) \in E\right] + \delta.$$

When $\delta = 0$, $\mathcal{A}$ is called ϵ-differentially private (ϵ-DP).

CDP focuses on constraining the output distribution rather than the credibility of the entities running them. The curator's role in simplifying algorithm design often results in a minimal loss of accuracy due to privacy protection, as described by Cai et al. (2021) (Cai et al., 2021).

2.2. Local Differential Privacy

LDP is conceptualized from a more cautious standpoint. In the LDP framework, there is no reliance on a trustworthy curator. Instead, the curator's role is limited to coordinating interactions among users, each possessing their own private information X_i. During each interaction round, the curator selects a user and assigns them a randomizer, R_t, which adheres to the definition below:

Definition 2.2. (Joseph et al., 2019) An (ϵ, δ)-randomizer $R : X \to Y$ is an (ϵ, δ)-differentially private function taking a single data point as input.

Participants then verify whether the given (ϵ, δ) parameters align with the experimental setup. If they meet the criteria, each user then applies the randomizer to their private data and shares the outcome with the curator. The degree of interaction in this process can range from fully adaptive (the least stringent level) to sequential (or one-shot) interaction, and finally, to non-interactive (the most stringent). In our research, we adhere to the strictest model of non-adaptivity, prohibiting adaptiveness and requiring the determination of all user-randomizer pairings prior to any data collection. For an in-depth discussion on adaptivity, readers are directed to Definitions 2.3 and 2.6 in Cheu et al. (2019) (Cheu et al., 2019). Contrasting with the CDP framework, where the curator deliberately adds noise to the output to fulfill the DP condition, in the LDP setting, the curator's objective is to construct estimations based on the data that has undergone randomization by the users.

2.3. The current status problem

Current status data emerges in studies where the primary measurement is the occurrence time of a specific event, but observations are confined to indicators that reveal if the event has transpired at the time of data collection. This type of data is particularly relevant in survival analysis, such as in research investigating the survival of patients with cancer during an observation period, which is highly related to isotonic regression (see (Durot & Lopuhaä, 2018)). In these cases, researchers may passively acquire the patient's status through hospital visits (alive) or loss of contact (presumably dead). However, the exact time of death is unobtainable, especially if it lies in the future. Refer to (Jewell & van der Laan, 1995; Rossini & Tsiatis, 1996; Aarts et al., 2000; Wang et al., 2008; Sal y Rosas & Hughes, 2011) for more information about the research of this type.

3. Methodology

3.1. Problem formulation

Let $X = \{X_1, \ldots, X_n\}$ be independently and identically distributed random variables defined on $[0, 1]$ representing the private information of each user. The goal is to estimate the underlying CDF of X_i (F) with inquiries to each user while conforming to the ϵ-LDP condition.

3.2. The LDP data collection

In contrast to the CDP setting, where user data is openly gathered for analysis, the development of an LDP protocol commences with the data collection process. This is due to the ϵ-LDP constraint, which presents a significant challenge for estimation problems. Initially, without the DP constraint, the CDF can be intuitively approximated by the ECDF, yielding a convergence rate of $\mathcal{O}_p(n^{-1/2})$. Nevertheless, in the LDP context, each data point's contribution is considerably restricted.

To illustrate this point, consider the canonical Laplace mechanism, which serves as the standard DP mechanism for bounded continuous variables. The noise variance (2.0) needed to achieve LDP with $\epsilon = 1$ is eight times larger than the highest possible variance (0.25) of the $[0, 1]$ bounded

variable. In addition, reconstructing the original distribution from the Laplace noise perturbed data will lead to a notoriously hard deconvolution problem (Fan, 1991) with terrible sample efficiency (Fan, 1992). This stringent condition compels us to constrain the inquiries directed toward end users, thereby reducing the scale of DP noise. To this end, we generate T_i from another distribution G and collect binary responses from users of the question below:

Compare T_i and your private number X_n: is T_i greater than or equal to X_n?

If the users are asked to answer this question truthfully, they still faces a serious privacy concern. However, we can ask the user to perturb the answer locally, leading to the following randomizer:

Definition 3.1. Random response randomizer:

$$\mathcal{E}_i(X_i) = \begin{cases} \mathbf{1}_{X_i \le T_i}, & w.p.\ \text{r}, \\ \text{Bernoulli}(0.5) & w.p.\ \text{1-r}. \end{cases} \tag{1}$$

As a special case of randomized response, $\mathcal{E}_i$ is a $(\epsilon, 0)$ randomizer for $\epsilon = \log((1+r)/(1-r))$ (Dwork et al., 2014). For certain values of r, the mechanism can be executed physically using a coin or dice. It is worth noting that in the definition provided, the value of T_i is generated by the curator and distributed to users. This is due to concerns that end-users may lack the knowledge or equipment to accurately produce the necessary randomness. However, if the entire process is automated on digital devices, the generation of T_i can be shifted to the user side. This approach reduces communication costs and offers additional privacy advantages since the data point be no longer trace by the assigned T_i. For the remainder of this paper, we will represent the privacy budget using $r = \tanh(\epsilon/2)$, as it affords a more intuitive interpretation and a simpler form in our results (refer to Table 2 for a conversion table).

Following the definitions above, the curator can collect a private data view of X namely $\{(\Delta_1, T_1), \ldots, (\Delta_i, T_i)\}$, where $\Delta_i = \mathcal{E}_i(X_i)$. By the post-processing property, any function of the private view or even the view itself can be safely released without violation of the ϵ-LDP condition.

Compared to the current status problem, where control over T_i is limited, the T_i employed in our LDP mechanism can be fully tailored. Initially, by either sending i.i.d. T_i to users or requesting users to generate T independently, it is clear that T_i is independent of all X_i and all other T_j for $j \neq i$. This scenario is atypical in medical studies concerning current status. In practice, patients' visits often correlate with their own status and even the status of others (for instance, the weather may affect hospital visits, leading to correlated T_i. In another example, when observing patients' lifespans, since future deaths cannot be observed, the censoring is related to the current time and patients' birth data).

The ability to freely design G provides a significant advantage. Firstly, the independence between T_i and X_i simplifies the analysis of estimation. The known G also eliminates the need for estimation and the potential errors that may arise from it. However, the most crucial aspect we can design is the control of G to generate better estimations in areas of interest. Intuitively, the estimation of F will be more accurate when G samples more. Here, we introduce two types of sampling methods:

Density-based sampling: In this type of sampling, we let G correspond to a density g that is uniformly bounded away from 0, ensuring that every open set within the domain can be sampled with a non-zero probability. This approach provides a more comprehensive representation of the underlying distribution. The simplest G in this form is the CDF of the uniform distribution, given by $G(x) = x$. Such a choice will lead to uniform sampling. Other reasonable choices include weighted sampling, where G is selected to be denser in regions of particular interest, or if we possess prior knowledge about F, we can choose $G \approx F$ to achieve improved estimation outcomes.

Preselected sampling: In this sampling method, we let G be a discrete distribution on the interval $[0, 1]$. This means we pre-select a subset of values $\{x_1, \ldots, x_\kappa\} \subset [0, 1]$ and define $G(x) = \sum_{i=1}^{\kappa} p_i 1_{x > x_i}$. Preselected sampling focuses the estimation on the chosen nodes, making it particularly useful when there are specific, exact x values of interest. For instance, when estimating income distribution, we might be especially interested in the proportion of people below the poverty line or above a certain income threshold. Alternatively, we might be interested in an evenly distributed grid of X_i to provide a plausible plot of the CDF curve.

3.2.1. Additional privacy guarantee from shuffling

The shuffle model (Balle et al., 2019) is distinguished by a shuffling step that follows the LDP perturbation applied by each user, after which the datapoints can't be traced back to specific users, resulting in a significant privacy amplification effect (Cheu et al., 2019). According to theorem 2 and corollary 2 (Chen et al., 2024), a shuffled ϵ-LDP process can be approximated as $\frac{2e^{\epsilon/2}}{\sqrt{n-1}}$-GDP and $(\lambda, \frac{2e^{\epsilon}\lambda}{n-1})$-RDP, for any $\lambda \geq 2$.

The results from the shuffling model offer a central privacy assessment ($\frac{2e^{\epsilon/2}}{\sqrt{n-1}}$-GDP and $(\lambda, \frac{2e^{\epsilon}\lambda}{n-1})$-RDP) for our estimators under two scenarios. The first involves a straightforward application of the shuffle model, assuming the existence of a trustworthy shuffler to eliminate order information from the private data. The second scenario is more subtle; we

will later see that the estimator and its algorithmic implementation are order-invariant. If the algorithm is executed faithfully and the private data remains secure, the distribution of the output will resemble that of private data processed through a shuffler. Consequently, it will exhibit the same CDP property as in the previous case.

3.3. The constrained isotonic estimator

The protocol in the last chapter provided an LDP view of the data. In this chapter, we construct an estimator from the LDP private view. Define:

$$X_i^\star = \begin{cases} X_i, & w.p.\ \text{r}, \\ \text{Bernoulli}(0.5) & w.p.\ \text{1-r}. \end{cases} \tag{2}$$

The CDF of $X_i^\star$ can be derived from F as below:

$$F^\star(x) = \left(rF(x) + \frac{1-r}{2}\right)\mathbf{1}_{0<x<1} + \mathbf{1}_{x=1}. \tag{3}$$

The introduction of the notations $F^\star$ and $X^\star$ brings to light a pivotal observation that paves the way to the solution. We define $\Delta_i^\star = \mathbf{1}_{X_i^\star \le T_i}$. Notice that:

$$\mathbb{P}(\Delta_i^\star = 1 | X_i, T_i) = \mathbb{P}(\Delta_i = 1 | X_i, T_i).$$

Consequently, rather than interpreting the observed Δ_i as a noise-afflicted response of the indicator function $\mathbf{1}_{X_i \le T_i}$, it is feasible to regard Δ_i as accurate responses of $\mathbf{1}_{X_i^\star \le T_i}$, derived from an alternative variable, $X_i^\star$. This method significantly reduces the conflict between local differential privacy treatments and statistical analysis techniques. As a result, the challenge shifts from estimating highly noisy observations to accurately estimating responses from a different distribution. With the help of the notations of $F^\star$, the log-likelihood can be expressed as

$$\begin{aligned} L(F, \Delta, T) :&= \sum_{i=1}^{n} \Delta_i^\star \log F^\star(T_i) \\ &\quad + (1 - \Delta_i^\star)\log(1 - F^\star(T_i)) \\ &= \sum_{i=1}^{n} \Delta_i \log\left(rF(T_i) + \frac{1-r}{2}\right) \\ &\quad + (1-\Delta_i)\log\left(\frac{1+r}{2} - rF(T_i)\right). \end{aligned} \tag{4}$$

It may seem appealing to determine $F^\star$ and F through the naive maximization of the log-likelihood function. Nonetheless, several issues arise from this approach. For instance, the maximizing function $F^\star$ may not necessarily represent a CDF as there is no assurance that the estimation will exhibit monotonic behavior. Additionally, the error associated with $F^\star$ is likely to be significantly high around the values of 0 and 1 due to the lack of relevant samples. To address this issue, we define D as the function family of all non-decreasing functions mapping from $[0,1]$ to $[0,1]$, and propose our constrained isotonic estimator as follows:

$$\widehat{F} \in \arg\max_{\widehat{F} \in D} L(F, \Delta, T).$$

We remark that the $\widehat{F}$ satisfying the definition is not unique. The right-hand side represents an equivalence class, wherein two distribution functions are considered equivalent if and only if they agree on all T_i. Consequently, the maximization process can be performed solely on the nodes T_i. The remaining part of the function can be arbitrarily monotonically interpolated. For instance, the function values can be determined by the nearest T_i to the left or right, or they can be linearly interpolated. Regardless of the interpolation technique, the properties presented in the following chapter remain applicable. In the numerical experiments, the function values are filled using the nearest T_i to the left, resulting in a left-continuous staircase function to avoid unfair advantages. It's also worth noting that $\widehat{F}$ is an order-invariant M-estimator, enabling the Central DP accounting discussed in Section 3.2.1.

3.4. Algorithm

The disentanglement between the LDP treatment and the data analysis allows our to make use of the nonparametric likelihood estimation algorithm in the survival analysis (Huang & Wellner, 1997) with minor tweaking. A detailed description of the full algorithm is presented in Algorithm 1. The initial four steps adhere to the standard procedure of isotonic regression as delineated by Runlong (Tang et al., 2012a). In the intermediate stage, the estimation $\widehat{F}^\star(x)$ optimizes the log-likelihood function, disregarding the constraint that $\widehat{F}(x) \in [0,1]$ (or equivalently $\widehat{F}^\star(x) \in [(1-r)/2, (1+r)/2]$). Notably, although the range constraint is not taken into account during the evaluation of $\widehat{F}^\star(x)$, it is effectively satisfied in step 6 through a clipping procedure. This process not only ensures compliance with the range constraint but also maintains optimality under this constraint (refer to Appendix 6 for a detailed proof). The Greatest Convex Minorant can be deterministically computed, eliminating the requirement for iterative optimization, as demonstrated by Robertson et al. (Robertson, 1988) and referenced in (Klaus & Strimmer., 2015). This results in an overall deterministic algorithm free of hyperparameters for the analysis. In addition, the Algorithm 1 demonstrates excellent performance. For $n \le 10^7$, it took less than 1s to execute on a single core of an AMD Threadripper PRO 3995WX CPU. For comprehensive details regarding computation times, refer to Table 7.

Algorithm 1 Constrained isotonic estimation

1: Compute the function $H_1(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{T_i \le x\}}$, $H_2(x) = \frac{1}{n}\sum_{i=1}^{n} \Delta_i \mathbf{1}_{\{T_i \le x\}}$.
2: Plot $M = (H_1(x), H_2(x))$, for $x \in [0,1]$.
3: Compute Greatest Convex Minorant of M as Z.
4: Compute $\widehat{F}^{\star}(x) =$ left-derivative of Z at $H_1(x)$ for $x \in [0,1]$.
5: Invert the linear relationship by $\widetilde{F} = r^{-1}\left(\widehat{F}^{\star}(x) - \frac{1-r}{2}\right)$.
6: Give $\widehat{F}(x) = 0 \bigvee (1 \bigwedge \widetilde{F}(x))$.

4. Asymtotic properties

In the following section, we will explore the mathematical underpinnings of the data processing algorithm proposed in the previous section. To facilitate our analysis, we will start by introducing consistency results under density-based sampling:

Theorem 4.1. *Consistency under density-based sampling:*

(i) L_2 consistency: If there exists $g(x) > 0$ that are the corresponding density function of $G(x)$, one has that

$$\left\|\widehat{F} - F\right\|_2 = \mathcal{O}_p(r^{-1}n^{-1/3}),$$

where $\|h\|_2^2 = \int_0^1 h^2(x)dx$.

(ii) Uniform consistency: With further assumptions that there exists $f(x) > 0$ that are the corresponding density function of $F(x)$, one has that

$$\sup_{x\in[0,1]} \left|\widehat{F}(x) - F(x)\right| = \mathcal{O}_p(r^{-1}n^{-1/3}\log n).$$

From Theorem 4.1, one finds that L_2 consistency requires the existence of the density function of $G(x)$ only, and the convergence rate will be slightly faster than uniform consistency. The asymptotic results are influenced by the truthful response rate r, where a smaller r necessitates a larger sample size to achieve the same level of finite sample performance as r^{-1} acts as a multiplicative factor in the L_2 and L_∞ norms of the error. In the context of differential privacy, this factor r^{-1} can alternatively be expressed as $\coth(\epsilon/2)$ or $(1+e^{\epsilon})/(1-e^{\epsilon})$. Next, we will discuss the point-wise asymptotic distribution of the proposed estimator.

Theorem 4.2. *Under the assumption in Theorem 4.1, for any $x_0 \in (0,1)$, one obtains that*

$$\frac{(r^2 g(x_0)n)^{1/3}(\widehat{F}(x_0) - F(x_0))}{\left\{4\left(rF(x_0) + \frac{1-r}{2}\right)\left(\frac{1+r}{2} - rF(x_0)\right)f(x_0)\right\}^{1/3}} \Rightarrow \mathcal{Z} := \arg\max_{t\in\mathbb{R}}\left\{W(t) - t^2\right\},$$

where $W(t)$ is standard two-sided Brownian motion, and $\mathcal{Z}$ is referred as Chernoff distribution.

The proposed CDF estimator exhibits significantly different asymptotic behaviour compared to non-DP cases. While the convergence rate of ECDF is up to order $\mathcal{O}_p(n^{-1/2})$, the convergence rate of the proposed estimator is much slower. Furthermore, we note that the sequence of stochastic processes $\{n^{1/3}(\widehat{F}(x) - F(x)), x \in [0,1]\}$ is not tight in $D[0,1]$ leading to a problematic topological structure caused by infinite dimensionality (see (Huang & Wellner, 1997)). Such underlying difficulty limits us to point-wise asymptotic distribution, as opposed to the weak convergence results of $\widehat{F}$ and related goodness-of-fit statistics, such as KS statistics. However, in most practical scenarios, we are interested in estimating pre-design points on the CDF rather than the entire curve, due to computational accuracy constraints.

To achieve this, we assume that the observation times T_i are i.i.d. random variables sampled from a discrete probability measure G_n supported on $[0,1]$. We denote the support of G_n by $\{x_{i,n}\}_{i=1}^{\kappa_n}$, where the ith grid point is given by $x_{i,n} = in^{-\gamma}$, $i = 1, \dots \kappa_n = [n^{\gamma}]$, and $\gamma \in (0,1]$. We view the distribution G_n as a discretization of an absolutely continuous distribution G', with $G_n\{x_{i,n}\} = G'(x_{i,n}) - G'(x_{i-1,n})$ for $i = 2, 3, \dots, \kappa_n - 1$, $G_n\{x_{1,n}\} = G'(x_{1,n})$, and $G_n\{x_{\kappa_n,n}\} = 1 - G'(x_{\kappa_n-1,n})$, which will allow us to unify the theoretical framework of density-based sampling and preselected sampling mentioned above while obtaining the relationship between their convergence rates. Our focus is on estimating F at a grid point. To this end, we choose a grid point with respect to a fixed time $x_0 \in (0,1)$ that does not depend on n and can be viewed as an anchor point. We define x_l as the largest grid point less than or equal to x_0, and x_r as the first grid point to the right of x_l.

Theorem 4.3. *Consistency and asymptotic distribution under preselected sampling:*

Under the assumptions in Theorem 4.1, and $f(x) \in C[0,1]$, for any $x_0 \in (0,1)$, if $\gamma \in (0, 1/3)$, as $n \to \infty$, one has that

$$n^{1/2-\gamma/2}\left(\widehat{F}(x_l) - F(x_l), \widehat{F}(x_r) - F(x_r)\right) \Rightarrow \sqrt{\frac{\left(rF(x_0) + \frac{1-r}{2}\right)\left(\frac{1+r}{2} - rF(x_0)\right)}{r^2 g'(x_0)}} N(0, I_2),$$

where $g'(x_0)$ is the density of G' on x_0.

Specially, if $\kappa_n = \kappa < \infty$, then for increasing sequence

$\{x_j\}_{j=1}^{\kappa}$, *on has that*

$$\sqrt{n}\left(\widehat{F}(x_j) - F(x_j)\right)_{j=1}^{\kappa} \Rightarrow N\left(0, \operatorname{diag}\left\{\frac{\left(rF(x_l) + \frac{1-r}{2}\right)\left(\frac{1+r}{2} - rF(x_l)\right)}{r^2(G'(x_j) - G'(x_{j-1}))}\right\}_{j=1}^{\kappa}\right). \tag{5}$$

It is noteworthy that the asymptotic results in Theorem 4.3 get rid of all nuisance parameters and can be evaluated directly. Theorem 4.3 establishes the connection between density-based sampling and preselected sampling. The convergence rate of the estimator $\widehat{F}(x)$ is determined by the density of grids (relative to sample size n), and the first statement in the theorem describes how the convergence rate varies continuously from $\mathcal{O}_p(n^{-1/3})$ to $\mathcal{O}_p(n^{-1/2})$ when γ varies from $1/3$ to 0. When $\gamma \geq 1/3$, the convergence rate is still no slower than $\mathcal{O}_p(n^{-1/3})$, but the asymptotic distribution will be more complex, which is neither normal distribution in Theorem 4.3 or Chernoff distribution defined in Theorem 4.2. The second statement mainly focuses on estimating $F(x)$ on finite grids. The convergence rate approaches the order of $\mathcal{O}_p(n^{-1/2})$, as same as the convergence rate of ECDF in non-DP cases, and the asymptotic normality holds for the whole sequence $\{x_j\}_{j=1}^{\kappa}$. In the finite grids case, the observation grids are not necessarily uniform like infinite ones, which will be more fixable in practice. Also, the number of grids will not be increasing as the sample size in many scenarios, which will fall into the discussion of the second assertion. The covariance matrix in (5) is a diagonal matrix, which implies that the estimators $\{\widehat{F}(x_j)\}_{j=1}^{\kappa}$ are asymptotically independent. This may seem counter-intuitive, but (Groeneboom, 1984) studied the local dependence structure of this type of process in a closely related problem, and one can construct i.i.d. random variables with the distribution of the estimators $\{\widehat{F}(x_j)\}_{j=1}^{\kappa}$ (see the Appendix), which simplifies the results. Therefore, we can conduct statistical inference on $\{F(x_j)\}_{j=1}^{\kappa}$ based on (5), such as constructing confidence intervals and hypothesis testing for $F(x)$ on the grids $\{x_j\}_{j=1}^{\kappa}$.

5. Experiments

In this chapter, we assess the performance of our proposed algorithms by employing various probability distributions. The datasets are derived from four distinct cases: Uniform distribution $U(0,1)$, Truncated normal distribution $N_c(0,1,\mu,\sigma^2)$, and Continuous Bernoulli distribution $CB(\lambda)$.

For the Truncated normal distribution, the parameters are set as $\mu = 1/2$ and $\sigma^2 = 1/4$. This results in a distribution equivalent to $Y/2 + 1/2$, conditioned on the absolute value of Y being less than 1, where Y follows a standard normal distribution. In the case of the Continuous Bernoulli distribution, the parameter λ is selected to be $1/4$, which yields a non-symmetric density function. The density functions for the specified distributions are illustrated in Figure 2.

We consider the truthful response rate $r = 0.25, 0.5, 0.9$, which means the privacy budget is $\epsilon = \log(1 + 2r/(1-r))$ corresponding to $0.51, 1.09, 2.94$ respectively. These values indicate varying levels of privacy protection, ranging from strong to moderate. For comparison, Apple's implementation of differential privacy employs privacy budgets of 8 for QuickType and auto-play intent, 4 for emoji usage and crash reports in Safari, and 2 for highly sensitive health data (Apple, 2020; Tang et al., 2017).

The sample size ranges n spans from 10^3 to 10^7, with a total of $10,000$ replications (and reported means). To eliminate any correlation between experiments, the results from different sample sizes are independently conducted from scratch. Before delving into a more detailed presentation of the results, we first showcase a plot of our proposed estimator for the uniform distribution under density-based sampling. As depicted in Figure 3, our estimator, represented by the staircase functions, converges to the true CDF as n increases, resulting in diminishing absolute errors in the form of spikes.

5.1. Density based sampling performance

Next, we discuss the performance of our proposed estimator under density-based sampling. As for the sampling density, we consider two types of G. The first type is $G(x) = x$, which corresponds to uniform sampling. This is the preferred choice in situations where we do not have explicit preferences or knowledge about the distribution and domain. The second type of G is chosen as $G = F$. Although it is unlikely to occur in real practice, this represents the best possible case when we already have some prior knowledge about the distribution (as an extreme case of $G \approx F$). Results under the second type of G are marked with an additional $*$ and is given in the appendix. The improvement over uniform sampling is marginal. The table below presents the empirical results for both uniform consistency (represented by the maximum absolute error) and L_2 consistency (represented by the L_2 error) of the estimator.

As observed in the table, both the maximum absolute error and the L_2 error decrease as n increases and the privacy budget increase (larger r) as expected. The results for second type of G show a slight advantage over the first one, but the difference is negligible. This observation suggests that a uniform sample would be sufficient, and while sampling closer to the true distribution can be helpful, the improvement is only marginal. Therefore, we recommend using uniform sampling with the density-based approach, as it avoids the issues associated with acquiring prior knowledge.

Table 1. Empirical results of uniform consistency (L_2 consistency) of the proposed estimator under uniform sampling.

n	r	$U(0,1)$	$N_c(0,1,\mu,\sigma^2)$	$CB(\lambda)$
10^3	0.25	0.262(0.118)	0.289(0.116)	0.270(0.120)
	0.5	0.183(0.076)	0.199(0.074)	0.185(0.075)
	0.9	0.127(0.050)	0.137(0.047)	0.129(0.049)
10^4	0.25	0.143(0.057)	0.156(0.057)	0.147(0.057)
	0.5	0.096(0.036)	0.104(0.035)	0.100(0.036)
	0.9	0.065(0.023)	0.073(0.022)	0.067(0.022)
10^5	0.25	0.074(0.027)	0.081(0.027)	0.077(0.027)
	0.5	0.048(0.017)	0.054(0.017)	0.050(0.017)
	0.9	0.033(0.011)	0.037(0.010)	0.034(0.010)
10^6	0.25	0.038(0.013)	0.041(0.013)	0.039(0.013)
	0.5	0.024(0.008)	0.027(0.008)	0.025(0.008)
	0.9	0.016(0.005)	0.019(0.005)	0.017(0.005)
10^7	0.25	0.019(0.006)	0.021(0.006)	0.020(0.006)
	0.5	0.012(0.004)	0.013(0.004)	0.013(0.004)
	0.9	0.008(0.002)	0.009(0.002)	0.008(0.002)

We remark that the maximum errors for the type 2 groups are nearly identical for larger samples; this is because the effects of f and g cancel each other out in Theorem 4.2. To verify our claimed convergence rate we give a graphical illustration comparison between results from different sample sizes and the convergence rate and we showcase a term what we call standardized maximum absolute error (SMAE), which is defined as MAE multiplied by $rn^{1/3}/\log n$. Under Theorem 4.1, the SMAE should remain bounded as $n \to \infty$ and varying r. We show this in the plot of standardized and unstandardized maximum absolute error (Figure 4 in Appendix). The three bundles in the curve of SMAE representing the results from the same F tend to be similar (not r), suggesting that the effect of privacy budget r is also properly modelled the standardization factor $rn^{1/3}/\log n$. This supports our claim in Theorem 4.1.

5.2. Preselected sampling performance

Theorem 4.3 predicts a multivariate asymptotically normal distribution for the residual. To condense the results into interpretable numerical outcomes, we define the following standardized weighted L^2 error $WMSE(\widehat{F})$:

$$\sqrt{n}\sum_{j=1}^{\kappa}\frac{r^2\left(G'(x_j)-G'(x_{j-1})\right)\left(\widehat{F}\left(x_j\right)-F\left(x_j\right)\right)^2}{\left(rF\left(x_l\right)+(1-r)/2\right)\left((1+r)/2-rF\left(x_l\right)\right)}, \tag{6}$$

which takes the sum of each square error divided by their corresponding predicted variance. According to theorem 4.3, the $WMSE(\widehat{F})$ asymptotically follows a χ^2 distribution with a degree of freedom κ. We proceed to compare the empirical $WMSE(\widehat{F})$ with the theoretical $\chi^2(\kappa)$ distribution from two perspectives. First, we examine the relative χ^2 error (RCE), which we define as $WMSE(\widehat{F})/\kappa$. An RCE value greater than 1 indicates that the actual weighted error is larger than expected, and vice versa. Second, we consider the coverage rate, defined as $\mathbb{P}(WMSE(\widehat{F}) < \chi^2_{0.95,\kappa})$. If the distribution of the residuals aligns with expectations, the coverage rate should converge to 0.95. In the following, we present a plot illustrating the relative χ^2 error and coverage rate for the uniform distribution and sampling when $\kappa = 10$:

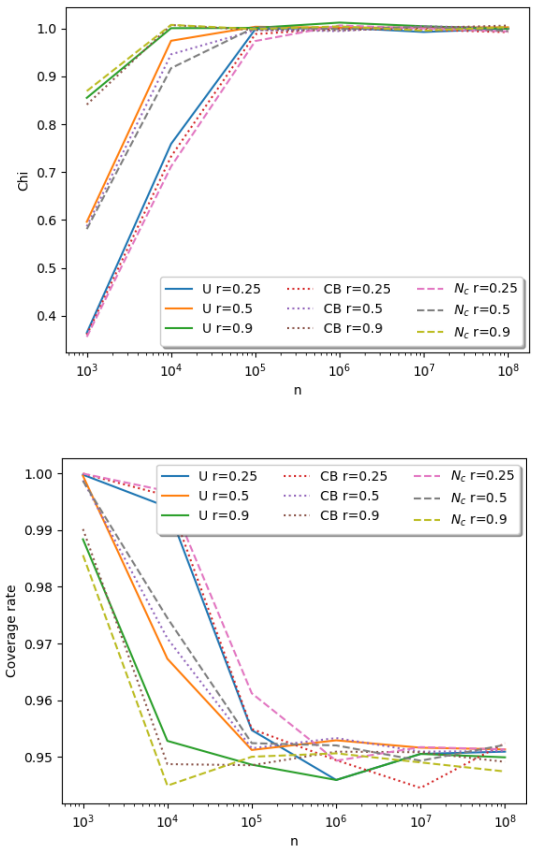

Figure 1. Left: The plot of relative χ^2 error compared to the true value Right: the plot of coverage rate

The results for large samples ($n \geq 10^5$) align with our claim in Theorem 4.3. Further, the small sample performance is actually better than our prediction. Consequently, our error bound proves to be numerically valid in this experiment, and for large samples, our asymptotic distribution permits statistical inference. For larger values of κ, the error bounds and asymptotic distribution remain valid, but a greater number of samples will be required to converge to the asymptotic results. We provide the RCE and coverage rates in tables located in the appendix.

6. Conclusions and future works

In this paper, we developed a data collection procedure and estimator for CDF estimation under the LDP framework, analyzed its asymptotic properties, and demonstrated its practical utility and accuracy through numerical experiments. Our work provides a comprehensive approach to CDF estimation while preserving privacy, offering valuable insights and applications for the field of privacy-preserving data analysis. Despite the contributions, there remain several intriguing unanswered questions, which pave the way for future research. Firstly, since the proposed estimator $\widehat{F}$ is non-differentiable, obtaining the density estimator directly becomes a challenge. Additionally, extending the estimation of the multivariate CDF to multivariate data poses further difficulties. Lastly, exploring the generation of bootstrap samples based on the proposed estimator for conducting further inference is an intriguing direction worth investigating.

Acknowledgements

We would like to thank the anonymous reviewers and area chairs for great feedback on the paper. Linglong Kong was partially supported by grants from the Canada CIFAR AI Chairs program, the Alberta Machine Intelligence Institute (AMII), the Natural Sciences and Engineering Council of Canada (NSERC), and the Canada Research Chair program from NSERC. Hu's research was supported partially by the National Natural Science Foundation of China award 12171269.

Impact Statement

The potential broader impact of this work is profound, especially in an era where data privacy concerns are paramount. Our algorithm forms a component of a reliable data analysis toolkit that adheres to LDP protocols, ensuring better protection of individuals' private information against unauthorized access and inference attacks. This is particularly crucial in sensitive domains such as healthcare, finance, and social networking, where the misuse of personal data can have severe consequences. Looking forward, the societal consequences of this work could be far-reaching. As LDP techniques become more sophisticated and widely adopted, we can expect a shift in how data is handled across industries, with a greater emphasis on privacy-by-design principles. This could lead to more robust data protection regulations and standards, enhancing public trust in digital systems.

References

Aarts, C., Kylberg, E., Hörnell, A., Hofvander, Y., Gebre-Medhin, M., and Greiner, T. How exclusive is exclusive breastfeeding? a comparison of data since birth with current status data. *International Journal of epidemiology*, 29(6):1041–1046, 2000.

Acharya, J., Sun, Z., and Zhang, H. Hadamard response: Estimating distributions privately, efficiently, and with little communication. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1120–1129. PMLR, 2019.

Anand, S. A., Wang, C., Liu, J., Saxena, N., and Chen, Y. Spearphone: A speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers. *arXiv preprint arXiv:1907.05972*, 2019.

Apple. Differential privacy overview - apple, 2020. URL https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

Asi, H., Feldman, V., and Talwar, K. Optimal algorithms for mean estimation under local differential privacy. In *International Conference on Machine Learning*, pp. 1046–1056. PMLR, 2022.

Ayyagari, R. An exploratory analysis of data breaches from 2005-2011: Trends and insights. *Journal of Information Privacy and Security*, 8(2):33–56, 2012.

Balle, B., Bell, J., Gascón, A., and Nissim, K. The privacy blanket of the shuffle model. In *Advances in Cryptology–CRYPTO 2019: 39th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 18–22, 2019, Proceedings, Part II 39*, pp. 638–667. Springer, 2019.

Barbaro, M. and T. Zeller, J. A face is exposed for aol searcher no. 4417749. *New York Times (Aug, 9, 2006)*, 2006.

Bassily, R. and Smith, A. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 127–135, 2015.

Cai, T. T., Wang, Y., and Zhang, L. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5): 2825–2850, 2021.

Chen, D. and Chua, G. A. Differentially private stochastic convex optimization under a quantile loss function. In *International Conference on Machine Learning*, pp. 4435–4461. PMLR, 2023.

Chen, E., Cao, Y., and Ge, Y. Renyi differential privacy in the shuffle model: Enhanced amplification bounds. *arXiv preprint arXiv:2401.04306*, 2024.

Cheu, A., Smith, A., Ullman, J., Zeber, D., and Zhilyaev, M. Distributed differential privacy via shuffling. In *Advances in Cryptology–EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part I 38*, pp. 375–403. Springer, 2019.

Dehling, H. and Philipp, W. *Empirical process techniques for dependent data*. Springer, 2002.

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

Durot, C. and Lopuhaä, H. P. Limit theory in monotone function estimation. *Statistical science*, 33(4):547–567, 2018.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer, 2006.

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

Erlingsson, Ú., Pihur, V., and Korolova, A. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.

Fan, J. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, pp. 1257–1272, 1991.

Fan, J. Deconvolution with supersmooth distributions. *Canadian Journal of Statistics*, 20(2):155–169, 1992.

Gillenwater, J., Joseph, M., and Kulesza, A. Differentially private quantiles. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3713–3722. PMLR, 18–24 Jul 2021.

Groeneboom, P. Estimating a monotone density. *Department of Mathematical Statistics*, (R 8403), 1984.

Groeneboom, P. and Wellner, J. A. *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media, 1992.

Hua, J., Shen, Z., and Zhong, S. We can track you if you take the metro: Tracking metro riders using accelerometers on smartphones. *IEEE Transactions on Information Forensics and Security*, 12(2):286–297, 2016.

Huang, J. and Wellner, J. A. Asymptotic normality of the npmle of linear functionals for interval censored data, case 1. *Statistica Neerlandica*, 49(2):153–163, 1995.

Huang, J. and Wellner, J. A. Interval censored survival data: a review of recent progress. In *Proceedings of the first Seattle symposium in biostatistics: survival analysis*, pp. 123–169. Springer, 1997.

Jewell, N. P. and van der Laan, M. Generalizations of current status data with applications. *Lifetime data analysis*, 1: 101–109, 1995.

Joseph, M., Mao, J., Neel, S., and Roth, A. The role of interactivity in local differential privacy. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 94–105, 2019. doi: 10.1109/FOCS.2019.00015.

Klaus, B. and Strimmer., K. *fdrtool: Estimation of (Local) False Discovery Rates and Higher Criticism*, 2015. URL https://CRAN.R-project.org/package=fdrtool. R package version 1.2.15.

Komlós, J., Major, P., and Tusnády, G. An approximation of partial sums of independent rv'-s, and the sample df. i. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 32:111–131, 1975.

Lahiri, S. N., Kaiser, M. S., Cressie, N., and Hsu, N.-J. Prediction of spatial cumulative distribution functions using subsampling. *Journal of the American Statistical Association*, 94(445):86–97, 1999.

Lee, I. An analysis of data breaches in the us healthcare industry: diversity, trends, and risk profiling. *Information Security Journal: A Global Perspective*, 31(3):346–358, 2022.

Li, Z., Wang, T., Lopuhaä-Zwakenberg, M., Li, N., and Škoric, B. Estimating numerical distributions under local differential privacy. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pp. 621–635, 2020.

Liu, J., Zou, X., Zhao, L., Tao, Y., Hu, S., Han, J., and Ren, K. Privacy leakage in wireless charging. *IEEE Transactions on Dependable and Secure Computing*, 2022.

Liu, R. and Yang, L. Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics*, 20(8):661–677, 2008.

Liu, Y., Hu, Q., Ding, L., and Kong, L. Online local differential private quantile inference via self-normalization. In *International Conference on Machine Learning*, pp. 21698–21714. PMLR, 2023.

Narayanan, A. and Shmatikov, V. How to break anonymity of the Netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

Narayanan, A. and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pp. 111–125. IEEE, 2008.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Quach, S., Thaichon, P., Martin, K. D., Weaven, S., and Palmatier, R. W. Digital technologies: Tensions in privacy and data. *Journal of the Academy of Marketing Science*, 50(6):1299–1323, 2022.

Robertson, T. Order restricted statistical inference. Technical report, 1988.

Rossini, A. and Tsiatis, A. A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, 91 (434):713–721, 1996.

Sal y Rosas, V. G. and Hughes, J. P. Nonparametric and semiparametric analysis of current status data subject to outcome misclassification. *Statistical Communications in Infectious Diseases*, 3(1), 2011.

Shorack, G. R. and Wellner, J. A. *Empirical processes with applications to statistics*. SIAM, 2009.

Smith, A. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 813–822, 2011.

Tang, J., Korolova, A., Bai, X., Wang, X., and Wang, X. Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*, 2017.

Tang, R., Banerjee, M., and Kosorok, M. R. Likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure. *The Annals of Statistics*, 40(1):45 – 72, 2012a. doi: 10.1214/11-AOS942. URL https://doi.org/10.1214/11-AOS942.

Tang, R., Banerjee, M., and Kosorok, M. R. Likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure. *The Annals of Statistics*, 40(1):45–72, 2012b.

Tang, R., Banerjee, M., and Kosorok, M. R. Supplement to "likelihood based inference for current status data on a grid: A boundary phenomenon and an adaptive inference procedure". 2012c.

Wang, L., Sun, J., and Tong, X. Efficient estimation for the proportional hazards model with bivariate current status data. *Lifetime Data Analysis*, 14:134–153, 2008.

Xue, L. and Wang, J. Distribution function estimation by constrained polynomial spline regression. *Journal of Nonparametric Statistics*, 22(4):443–457, 2010.

Zhang, L., Pathak, P. H., Wu, M., Zhao, Y., and Mohapatra, P. Accelword: Energy efficient hotword detection through accelerometer. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 301–315, 2015.

Appendix

Figures and tables

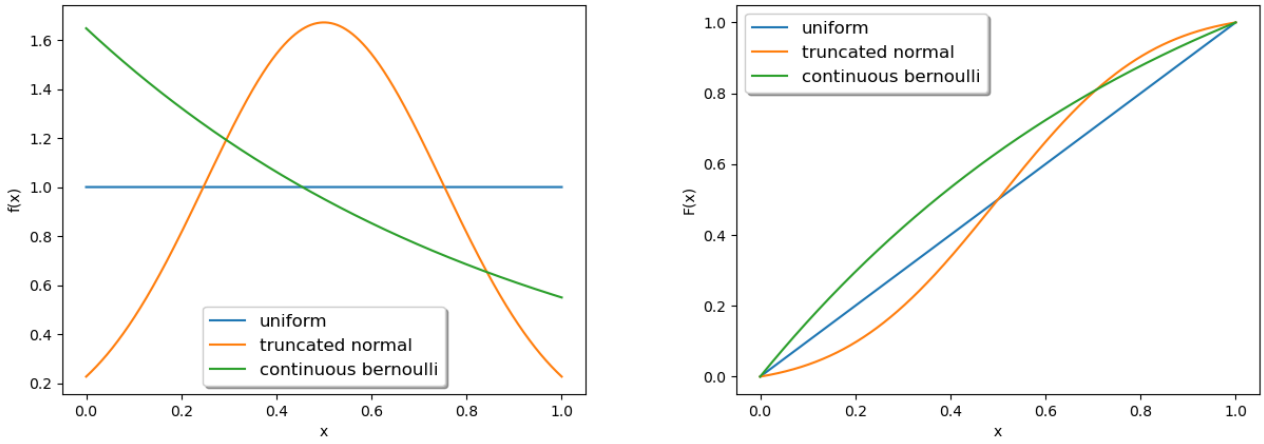

Figure 2. Left: Plot of CDF of the distributions, Right: Plot of PDF of the distributions

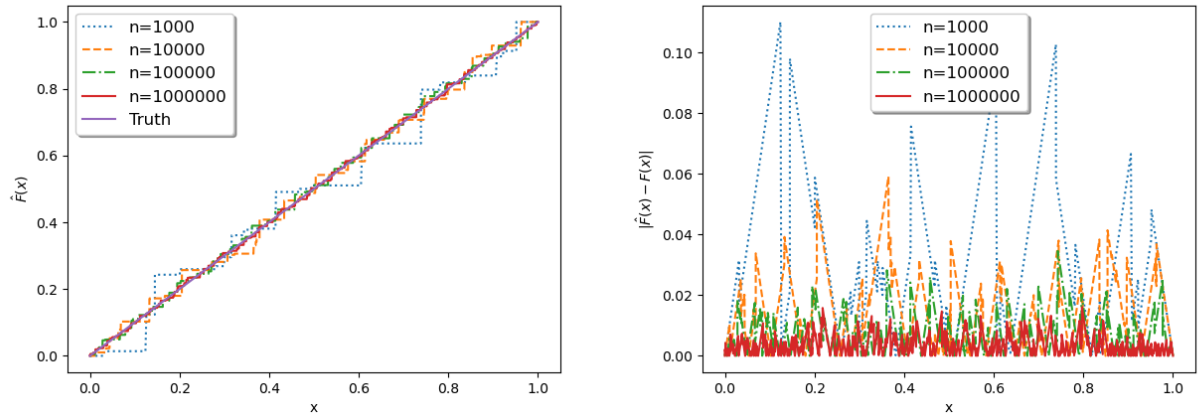

Figure 3. Left: The plot of the estimation and true value Right: The plot of absolute error

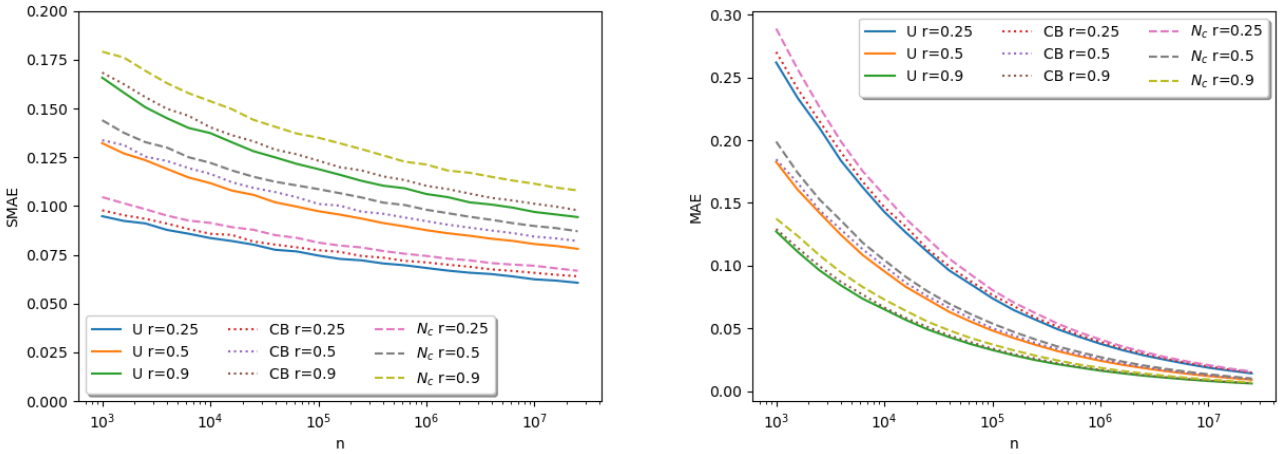

Figure 4. Plot of the maximum absolute error: standardized (left) and unstandardized(right)

Table 2. Conversion table between r and ϵ

r	ϵ	r	ϵ
0	0	0.5	1.10
0.05	0.10	0.55	1.24
0.1	0.20	0.6	1.39
0.15	0.30	0.65	1.55
0.2	0.40	0.7	1.73
0.25	0.51	0.75	1.95
0.3	0.62	0.8	2.20
0.35	0.73	0.85	2.51
0.4	0.85	0.9	2.94
0.45	0.97	0.95	3.66

Table 3. Empirical results of uniform consistency (L_2 consistency) of the proposed estimator under $G = F$.

n	r	$U(0,1)$	$N_c(0,1,\mu,\sigma^2)$	$CB(\lambda)$
10^3	0.25	0.262(0.118)	0.261(0.111)	0.265(0.116)
	0.5	0.183(0.076)	0.182(0.073)	0.183(0.076)
	0.9	0.127(0.050)	0.126(0.045)	0.127(0.049)
10^4	0.25	0.143(0.057)	0.142(0.055)	0.143(0.057)
	0.5	0.096(0.036)	0.096(0.035)	0.096(0.036)
	0.9	0.065(0.023)	0.065(0.021)	0.065(0.023)
10^5	0.25	0.074(0.027)	0.074(0.026)	0.074(0.027)
	0.5	0.048(0.017)	0.049(0.017)	0.049(0.017)
	0.9	0.033(0.011)	0.033(0.010)	0.033(0.010)
10^6	0.25	0.038(0.013)	0.038(0.013)	0.038(0.013)
	0.5	0.024(0.008)	0.024(0.008)	0.024(0.008)
	0.9	0.016(0.005)	0.016(0.005)	0.016(0.005)
10^7	0.25	0.019(0.006)	0.019(0.006)	0.019(0.006)
	0.5	0.012(0.004)	0.012(0.004)	0.012(0.004)
	0.9	0.008(0.002)	0.008(0.002)	0.008(0.002)

Table 4. Empirical coverage rate(RCE) of the proposed estimator with G is uniform distribution when $\kappa = 10$.

n	r	$U(0,1)$	$N_c(0,1,\mu,\sigma^2)$	$CB(\lambda)$
10^3	0.25	1.000(0.365)	1.000(0.355)	1.000(0.360)
	0.5	0.999(0.598)	0.999(0.578)	0.999(0.586)
	0.9	0.986(0.859)	0.985(0.865)	0.988(0.839)
10^4	0.25	0.995(0.757)	0.997(0.718)	0.996(0.731)
	0.5	0.964(0.974)	0.974(0.916)	0.973(0.941)
	0.9	0.948(0.996)	0.949(0.998)	0.949(0.999)
10^5	0.25	0.950(1.004)	0.959(0.968)	0.955(0.988)
	0.5	0.951(1.002)	0.950(0.991)	0.946(1.003)
	0.9	0.952(1.001)	0.951(0.992)	0.947(1.010)
10^6	0.25	0.947(1.005)	0.949(0.999)	0.948(1.005)
	0.5	0.951(1.005)	0.948(1.005)	0.951(1.000)
	0.9	0.954(0.995)	0.949(1.005)	0.953(1.005)
10^7	0.25	0.950(1.003)	0.956(1.001)	0.951(1.009)
	0.5	0.953(0.996)	0.950(1.000)	0.952(1.002)
	0.9	0.947(1.002)	0.948(1.004)	0.949(1.002)

Table 5. Empirical coverage rate(RCE) of the proposed estimator with G is uniform distribution when $\kappa = 20$.

n	r	$U(0,1)$	$N_c(0,1,\mu,\sigma^2)$	$CB(\lambda)$
10^3	0.25	1.000(0.184)	1.000(0.182)	1.000(0.183)
	0.5	1.000(0.314)	1.000(0.311)	1.000(0.313)
	0.9	1.000(0.522)	1.000(0.525)	1.000(0.510)
10^4	0.25	1.000(0.420)	1.000(0.403)	1.000(0.411)
	0.5	1.000(0.662)	1.000(0.631)	1.000(0.644)
	0.9	0.987(0.896)	0.984(0.900)	0.986(0.891)
10^5	0.25	0.996(0.814)	0.997(0.764)	0.996(0.792)
	0.5	0.963(0.985)	0.976(0.933)	0.966(0.975)
	0.9	0.949(0.997)	0.953(0.997)	0.952(0.996)
10^6	0.25	0.950(1.000)	0.962(0.974)	0.949(0.998)
	0.5	0.947(1.002)	0.951(1.005)	0.953(0.998)
	0.9	0.951(1.000)	0.949(1.002)	0.949(0.995)
10^7	0.25	0.949(0.997)	0.950(1.002)	0.951(1.000)
	0.5	0.950(1.003)	0.948(1.001)	0.950(1.000)
	0.9	0.947(1.004)	0.952(0.999)	0.954(0.996)

Table 6. Empirical coverage rate(RCE) of the proposed estimator with G is uniform distribution when $\kappa = 30$.

n	r	$U(0,1)$	$N_c(0,1,\mu,\sigma^2)$	$CB(\lambda)$
10^3	0.25	1.000(0.124)	1.000(0.122)	1.000(0.123)
	0.5	1.000(0.213)	1.000(0.210)	1.000(0.210)
	0.9	1.000(0.361)	1.000(0.359)	1.000(0.356)
10^4	0.25	1.000(0.284)	1.000(0.274)	1.000(0.278)
	0.5	1.000(0.461)	1.000(0.447)	1.000(0.454)
	0.9	1.000(0.713)	0.999(0.713)	1.000(0.695)
10^5	0.25	1.000(0.601)	1.000(0.571)	1.000(0.587)
	0.5	0.994(0.866)	0.997(0.808)	0.996(0.836)
	0.9	0.956(0.995)	0.963(0.980)	0.958(0.987)
10^6	0.25	0.971(0.968)	0.986(0.894)	0.978(0.938)
	0.5	0.950(1.003)	0.956(0.985)	0.952(0.997)
	0.9	0.953(0.996)	0.950(1.002)	0.950(0.999)
10^7	0.25	0.949(1.001)	0.952(0.996)	0.948(1.001)
	0.5	0.955(0.996)	0.953(0.998)	0.950(0.997)
	0.9	0.953(0.999)	0.953(0.999)	0.948(1.005)

Table 7. Computation times and standard derivation under different sample sizes

n	10^3	10^4	10^5	10^6	10^7	10^8
average time (ms)	0.081	0.345	4.011	42.28	527.5	5894
standard derivation	0.001	0.005	0.021	0.171	1.5899	40.12

Proof of Theorem 4.1

For the L_2 consistency, applying the Lemma 4.1 in (Huang & Wellner, 1995), one derives that

$$\int_0^1 (\sqrt{\widehat{F}^\star}(x) - \sqrt{F}(x))^2 dG(x) = \mathcal{O}_p(n^{-2/3}).$$

By the assumption of G, one has that

$$\int_0^1 (\sqrt{\widehat{F}^\star}(x) - \sqrt{F}(x))^2 dx = \mathcal{O}_p(n^{-2/3}).$$

Note that

$$\begin{aligned}\int_0^1 (\widehat{F}^\star(x) - F(x))^2 dx &= \int_0^1 (\sqrt{\widehat{F}^\star}(x) - \sqrt{F}(x))^2 (\sqrt{\widehat{F}^\star}(x) + \sqrt{F}(x))^2 dx \\ &< 4\int_0^1 (\sqrt{\widehat{F}^\star}(x) - \sqrt{F}(x))^2 dx = \mathcal{O}_p(n^{-2/3}).\end{aligned}$$

By the linear transformation between $(\widehat{F}(x), F(x))$ and $(\widehat{F}^\star(x), F^\star(x))$, the frist assertion of Theorem 4.2 holds

For the uniform consistency, we divided it into three parts, i.e.,

$$\begin{aligned}\sup_{x\in[0,1]} \left|\widehat{F}(x) - F(x)\right| &\le \sup_{x\in[0,2n^{-1/3}\log n]} \left|\widehat{F}(x) - F(x)\right| \\ &+ \sup_{x\in[2n^{-1/3}\log n, 1-2n^{-1/3}\log n]} \left|\widehat{F}(x) - F(x)\right| + \sup_{x\in[1-2n^{-1/3}\log n, 1]} \left|\widehat{F}(x) - F\right|.\end{aligned}$$

For any $x \in [2n^{-1/3}\log n, 1 - 2n^{-1/3}\log n]$, we first consider the estimator $\widehat{F}^*(x) \in [(1-r)/2, (1+r)/2]$. Since the CDF $F^\star(x)$ has a strictly positive density on $[x - n^{-1/3}\log n, x + n^{-1/3}\log n]$, one obtains that, for some positive constants c_1, c_2,

$$\mathbb{P}\left(\left|\widehat{F}^\star(x) - F^\star(x)\right| \ge n^{-1/3}\log n\right) \le c_1 \exp\{-c_2(\log n)^2\}$$

based on the Lemma 5.9 in (Groeneboom & Wellner, 1992). If there exists a sub-interval $\mathcal{I} \subset [2n^{-1/3}\log n, 1 - 2n^{-1/3}\log n]$ such that $\widehat{F}^*(x) \notin [(1-r)/2, (1+r)/2]$, the refinement of $\widehat{F}(x)$ does not lead to a worse convergence rate obviously. Hence, for any $r \in (0, 1]$

$$\mathbb{P}\left(\left|\widehat{F}(x) - F(x)\right| \ge r^{-1}n^{-1/3}\log n\right) \le c_1 \exp\{-c_2(\log n)^2\}$$

Now, for $x_i = in^{-1/3}\log n$, $i = 2, \ldots, [n^{1/3}\log n] - 1$, one has that

$$\mathbb{P}\left(\left|\widehat{F}(x_i) - F(x_i)\right| \ge r^{-1}n^{-1/3}\log n\right) \le c_1 \exp\{-c_2(\log n)^2\},$$

and

$$\mathbb{P}\left(\max_{2\le i\le [n^{1/3}\log n]-1} \left|\widehat{F}(x_i) - F(x_i)\right| \ge r^{-1}n^{-1/3}\log n\right) \le c_1 \exp\{-c_2(\log n)^2/2\}.$$

Due to the monotonic incrementality of $\widehat{F}(x)$ and $F(x)$, one obtains that

$$\mathbb{P}\left(\sup_{x\in[2n^{-1/3}\log n,1-2n^{-1/3}\log n]}\left|\widehat{F}(x)-F(x)\right|\geq r^{-1}n^{-1/3}\log n\right)\leq c_1\exp\{-c_2(\log n)^2/2\}.$$

For $x\in[0,2n^{-1/3}\log n]$, for the same arguments, one has that

$$\begin{aligned}&\sup_{x\in[0,2n^{-1/3}\log n]}\left|\widehat{F}(x)-F(x)\right|\leq\left|\widehat{F}(2n^{-1/3}\log n)-F(0)\right|\\&\leq\left|\widehat{F}(2n^{-1/3}\log n)-F(2n^{-1/3}\log n)\right|+\left|F(2n^{-1/3}\log n)-F(0)\right|=\mathcal{O}_p(r^{-1}n^{-1/3}\log n),\end{aligned}$$

for the reason that $F(x)$ has a strictly positive density on $[0,1]$. The left part holds for the same derivation, and the proof is complete.

Proof of Theorem 4.2

Under the assumption of Theorem 4.1, for any $x_0\in[0,1]$, such that $0<F(x_0),G(x_0)<1$, the CDF $F^\star(x)$ has positive density $f^\star(x_0)$. Then, following Theorem 5.1 in (Groeneboom & Wellner, 1992), one obtains that

$$\left\{\frac{g(x_0)n}{4F^\star(x_0)\left(1-F^\star(x_0)\right)f^\star(x_0)}\right\}^{1/3}\left(\widehat{F}^\star(x_0)-F^\star(x_0)\right)\Rightarrow\mathcal{Z}:=\underset{t\in\mathbb{R}}{\arg\max}\left\{W(t)-t^2\right\}.$$

Therefore, by the linear transformation between $(\widehat{F}(x),F(x))$ and $(\widehat{F}^\star(x),F^\star(x))$, the assertion of Theorem 4.2 holds and the proof is complete.

Proof of Theorem 4.3

The first assertion will be confirmed by the careful check of the proof Theorem 3.1 in (Tang et al., 2012b), and $F^\star(x_0)$ has positive density on $(x_0,x_0+n^{-\gamma})$, which satisfies the assumptions of Theorem 3.1 in (Tang et al., 2012b). Then, one obtains

$$n^{1/2-\gamma/2}\left(\widehat{F}^\star(x_l)-F^\star(x_l),\widehat{F}^\star(x_r)-F^\star(x_r)\right)\Rightarrow\sqrt{\frac{F^\star(x_0)\left(1-F^\star(x_0)\right)}{g(x_0)}}N\left(0,I_2\right).$$

Therefore, by the linear transformation between $(\widehat{F}(x),F(x))$ and $(\widehat{F}^\star(x),F^\star(x))$, the first assertion of Theorem holds.

If $\kappa_n=\kappa<\infty$, let $Z_l=\sum_{i=1}^n\Delta_i^\star\mathbf{1}_{T_i=t_l}$, $N_l=\sum_{i=1}^n$ and $\bar{Z}_l=Z_l/N_l$, $l=1,\ldots,\kappa$. Following proposition 3.4 in (Tang et al., 2012b), one has that, as $n\to\infty$,

$$\mathbb{P}\left(\bar{Z}_1\leq\cdots\leq\bar{Z}_\kappa\right)=1. \tag{7}$$

Given $\{N_l\}_{l=1}^K$, for each i draw an i.i.d. sample $\{Y_{lj}\}_{j=1}^{N_l}$ from Bernoulli $(1,F^\star(t_l))$. Denote $\bar{Y}_l=N_l^{-1}\sum_{j=1}^{N_l}Y_{lj}$, for each l. The second model is as follows. Suppose $\{t_l\}_{i=1}^\kappa,\{X_i^\star\}_{i=1}^n$ and $\{N_l\}_{i=1}^\kappa$ are defined as before. Let $\{Y'_{li}:1\leq l\leq\kappa,1\leq i\leq n\}$ be a family of mutually independent random variables, distributed independently of the variables in the previous sentence, such that for each i,Y'_{ij} follows Bernoulli $(1,F^\star(t_l))$ for $1\leq j\leq n$. Denote $\bar{Y}'_l=N_l^{-1}\sum_{j=1}^n Y'_{lj}\left\{X_j^\star=t_l\right\}$ for each l. Following Lemma 1.2 in (Tang et al., 2012c), one has that

$$\left(\{N_l\},\{\bar{Z}_l\}\right)\overset{d}{=}\left(\{N_l\},\{\bar{Y}_l\}\right)\overset{d}{=}\left(\{N_l\},\{\bar{Y}'_l\}\right).$$

Hence, combined with (7), we only need to prove the asymptotic properties of $\left(\{N_l\},\{\bar{Y}'_l\}\right)$. Then, by a triangular array version of the multivariate central limit theorem, it is sufficient to check the Lindeberg condition, and following the argument about Proof of Proposition S.1 in (Tang et al., 2012c), we will obtain the second assearations, after the linear transformation between $(\widehat{F}(x),F(x))$ and $(\widehat{F}^\star(x),F^\star(x))$.

Proof of the Algorithm 1

Firstly, $\widetilde{F}$ is the unconstrained maximizer of the log-likelihood function. If $\widetilde{F}(x) \in [0,1]$, then $\widehat{F} = \widetilde{F}$ is trivially the constrained maximizer, as it is the unconstrained maximizer that also satisfies the range constraint. If not, define x^- as $\inf x : \widetilde{F}(x) > 0$ and x^+ as $\sup x : \widetilde{F}(x) < 1$. Then $x^- > 0$ or $x^+ < 1$.

Suppose there is another function, $\widehat{F}_2$, such that $L(\widehat{F}_2, \Delta, T) > L(\widehat{F}, \Delta, T)$ and $\widehat{F}_2(x) \in [0,1]$. We then define a new function $\widehat{F}_3(x)$ as:

$$\widehat{F}_3(x) = (\mathbf{1}_{x<x^-} + \mathbf{1}_{x\geq x^+})\widetilde{F}(x) + \mathbf{1}_{x^-\leq x<x^+}\widehat{F}_2(x).$$

For simplicity, denote

$$L(F, \Delta, T, i) := \Delta_i \log\left(rF(T_i) + \frac{1-r}{2}\right) + (1-\Delta_i)\log\left(\frac{1+r}{2} - rF(T_i)\right).0$$

Now we compare $L(\widehat{F}_3, \Delta, T)$ and $L(\widetilde{F}, \Delta, T)$:

$$\begin{aligned}
L(\widehat{F}_3, \Delta, T) - L(\widetilde{F}, \Delta, T) &= \sum_{i=1}^{n}\Big[L(\widehat{F}_3, \Delta, T, i) - L(\widetilde{F}, \Delta, T, i)\Big] \\
&= \sum_{i=1}^{n}\mathbf{1}_{x^-\leq T_i<x^+}\Big[L(\widehat{F}_3, \Delta, T, i) - L(\widetilde{F}, \Delta, T, i)\Big] \\
&= \sum_{i=1}^{n}\mathbf{1}_{x^-\leq T_i<x^+}\Big[L(\widehat{F}_2, \Delta, T, i) - L(\widehat{F}, \Delta, T, i)\Big] \\
&\geq L(\widehat{F}_2, \Delta, T) - L(\widehat{F}, \Delta, T) \\
&> 0,
\end{aligned}$$

This implies that $\widehat{F}_3$ has a higher log-likelihood than $\widetilde{F}$, contradicting the assumption that $\widetilde{F}$ is the unconstrained maximizer. Therefore, the original claim holds.