

Differentially Private Regression for Discrete-Time Survival Analysis

THÔNG T. NGUYỄN, Nanyang Technological University

SIU CHEUNG HUI, Nanyang Technological University

In survival analysis, regression models are used to understand the effects of explanatory variables (e.g., age, sex, weight, etc.) to the survival probability. However, for sensitive survival data such as medical data, there are serious concerns about the privacy of individuals in the data set when medical data is used to fit the regression models. The closest work addressing such privacy concerns is the work on Cox regression which linearly projects the original data to a lower dimensional space. However, the weakness of this approach is that there is no formal privacy guarantee for such projection. In this work, we aim to propose solutions for the regression problem in survival analysis with the protection of differential privacy which is a golden standard of privacy protection in data privacy research. To this end, we extend the *Output Perturbation* and *Objective Perturbation* approaches which are originally proposed to protect differential privacy for the Empirical Risk Minimization (ERM) problems. In addition, we also propose a novel sampling approach based on the Markov Chain Monte Carlo (MCMC) method to practically guarantee differential privacy with better accuracy. We show that our proposed approaches achieve good accuracy as compared to the non-private results while guaranteeing differential privacy for individuals in the private data set.

CCS Concepts: •Mathematics of computing → Survival analysis; •Security and privacy → Privacy protections;

Additional Key Words and Phrases: survival analysis; discrete-time models; differential privacy; regression models

ACM Reference format:

Thông T. Nguyễn and Siu Cheung Hui. 2016. Differentially Private Regression for Discrete-Time Survival Analysis. 1, 1, Article 1 (January 2016), 19 pages.

DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Survival analysis studies and models probability of failure of time-related processes (e.g., time to death of HIV patients, time to divorce of married couples, time to graduation of Ph.D. students, etc.). Two important concepts in survival analysis are (1) the hazard rate function $h(t)$ which is the probability of failure (death) at time t , and (2) the survival function $S(t)$ which is the probability of survival to time t . An example of survival data set is the electronic health records (EHRs) which have been widely used and collected at large scale in modern hospitals (Blumenthal and Tavenner 2010; DesRoches et al. 2008; Jha et al. 2009). These health records are very useful for fitting the regression models to assist doctors in the medical decision processes for treatment, diagnosis, etc. In general, regression models are used to analyze the effects of explanatory variables (e.g., age, sex, weight, etc.) to the survival probability of patients. However, these models may also have serious problems of breaching patient's privacy as there is no guarantee that these models do not leak any personal information of individual patients in the data set.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2016 ACM. Manuscript submitted to ACM

In this work, we focus on the privacy problems of regression models used in survival analysis. We consider the setting in which privacy-preserving algorithms use data in the private data set to fit a survival regression model. The model is then published and available to the public for the benefits of society. Therefore, in this setting, the adversaries are assumed to know the output model, i.e., the parameters of the regression model. The goal is to design algorithms that can fit the survival regression model to the data set with high accuracy while guaranteeing that the adversaries cannot learn much information about the individuals in the data set when knowing the output model.

There are two different kinds of regression models in survival analysis, namely continuous-time models and discrete-time models. For continuous-time models, time is a continuous variable and failure events can happen at any moment. Cox regression is a well-known continuous-time model (Andersen and Gill 1982; Cox 1992) which allows estimation without any assumption on the baseline hazard effects. However, we have to assume the proportional hazard property (i.e., a unit increase in an explanatory variable will cause a multiplicative effect on the hazard rate). For discrete-time models (Allison 1982; Cox and Oakes 1984; Muthén and Masyn 2005), time is discrete and failure events only happen at discrete values of time. Discrete-time regression models are better than Cox regression when dealing with tied events (i.e., events which have the same value of survival time) and unobserved population heterogeneity (i.e., unobserved explanatory variables may cause bias to the estimation). Moreover, it does not need the proportional hazard property assumption as Cox regression does (Hess and Persson 2012).

In this paper, we propose solutions for the problem of guaranteeing discrete-time models not to leak personal information of the patients. Our proposed approaches guarantee differential privacy protection, which is the state-of-the-art privacy-preserving technique in data privacy research. Informally, a differentially private algorithm guarantees that two neighboring data sets which are different at only one patient's record are guaranteed to produce two outputs whose probability densities are very similar. This prevents an adversary from recognizing a data set from the collection of its neighbors. Therefore, an adversary cannot infer the personal information of a particular patient in the data set even in the case when the adversary knew all the information of all other patients in the data set (if otherwise, then the adversary can easily distinct two neighboring data sets).

In our solutions, we use the maximum likelihood estimation to transform the estimation problem to the optimization problem of choosing parameters to maximize the log-likelihood of the observed data set with respect to the discrete-time model. Coincidentally, our problem has a similar likelihood form as a logistic regression problem. This allows us to use the Output Perturbation (Out-Pert) and Objective Perturbation (Obj-Pert) proposed by Chaudhuri et al. (Chaudhuri et al. 2011) for our problem. These methods were originally proposed to protect differential privacy for the Empirical Risk Minimization (ERM) problems which include the logistic regression problem. The Out-Pert approach adds noise to the optimization solution to protect differential privacy. The Obj-Pert approach randomly perturbs the objective function, thereby ensuring the randomness of its optimization solution which can guarantee differential privacy for the solution. However, these approaches cannot be applied directly to our problem due to the difference in the loss function. Especially, this is due to the fact that our loss function is not a logistic loss function but a sum of logistic loss functions as the result of the discrete-time models. Therefore, we propose generalized extensions of the Out-Pert and Obj-Pert approaches to cater for our loss function.

A disadvantage of the above perturbation approaches is that for them to work properly they require a non-negligible regularization term in the objective function which incurs bias to the output model. To tackle this, we propose a sampling approach which protects differential privacy by directly sampling parameters from the objective function without the need of a regularization term to guarantee differential privacy. Similar ideas on sampling the objective

functions to provide differential privacy are also proposed in (Bassily et al. 2014; Kifer et al. 2012; Wang et al. 2015) for the ERM problems. However, it is required that the loss function has to have a finite maximum value. The previous works guarantee this property by boxing the output parameters in a finite-volume space (e.g., a sphere). This approach does not work well when the optimal parameter has a large magnitude. In this work, to guarantee the finite constraint, we wrap the loss function inside a sanitizer function (i.e., a scaled *tanh* function) to create a new finite loss function. We intentionally pick the sanitizer function that can keep the loss function in its original form when the value of the loss function is small. Meanwhile, the sanitizer function deforms the loss function at large values to make the function finite. The advantage of this approach is that the sampled parameter can arbitrary large while the objective function is kept almost the same around the optimal parameter which minimizes the objective function.

In order to sample an output parameter from the posterior distribution, Bassily et al. (Bassily et al. 2014) proposed a polynomial run-time algorithm to sampling the log-concave objective function but their algorithm is still impractical due to the high degree of its polynomial run-time complexity. On the other hand, Wang et al. (Wang et al. 2015) proposed to use a stochastic gradient Nosé-Hoover thermostat algorithm (Ding et al. 2014) to sample the posterior distribution. In this work, we propose to use Preconditioned Stochastic Gradient Langevin Dynamics (pSGLD) sampling algorithm (Li et al. 2015) to sample the objective function due to its advantages in sampling multi-dimensional parameters with different scales. It is worth to note that even though the sampling approach gives better accuracy (as we will see in Section 6), due to the property of its Markov chain, it cannot sample the objective function exactly. Therefore, the sampling approach does not mathematically guarantee differential privacy but only guarantees it approximately in practice.

In summary, the main contributions of this paper are as follows:

- We propose two privacy-preserving approaches, namely the Extended Output Perturbation and Extended Objective Perturbation, for the discrete-time survival regression problem. The proposed approaches guarantee differential privacy for the survival regression models. We formally prove these guarantees based on the definition of differential privacy.
- We propose a sampling approach to output a random model from its posterior distribution. The proposed sampling approach is based on pSGLD, which is a particular kind of the Markov Chain Monte Carlo (MCMC) method, to efficiently sample the random output which guarantees differential privacy approximately in practice.
- We show the effectiveness of our proposed approaches on four real survival data sets. In addition, we show that the results obtained from the discrete-time models are very close to the results obtained from Cox regression. We also show experimentally the convergence of our proposed sampling approach.

The rest of the paper is organized as follows: In Section 2, we review the related work on differential privacy and discrete-time survival analysis. Section 3 presents the regression models used in this work. Sections 4 discusses the proposed approaches of the Extended Output Perturbation and Extended Objective Perturbation along with their privacy guarantees. Section 5 discusses the proposed sampling approach. Section 6 presents the experimental results from real data sets. Finally, we conclude the paper in Section 7.

2 RELATED WORK

Even though it is important to protect privacy in medical data, as far as we know the work of Yu et al. (Yu et al. 2008) is the only work on privacy protection for Cox regression. Their work considers the setting in which Cox regression is

executed on a distributed data set over many institutions. They proposed to project patient's data to a lower dimensional space by a linear projection. The projection is satisfied by an optimization constraint to preserve good properties of the original data. However, their work is not based on a formal privacy definition such as differential privacy. Our work on discrete-time models for survival analysis is the first to propose a solution for the privacy problem of discrete-time survival models and also the first to apply differential privacy to survival analysis.

2.1 Differential Privacy

The state-of-the-art technique for the data privacy problem is *differential privacy* (Dwork 2009, 2011; Dwork et al. 2014). Basically, differential privacy is a promise to individuals in the data set that their information will not influence much on the final published results from the analysis. Differential privacy is used in many applications such as histogram publication (Li et al. 2010; Zhang et al. 2014), graph analysis (Borgs et al. 2015; Kasiviswanathan et al. 2013; Lu and Miklau 2014), regression and classification (Bassily et al. 2014; Chaudhuri and Monteleoni 2009; Kifer et al. 2012; Wang et al. 2015), recommender systems (Machanavajjhala et al. 2011; McSherry and Mironov 2009), etc. Here, we give a brief overview of differential privacy, interested readers can refer to (Dwork et al. 2014) for a detailed discussion on this subject.

To formalize the definition of differential privacy, we first need to introduce the definition of *two neighboring data sets*.

Definition 2.1 (Neighboring data sets). Two data sets \mathcal{D} and \mathcal{D}' are neighbors (denoted as $d(\mathcal{D}, \mathcal{D}') = 1$) if they agree in all except one record.

From that, we have a formal definition of differential privacy.

Definition 2.2 (Differential privacy). An algorithm \mathcal{A} is ϵ -differentially private if for any output value x of \mathcal{A} and for any pair of neighboring data sets \mathcal{D} and \mathcal{D}' :

$$\text{pdf}(\mathcal{A}(\mathcal{D}) = x) \leq \exp(\epsilon) \cdot \text{pdf}(\mathcal{A}(\mathcal{D}') = x)$$

where ϵ is the privacy budget of the algorithm \mathcal{A} .

2.2 Discrete-time Survival Analysis

For discrete-time models, let time be divided into intervals $[a_0, a_1), [a_1, a_2), \dots, [a_{q-1}, a_q]$, $a_0 = 0, a_q = 1$, where q is the number of discrete times. The discrete time t refers to the interval $[a_{t-1}, a_t)$. A discrete random variable T represents the discrete failure time. $T = t$ denotes the failure within the time interval $t = [a_{t-1}, a_t)$. The characteristic function of T is the discrete hazard function:

$$h(t) = \Pr(T = t \mid T \geq t), \quad t = 1, \dots, q$$

which is the conditional probability for the risk of failure in interval t given the survival in all previous intervals. The discrete survival function for reaching interval t is:

$$S(t) = \Pr(T \geq t) = \prod_{s=1}^{t-1} (1 - h(s)) \quad (1)$$

Discrete-time data sets are given by $(x_i, \delta_i, t_i), i = 1, \dots, n$, where $t_i = \min(T_i, c_i)$ is the minimum of the survival time T_i and censoring time c_i , and δ_i is the indicator variable for failure ($\delta_i = 1$) or censoring ($\delta_i = 0$). When $\delta_i = 0$, the i^{th} patient is known to survive until time c_i but the survival time T_i is not observed ($T_i > c_i$). x_i is a real vector of

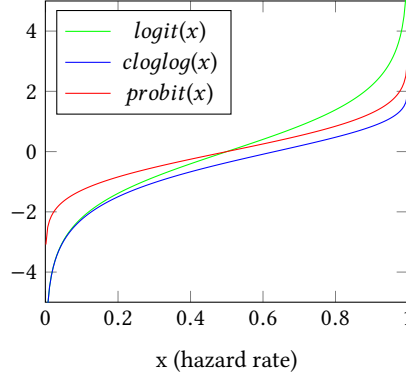


Fig. 1. The illustration of link functions: logit (green), cloglog (blue) and probit (red). We observe that these link functions are very similar in shape. Notably, the logit link function and cloglog link function are almost identical at x near 0. This explains why the output models from the logit link function and cloglog link function are very similar in practice when the number of discrete time q is large, or equivalently, the hazard rate x is small.

explanatory variables (e.g., sex, age, weight, etc.) which affect the survival probability. We assume that x_i is inside the unit-sphere, $\|x_i\| \leq 1$. This is actually a common practice in machine learning. Without loss of generality, we assume that $0 \leq t_i \leq 1$. For convenience, we use y_i to refer to the term $(2\delta_i - 1)$ and d_i to refer to the tuple (x_i, y_i, t_i) .

3 DISCRETE-TIME REGRESSION MODEL

In this section, we introduce the discrete-time regression models which are used to model the relationship between explanatory variables and the hazard rate, i.e., the predictive variable. From that, the subsequent sections will discuss the proposed differentially private approaches to guarantee that the estimated parameters from the regression model satisfy the definition of differential privacy.

3.1 Generalized Linear Models

We model the effects of explanatory variables x_i to the survival probability by using a generalized linear model:

$$g(h(t_i | x_i)) = \gamma(t_i) + x_i' \beta \quad (2)$$

where $g(\cdot)$ is the link function, β is the parameter vector representing the effects of explanatory variables and $\gamma(t_i)$ is a time-varying baseline hazard effect.

A commonly used link function in survival probability is the logit link function $g(x) = \text{logit}(x) = \log\left(\frac{x}{1-x}\right)$. The logit link function allows the model to have a nice interpretation of the proportional odds ratio. The other two link functions, which are also used in survival analysis, are the complementary log-log link function $g(x) = \text{cloglog}(x) = \log(-\log(1-x))$, and the probit link function $g(x) = \text{probit}(x) = \Phi^{-1}(x)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Interestingly, the complementary log-log link function has the same interpretation of proportional hazard ratio as the Cox regression. We refer interested readers to (Allison 1982) for more details.

As illustrated in Figure 1, the three link functions have similar shapes which lead to similar estimation results. In this work, we have selected the logit link function because it has a bounded derivative for the loss function which is

required by our proposed Extended Output Perturbation and Extended Objective Perturbation approaches. However, our proposed sampling approach can work with all three link functions.

3.2 Baseline Hazard Effect

We model the baseline hazard effect $\gamma(t)$ using natural cubic spline (Friedman et al. 2001) with e knots equally distributed over the interval $[0, 1]$, $0 = k_1 < k_2 < \dots < k_e = 1$.

Let

$$d_j(t) = \frac{\max(t - k_j, 0)^3 - \max(t - k_e, 0)^3}{k_e - k_j}$$

and

$$b_1(t) = 1, b_2(t) = t, b_{i+2} = d_i(t) - d_{e-1}(t)$$

The baseline hazard effect $\gamma(t)$ is approximated as a linear combination of e basis functions:

$$\gamma(t) = \alpha_1 b_1(t) + \dots + \alpha_e b_e(t)$$

In particular, let $A_i = [b_1(t_i), \dots, b_e(t_i)]'$ and $\alpha = [\alpha_1, \dots, \alpha_e]'$, then we can write $\gamma(t_i) = \alpha' A_i$.

3.3 Maximum Likelihood Estimation (MLE)

Traditionally, we use MLE to estimate parameters α and β in our models. The aim is to maximize the log-likelihood of the observed data. For simplicity, let $f = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ and $x_i^t = \begin{pmatrix} A_i \\ x_i \end{pmatrix}$. The log-likelihood function is:

$$\log \mathcal{L}(f) = \sum_{i=1}^n \log \left[h(t_i | f, x_i)^{\delta_i} (1 - h(t_i | f, x_i))^{1-\delta_i} S(t_i | f, x_i) \right]$$

Let $y_i = 2\delta_i - 1$, from (1), (2) and substituting $g(x) = \text{logit}(x)$, we can rewrite our problem as:

$$\log \mathcal{L}(f) = - \sum_{i=1}^n \left[\ell_{\text{LR}}(y_i f' x_i^{t_i}) + \sum_{s=1}^{t_i-1} \ell_{\text{LR}}(-f' x_i^s) \right]$$

where $\ell_{\text{LR}}(x) = \log(1 + \exp(-x))$ is the logistic loss function. To further simplify the formula, let $d_i = (x_i, y_i, t_i)$, $i = 1, \dots, n$, and let

$$\ell(f; d_i) = \ell_{\text{LR}}(y_i f' x_i^{t_i}) + \sum_{s=1}^{t_i-1} \ell_{\text{LR}}(-f' x_i^s) \quad (3)$$

be the loss function. Then, we get an ERM problem as follows:

$$f^* = \arg \min_f \sum_{i=1}^n \ell(f; d_i) \quad (4)$$

In this work, our main goal is to propose algorithms which protect differential privacy for f^* in Equation (4).

4 PERTURBATION APPROACHES

4.1 Extended Output Perturbation

In this section, we present our proposed algorithm which is the extension of the Output Perturbation approach in (Chaudhuri et al. 2011). For our problem, the loss function is a sum of logistic loss functions instead of a single logistic loss function as in (Chaudhuri et al. 2011). The proposed algorithm is in fact based on the generalized version of the

Algorithm 1 $\mathcal{A}_{\text{Ext-Out-Pert}}$: Extended Output Perturbation**Input:** Data set $\mathcal{D} = \{d_1, \dots, d_n\}$, loss function $\ell(f; d_i)$, privacy budget ϵ **Output:** f_{priv}

- 1: $J(f; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \ell(f; d_i) + \frac{\Lambda}{2} \|f\|^2$
- 2: Minimize $J(f; \mathcal{D})$ by using the BFGS algorithm to get the non-private solution f^*
- 3: Compute $t \leftarrow \frac{\sum_{s=1}^q \sqrt{4 + \|A_s\|^2 + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2 + 4}}}{n \cdot \Lambda}$
- 4: Sample a random vector b such that $\text{pdf}(b) \propto \exp\left(-\epsilon \frac{\|b\|}{t}\right)$
- 5: Compute and output $f_{\text{priv}} \leftarrow f^* + b$

Laplace mechanism (Dwork 2008) which is described as follows: Let $f^* = G(\mathcal{D})$ be the value that we want to guarantee differential privacy. f^* is the result of applying a function G on the private data set \mathcal{D} (e.g., it is in our case to minimize the objective function). We define the sensitivity of the function G as follows:

$$\text{sen}(G) = \max_{\mathcal{D}, \mathcal{D}'} \|G(\mathcal{D}) - G(\mathcal{D}')\|$$

where \mathcal{D} and \mathcal{D}' are two neighboring data sets. Then, the differentially private version of $f^* = G(x)$ is:

$$f_{\text{priv}} = f^* + \mu$$

where μ is a noisy random variable with probability density function $\text{pdf}(\mu) \propto \exp(-\epsilon \|\mu\| / \text{sen}(G))$.

As required by the Output Perturbation approach, we consider the following regularized objective function:

$$J(f; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \ell(f; d_i) + \frac{\Lambda}{2} \|f\|^2 \quad (5)$$

where $\mathcal{D} = \{d_i\}_{i=1}^n$, $\ell(\cdot)$ is the loss function as defined in (3) and Λ is the regularization parameter. In this approach, our goal is to compute the sensitivity of:

$$f^* = \arg \min_f J(f; \mathcal{D})$$

Then, we use the sensitivity to control the amount of noise added to f^* .

4.1.1 Proposed Algorithm. Algorithm 1 shows the proposed Extended Output Perturbation approach. It returns a vector f_{priv} as the minimizer of $J(\cdot)$ while guaranteeing differential privacy. At Line 2, we compute the non-private solution $f^* = \arg \min_f J(f; \mathcal{D})$ using the well-known BFGS algorithm (Fletcher 2013). f^* is guaranteed to exist due to the strongly convexity of $J(f; \mathcal{D})$. At Line 3, we compute t which is the sensitivity of f^* . Lines 4-5 add noise to the value of f^* .

In order to sample a random vector b in Algorithm 1 from the distribution $\text{pdf}(b) \propto \exp(-\epsilon \|b\| / t)$, we observe that the length of the vector b follows a Gamma distribution:

$$\|b\| \sim \Gamma(d, t/\epsilon)$$

where d is the number of components of b . Thus, in order to sample b we first sample its length $r = \|b\|$ from the Gamma distribution and then sample b as a uniform random point on the surface of a sphere with radius r .

4.1.2 Privacy Guarantee. In order to prove the differential privacy protection, we focus on proving that the sensitivity of f^* at Line 2 in Algorithm 1 is equal to the value of t which is computed at Line 3. Here, we use Lemma 4.1 from (Chaudhuri et al. 2011) to bound the sensitivity of f^* .

LEMMA 4.1. Let $G(f)$ and $g(f)$ be two vector-valued functions, which are continuous and differentiable at all points. In addition, let $G(f)$ and $G(f) + g(f)$ be λ -strongly convex. If $f_1 = \arg \min_f G(f)$ and $f_2 = \arg \min_f G(f) + g(f)$, then

$$\|f_1 - f_2\| \leq \frac{1}{\lambda} \max_f \|\nabla g(f)\|$$

From Lemma 4.1, our goal now is to bound the magnitude of the difference in the gradients of the objective function $J(\cdot)$ on any two neighboring data sets.

LEMMA 4.2. For any pair of patient's records $d_i = (x_i, y_i, t_i)$ and $d_j = (x_j, y_j, t_j)$, and for any f ,

$$\|\nabla \ell(f; d_i) - \nabla \ell(f; d_j)\| \leq \sum_{s=1}^q \sqrt{\|A_s\|^2 + 4} + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2 + 4}$$

PROOF.

$$\begin{aligned} \nabla \ell(f; d_i) &= \nabla \ell_{\text{LR}}(y_i f' x_i^{t_i}) + \sum_{s=1}^{t_i-1} \nabla \ell_{\text{LR}}(-f' x_i^s) \\ &= \frac{-y_i x_i^{t_i}}{1 + \exp(y_i f' x_i^{t_i})} + \sum_{s=1}^{t_i-1} \frac{x_i^s}{1 + \exp(-f' x_i^s)} \end{aligned}$$

Therefore, we can write $\nabla \ell(f; d_i) = \sum_{s=1}^q l_i^s$, where

$$l_i^s = \begin{cases} \frac{x_i^s}{1 + \exp(-f' x_i^s)}, & \text{if } s < t_i \\ \frac{-y_i x_i^{t_i}}{1 + \exp(y_i f' x_i^{t_i})}, & \text{if } s = t_i \\ \vec{0}, & \text{if } s > t_i \end{cases}$$

Similarly, we can also write $\nabla \ell(f; d_j) = \sum_{s=1}^q l_j^s$. Therefore,

$$\nabla \ell(f; d_i) - \nabla \ell(f; d_j) = \sum_{s=1}^q l_i^s - l_j^s$$

We have $|\frac{-y_i}{1 + \exp(y_i f' x_i^{t_i})}| \leq 1$, $\|x_i\| \leq 1$, $\|x_j\| \leq 1$, for any $s \in \{1 \dots q\}$, we consider four possible cases as follows:

Case 1: if $s < t_i$ and $s < t_j$, then

$$\begin{aligned} \|l_i^s - l_j^s\| &= \left\| \begin{pmatrix} (e_1 - e_2)A_s \\ e_1 x_i - e_2 x_j \end{pmatrix} \right\| \leq \sqrt{\|A_s\|^2 + (\|x_i\| + \|x_j\|)^2} \\ &\leq \sqrt{\|A_s\|^2 + 4} \end{aligned}$$

where $e_1 = \frac{1}{1 + \exp(-f' x_i^s)}$ and $e_2 = \frac{1}{1 + \exp(-f' x_j^s)}$.

Case 2: if $s > t_i$ or $s > t_j$, then $\|l_i^s - l_j^s\| \leq \max(\|x_i^s\|, \|x_j^s\|) \leq \sqrt{\|A_s\|^2 + 1} < \sqrt{\|A_s\|^2 + 4}$.

Case 3: if $l_i^s = \frac{-x_i^s}{1 + \exp(f' x_i^s)}$ and $l_j^s = \frac{x_j^s}{1 + \exp(-f' x_j^s)}$, then

$$\|l_i^s - l_j^s\| = \left\| - \begin{pmatrix} (e_1 + e_2)A_s \\ e_1 x_i + e_2 x_j \end{pmatrix} \right\| \leq \sqrt{\|2A_s\|^2 + 4}$$

where $e_1 = \frac{1}{1 + \exp(f' x_i^s)}$ and $e_2 = \frac{1}{1 + \exp(-f' x_j^s)}$.

Case 4: if $l_i^s = \frac{x_i^s}{1 + \exp(-f' x_i^s)}$ and $l_j^s = \frac{-x_j^s}{1 + \exp(f' x_j^s)}$, then $\|l_i^s - l_j^s\| \leq \sqrt{\|2A_s\|^2 + 4}$. This case is similar to **Case 3**.

We observe that there is at most one value of s , $1 \leq s \leq q$, belonging to **Case 3** or **Case 4** in which $\|l_i^s - l_j^s\| \leq \sqrt{\|2A_s\|^2 + 4}$. Therefore, from the triangle inequality:

$$\left\| \sum_{s=1}^q l_i^s - l_j^s \right\| \leq \sum_{s=1}^q \sqrt{\|A_s\|^2 + 4} + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2 + 4}$$

Therefore, the lemma follows. \square

Finally, we can bound the sensitivity of $f^* = \arg \min_f J(f; \mathcal{D})$ by the following lemma.

LEMMA 4.3. *The ℓ_2 -sensitivity of $f^* = \arg \min_f J(f; \mathcal{D})$ is at most $\frac{\sum_{s=1}^q \sqrt{4 + \|A_s\|^2} + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2 + 4}}{n\Lambda}$.*

PROOF. Without loss of generality, we assume that two neighboring data sets \mathcal{D} and \mathcal{D}' are different at n^{th} patient with $(x_n, y_n, t_n) \in \mathcal{D}$ and $(x'_n, y'_n, t'_n) \in \mathcal{D}'$.

Let $G(f) = J(f; \mathcal{D})$, $g(f) = J(f; \mathcal{D}') - J(f; \mathcal{D}) = \frac{1}{n}(\ell(f; d'_n) - \ell(f; d_n))$, $f_1 = \arg \min_f J(f; \mathcal{D})$, and $f_2 = \arg \min_f J(f; \mathcal{D}')$. Because $\frac{1}{2}\|f\|^2$ is 1-strongly convex, $G(f) = J(f; \mathcal{D})$ is Λ -strongly convex and $G(f) + g(f) = J(f; \mathcal{D}')$ is also Λ -strongly convex. From Lemma 4.2,

$$\begin{aligned} \|\nabla g(f)\| &= \left\| \frac{1}{n} (\nabla \ell(f; d'_n) - \nabla \ell(f; d_n)) \right\| \\ &\leq \frac{\sum_{s=1}^q \sqrt{4 + \|A_s\|^2} + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2 + 4}}{n} \end{aligned}$$

From Lemma 4.1,

$$\|f_1 - f_2\| \leq \frac{1}{\Lambda} \frac{\sum_{s=1}^q \sqrt{4 + \|A_s\|^2} + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2 + 4}}{n}$$

Therefore, the lemma follows. \square

THEOREM 4.1. *Algorithm 1 is ϵ -differentially private.*

PROOF. For any pair of neighboring data sets \mathcal{D} and \mathcal{D}' and for any f_{priv} ,

$$\frac{\text{pdf}(f_{priv} | \mathcal{D})}{\text{pdf}(f_{priv} | \mathcal{D}')} = \frac{\text{pdf}(b_1)}{\text{pdf}(b_2)} = \exp(-\epsilon/t(\|b_1\| - \|b_2\|))$$

where b_1 and b_2 are the corresponding noise vectors at Line 4 in Algorithm 1 with respect to the data sets \mathcal{D} and \mathcal{D}' . If f_1^* (resp., f_2^*) is the solution at Line 2 of Algorithm 1 on the data set \mathcal{D} (resp., \mathcal{D}'), then $f_1^* + b_1 = f_2^* + b_2 = f_{priv}$. From Lemma 4.3 and the triangle inequality:

$$\|b_1\| - \|b_2\| \leq \|b_1 - b_2\| = \|f_1 - f_2\| \leq t$$

where $t = \frac{\sum_{s=1}^q \sqrt{4 + \|A_s\|^2} + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2 + 4}}{n \cdot \Lambda}$. Therefore, $\frac{\text{pdf}(b_1)}{\text{pdf}(b_2)} \leq \exp(\epsilon)$. Thus, Algorithm 1 is ϵ -differentially private. \square

4.2 Extended Objective Perturbation

In this section, we present a solution based on the Objective Perturbation approach proposed in (Chaudhuri et al. 2011). Similarly to the Extended Objective Perturbation approach, we also consider the objective function as described in Equation (5). In this approach, instead of adding noise to the solution of the optimization problem as the output perturbation does, it adds noise to the objective function.

Algorithm 2 $\mathcal{A}_{\text{Ext-Obj-Pert}}$: Extended Objective Perturbation**Input:** Data set $\mathcal{D} = \{d_1, \dots, d_n\}$, objective function $J(f; \mathcal{D})$, privacy budget ϵ , parameter Λ **Output:** f^*

- 1: $\Delta \leftarrow 0$
- 2: Compute $\epsilon' \leftarrow \epsilon - 2 \sum_{s=1}^q \log \left(1 + \frac{\frac{1}{4} \sqrt{\|A_s\|^2 + 1}}{n(\Lambda + \Delta)} \right)$
- 3: **if** $\epsilon' < \epsilon/2$ **then**
- 4: Binary search value of Δ such that $2 \sum_{s=1}^q \log(1 + \frac{\frac{1}{4} \sqrt{\|A_s\|^2 + 1}}{n(\Lambda + \Delta)}) = \epsilon/2$ and set $\epsilon' \leftarrow \epsilon/2$
- 5: Compute $t \leftarrow \sum_{s=1}^q \sqrt{4 + \|A_s\|^2} + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2 + 4}$
- 6: Sample a random vector b such that $\text{pdf}(b) \propto \exp(-\epsilon' \|b\|/t)$
- 7: $f^* \leftarrow \arg \min_f J(f; \mathcal{D}) + \frac{1}{n} \langle b, f \rangle + \frac{1}{2} \Delta \|f\|^2$
- 8: Output f^*

4.2.1 Proposed Algorithm. Algorithm 2 shows the solution in pseudo-code. At Line 2, we compute ϵ' which is used to calibrate the magnitude of a random variable b . Here, the regularization parameter is equal to Λ . At Line 3, if $\epsilon' < \epsilon/2$, then it indicates that Λ is not large enough. In this case, an additional positive regularization parameter Δ is picked to set the value of ϵ' equals to $\epsilon/2$ (Line 4). At Line 5, we compute t which is the sensitivity of $\nabla J(f; \mathcal{D})$. Line 6 samples a random vector b using the same method described in Subsection 4.1.1. Lines 7-8 return the solution of the noisy objective function using the BFGS algorithm.

4.2.2 Privacy Guarantee. In this section, we will prove that the probability density of f^* from Algorithm 2 satisfies the differential privacy definition.

THEOREM 4.4. *Algorithm 2 is ϵ -differentially private.*

PROOF. The noisy objective function from Algorithm 2 is:

$$f^* = \arg \min_f \frac{1}{n} \sum_{i=1}^n \ell(f; d_i) + \frac{1}{n} \langle b, f \rangle + \frac{1}{2} (\Lambda + \Delta) \|f\|^2$$

Due to the convexity of $\ell(\cdot)$, the gradient is zero at the minimal point f^* , equivalently,

$$b = -n(\Lambda + \Delta)f^* - \sum_{i=1}^n \nabla \ell(f^*; d_i)$$

Due to the strongly convexity of the objective function, there is a bijective (injective and surjective) mapping from f to b (denoted as $f \rightarrow b$). Therefore, we can transform the probability density function of random variable f to the probability density function of random variable b by a multiplication factor of the Jacobian determinant (Billingsley 2008). From that, the probability density ratio in differential privacy can be rewritten as:

$$\frac{\text{pdf}(f \mid \mathcal{D})}{\text{pdf}(f \mid \mathcal{D}')} = \frac{\text{pdf}(b \mid \mathcal{D})}{\text{pdf}(b' \mid \mathcal{D}')} \cdot \frac{|\det(\text{Jacob}(f \rightarrow b \mid \mathcal{D}))|^{-1}}{|\det(\text{Jacob}(f \rightarrow b' \mid \mathcal{D}'))|^{-1}} \quad (6)$$

We first bound the ratio of the Jacobian determinants. Without loss of generality, we assume that the two data sets \mathcal{D} and \mathcal{D}' are different at n^{th} record with $d_n \in \mathcal{D}$ and $d_{n'} \in \mathcal{D}'$. Let

$$A = -\text{Jacob}(f \rightarrow b \mid \mathcal{D}) = n(\Lambda + \Delta)\mathbb{I} + \sum_{i=1}^n \nabla^2 \ell(f^*; d_i)$$

and $E = \nabla^2 \ell(f^*; d_n) - \nabla^2 \ell(f^*; d'_n)$, then

$$\frac{|\det(\text{Jacob}(f \rightarrow b \mid \mathcal{D}))|^{-1}}{|\det(\text{Jacob}(f \rightarrow b' \mid \mathcal{D}'))|^{-1}} = \frac{|\det(A + E)|}{|\det(A)|} = |\det(\mathbb{I} + A^{-1}E)|$$

Moreover, $E = \sum_{s=1}^q E_n^s - \sum_{s=1}^q E_{n'}^s$, where

$$E_n^s = \begin{cases} \frac{(x_n^s)(x_n^s)'}{(1+\exp(f'x_n^s))(1+\exp(-f'x_n^s))}, & \text{if } s < t_n \\ \frac{-y_n^2(x_n^s)(x_n^s)'}{(1+\exp(y_n f'x_n^s))(1+\exp(-y_n f'x_n^s))}, & \text{if } s = t_n \\ 0, & \text{if } s > t_n \end{cases}$$

Similarly, we can define $E_{n'}^s$ by replacing n by n' . From (Seiler and Simon 1975), for any square matrices A and B ,

$$\det(\mathbb{I} + A + B) \leq \det(\mathbb{I} + |A|) \cdot \det(\mathbb{I} + |B|)$$

where $|A| = (A'A)^{\frac{1}{2}}$. Moreover, $A^{-1}E_n^s$ and $A^{-1}E_{n'}^s$ are symmetric, thus

$$\det(\mathbb{I} + A^{-1}E) \leq \prod_{s=1}^q \det(\mathbb{I} + A^{-1}E_n^s) \cdot \det(\mathbb{I} + A^{-1}E_{n'}^s)$$

We now prove that $|\det(\mathbb{I} + A^{-1}E_n^s)| \leq 1 + \frac{\frac{1}{4}\sqrt{\|A_s\|^2+1}}{n(\Lambda+\Delta)}$. Because $\left| \frac{-y_n^2}{(1+\exp(y_n f'x_n^s))(1+\exp(-y_n f'x_n^s))} \right| \leq \frac{1}{4}$, and E_n^s is either a zero matrix or 1-rank matrix. The only non-zero eigenvalue of E_n^s if exist satisfies $|\lambda_1(E_n^s)| \leq \frac{1}{4}\|x_n^s\| \leq \frac{1}{4}\sqrt{\|A_s\|^2+1}$. As the objective function is $(\Lambda + \Delta)$ -strongly convex, A is a full-rank matrix with each eigenvalue greater than $n(\Lambda + \Delta)$. Therefore, $|\det(\mathbb{I} + A^{-1}E_n^s)| \leq 1 + \frac{\frac{1}{4}\sqrt{\|A_s\|^2+1}}{n(\Lambda+\Delta)}$. Similarly, $|\det(\mathbb{I} + A^{-1}E_{n'}^s)| \leq 1 + \frac{\frac{1}{4}\sqrt{\|A_s\|^2+1}}{n(\Lambda+\Delta)}$. Therefore,

$$\frac{|\det(\text{Jacob}(f \rightarrow b \mid \mathcal{D}))|^{-1}}{|\det(\text{Jacob}(f \rightarrow b' \mid \mathcal{D}'))|^{-1}} \leq \exp\left(2 \sum_{s=1}^q \log\left(1 + \frac{\frac{1}{4}\sqrt{\|A_s\|^2+1}}{n\Lambda}\right)\right) \quad (7)$$

Next, we bound the ratio of the probability density of random vector b with respect to two neighboring data sets. We have:

$$b - b' = \nabla \ell(f^*; d_n) - \nabla \ell(f^*; d'_n)$$

From Lemma 4.2,

$$\|b - b'\| \leq \sum_{s=1}^q \sqrt{\|A_s\|^2+4} + \max_{s \in \{1, \dots, q\}} \sqrt{\|2A_s\|^2+4}$$

Therefore,

$$\frac{\text{pdf}(b \mid \mathcal{D})}{\text{pdf}(b' \mid \mathcal{D}')} \leq \exp(\epsilon' \|b - b'\|/t) \leq \exp(\epsilon') \quad (8)$$

From (6), (7), (8), and $\epsilon = \epsilon' + 2 \sum_{s=1}^q \log\left(1 + \frac{\frac{1}{4}\sqrt{\|A_s\|^2+1}}{n\Lambda}\right)$, the theorem follows. \square

5 PROPOSED SAMPLING APPROACH

In this section, we propose a solution which guarantees differential privacy by directly sampling a random output from a modified version of the posterior distribution. In this work, we pick a normal distribution as the prior distribution. This is equivalent to using:

$$\mathcal{U}(f; \mathcal{D}) = -\frac{1}{2}\sigma\|f\|^2 - \sum_{i=1}^n \ell(f; d_i)$$

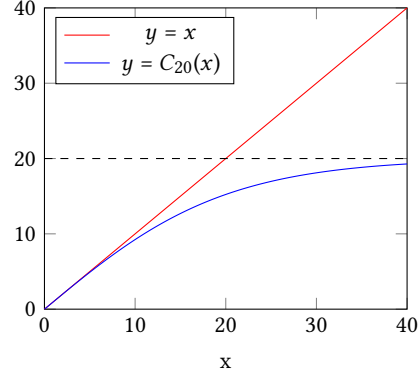


Fig. 2. The illustration of the sanitizer function (blue) with maximum value 20 and the identity function (red).

as the utility function in the *exponential mechanism* (McSherry and Talwar 2007) where the parameter σ is used to control the variance of the prior normal distribution. Then, the differentially private output is sampled from the following distribution:

$$\text{pdf}(f) \propto \exp\left(\frac{\epsilon \mathcal{U}(f; \mathcal{D})}{2\Delta \mathcal{U}}\right)$$

where $\Delta \mathcal{U} = \max_{d(\mathcal{D}, \mathcal{D}')=1, f} \|\mathcal{U}(f; \mathcal{D}) - \mathcal{U}(f; \mathcal{D}')\|$ is the sensitivity of \mathcal{U} . The reason we pick a normal prior distribution instead of a uniform prior distribution is not because our proposed solution required so to guarantee differential privacy but we observe that with a normal prior distribution the sampling algorithm converges better and is more stable.

Moreover, this approach requires the utility function $\mathcal{U}(f; \mathcal{D})$ has to have a bounded sensitivity. However, the loss function $\ell(\cdot)$ is not bounded. Therefore, the function $\mathcal{U}(f; \mathcal{D})$ has unbounded sensitivity. In order to overcome this difficulty, we propose a smooth sanitizer function $C(x)$ which is used to control the maximum value of the loss function $\ell(\cdot)$. The definition of $C(x)$ is given as follows:

$$C_v(x) = v \cdot \tanh\left(\frac{x}{v}\right)$$

which is illustrated in Figure 2. We now take the composition of $C_v(\cdot)$ with $\ell(f; d_i)$ to have a bounded-sensitivity utility function:

$$\mathcal{U}(f; \mathcal{D}) = -\frac{1}{2}\sigma \|f\|^2 - \sum_{i=1}^n C_v(\ell(f; d_i))$$

We intentionally pick the $\tanh(\cdot)$ function as the sanitizer because it nicely keeps the loss function in its original form when the value of the loss function is near 0. Meanwhile, it deforms the loss function at large values to make the function finite. The advantage of this approach is that the sampled parameter can arbitrary large while the objective function is kept almost the same around the optimal parameter which maximizes the posterior probability. We describe the pseudo-code of our approach in Algorithm 3.

Algorithm 3 $\mathcal{A}_{\text{Sanitized-EXP}}$: Sanitized Loss Mechanism**Input:** Data set $\mathcal{D} = \{d_i\}_{i=1}^n$, loss function $\ell(f; d_i)$, privacy budget ϵ , maximum value v , parameter Λ **Output:** f 1: $\mathcal{U}(f; \mathcal{D}) = -\frac{1}{2}\sigma\|f\|^2 - \sum_{i=1}^n C_v(\ell(f; d_i))$ 2: Sample a random vector f with the probability density

$$\text{pdf}(f) \propto \exp\left(\frac{\epsilon}{2v}\mathcal{U}(f; \mathcal{D})\right)$$

THEOREM 5.1 (PRIVACY GUARANTEE). *Algorithm 3 is ϵ -differentially private.*

PROOF. For two neighboring data sets \mathcal{D} and \mathcal{D}' ,

$\Delta\mathcal{U} = \max_{f, d(\mathcal{D}, \mathcal{D}')=1} |\mathcal{U}(f; \mathcal{D}) - \mathcal{U}(f; \mathcal{D}')| \leq v$. Therefore, at any point f , we have

$$\begin{aligned} \frac{\text{pdf}(f | \mathcal{D})}{\text{pdf}(f | \mathcal{D}')} &= \frac{\exp\left(\frac{\epsilon}{2v}\mathcal{U}(f; \mathcal{D})\right) / \int \exp\left(\frac{\epsilon}{2v}\mathcal{U}(f; \mathcal{D})\right) df}{\exp\left(\frac{\epsilon}{2v}\mathcal{U}(f; \mathcal{D}')\right) / \int \exp\left(\frac{\epsilon}{2v}\mathcal{U}(f; \mathcal{D}')\right) df} \\ &\leq \exp\left(\frac{2\epsilon}{2v} |\mathcal{U}(f; \mathcal{D}) - \mathcal{U}(f; \mathcal{D}')|\right) \\ &\leq \exp(\epsilon) \end{aligned}$$

Therefore, Algorithm 3 is ϵ -differentially private. \square

The problem with Algorithm 3 is that there is no run-time efficient algorithm to sample the distribution of f exactly. Bassily et al. (Bassily et al. 2014) proposed a polynomial run-time sampling algorithm. However, their proposed algorithm is still impractical due to the high degree of the polynomial run-time complexity and only apply for the log-convex function. Recently, there are developments (Ahn et al. 2012; Chen et al. 2014; Ma et al. 2015) in Markov Chain Monte Carlo (MCMC) method which can be applied to machine learning problems with large data sets. The idea is to construct Markov chains to simulate dynamical systems with stochastic gradients. At each step, we compute the gradient at the current location, then add a controlled amount of noise to the gradient and follow the noisy gradient to a new location. Asymptotically, the stationary distribution of this process converges to the true distribution from which the gradient is computed.

In this work, we propose to use an MCMC sampling algorithm, namely Preconditioned Stochastic Gradient Langevin Dynamics (pSGLD) (Li et al. 2015), to approximately sample the posterior distribution. pSGLD is good at sampling variables with differences in scale which is useful for our problem because the parameter α is usually much larger in magnitude than the parameter β (recall that $f = [\alpha, \beta]'$). The pseudo-code of pSGLD is described in Algorithm 4. At Line 1, we initialize the values of V_0 and f_1 . Line 3 computes the learning rate ϵ_t . It is required that $\lim_{t \rightarrow \infty} \sum_t \epsilon_t \rightarrow \infty$ and $\lim_{t \rightarrow \infty} \sum_t \epsilon_t^2 < \infty$ to guarantee the convergence. We sample uniformly k records from \mathcal{D} for estimating the average gradient \bar{g}^t (Line 5). We then compute the variance of the gradient at Line 6 (\odot is the element-wise product) and convert it to the preconditioned matrix G^t at Line 7. We update the parameter at Line 8 with a noise variable $\mathcal{N}(0, \epsilon_t G^t)$. It is worth to note that there is a permanent bias in pSGLD due to excluding a correction term in the updating step (Line 8). However, this bias is negligible and excluding the correction term helps to speed up the sampling algorithm which then helps to reduce the finite-sample bias as more steps are executed in a finite amount of time.

Algorithm 4 $\mathcal{A}_{\text{pSGLD}}$: pSGLD Sampling Algorithm**Input:** Data set $\mathcal{D} = \{d_i\}_{i=1}^n$, loss function ℓ , privacy parameter ϵ , μ , k , bounded value v and learning rate τ **Output:** f^{T+1}

```

1:  $V_0 \leftarrow \vec{0}, f_1 \leftarrow \vec{0}$ 
2: for  $t = 1$  to  $T$  do
3:   Compute  $\epsilon_t \leftarrow t^{-\tau}$ 
4:   Uniformly sample  $\Omega_k^t = \{d_{t_1}, \dots, d_{t_k}\} \subset \mathcal{D}$ 
5:   Compute  $\bar{g}^t = \frac{\epsilon}{2v}(\frac{\sigma f^t}{n} + \frac{1}{k} \sum_{i=1}^k \nabla C_v(\ell(f^t, d_{t_i})))$ 
6:    $V^t \leftarrow \mu V^{t-1} + (1 - \mu)(\bar{g}^t \odot \bar{g}^t)$ 
7:    $G^t \leftarrow 1/(\lambda \mathbb{I} + \text{diag}(\sqrt{V^t}))$ 
8:    $f^{t+1} \leftarrow f^t - \epsilon_t (G^t \cdot n \bar{g}^t) + \mathcal{N}(0, \epsilon_t G^t)$ 
9: Output  $f^{T+1}$ 

```

6 EXPERIMENTAL EVALUATION

In this section, we present the results of our experiments on four real data sets. We focus on answering the following three important research questions: (1) Does the sampling approach converge to its stationary distribution? (2) What is the trade-off between privacy and accuracy as compared to the non-private estimation? (3) Are the discrete-time regression models good alternatives to the Cox regression model? In the following sections, we address the above research questions accordingly.

6.1 Data Sets

Table 1. Statistics of the data sets.

Data set	Size	#uncensored	#explanatory variables
FL	7874	2169	8
TB	16116	1761	3
WT	21685	18615	3
SB	53558	16341	3

We use four real data sets in our experiments. Table 1 gives the statistics of these data sets.

- *The FLchain data set* (FL) - It is obtained from a study on the association of the serum free light chain with higher death rates (Dispenzieri et al. 2012; Kyle et al. 2006). The survival time of a patient is measured in days from enrollment until death. The censored cases are patients who are still alive at the last contact. The explanatory variables are age, sex, creatinine, mgus, etc.
- *The time-to-second-birth* (SB) and *time-to-third-birth* (TB) data sets - They are obtained from The Medical Birth Registry of Norway (Irgens 2000). The survival time is the time between the first and second births, and between the second and third births respectively. The censored cases are women who do not have the second birth, and the third birth respectively, at the time the data are collected. The explanatory variables in SB (resp., TB) are age, sex and death of earlier children (resp., age, spacing and sibs).

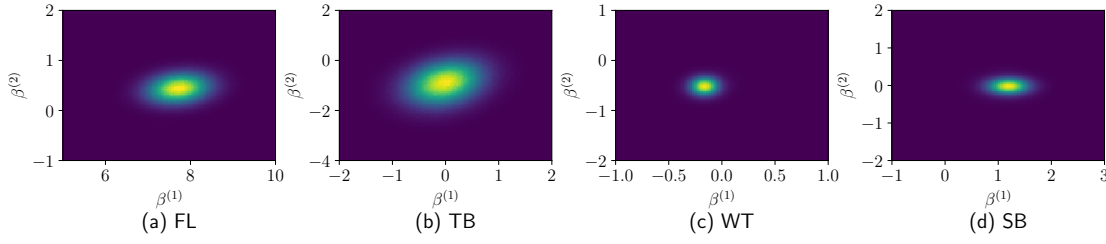


Fig. 3. An illustration of the probability densities of the sampling posteriors after 250 epochs at privacy budget $\epsilon = 6.4$.

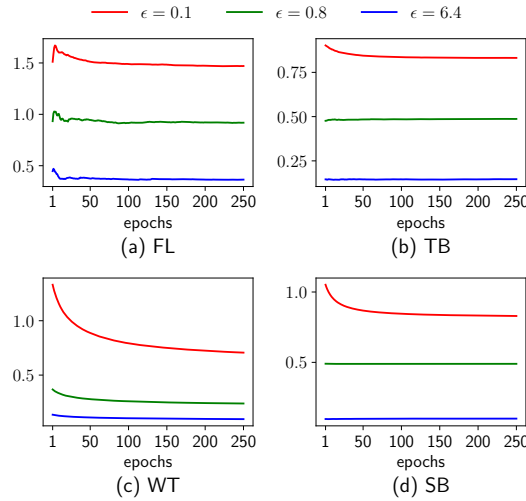


Fig. 4. An illustration of MRE as a statistical test for samples from the pSGLD sampling algorithm.

- *The Wichert data set (WT)* - It contains records on unemployment duration of people in Germany (Wichert and Wilke 2008). The survival time is the duration of unemployment until having a job again. The censored cases are the ones who do not have a new job at the time the data are collected. The explanatory variables are sex, age and wage.

The survival times in these four data sets are normalized to the interval $[0, 1]$. We set the number of discrete-time intervals $q = 200$. All the vectors of the explanatory variables are normalized to have zero mean and fitted inside the unit sphere. We use the natural cubic spline with $e = 3$ knots to model the baseline hazard effect.

6.2 Convergence of the Proposed Sampling Approach

This section reports on the convergence of our proposed sampling approach. The aim is to check whether it converges to the stationary distribution. The loss function is bounded by the value $v = 2 \log(n)$ where n is the size of the data set. We set the parameter $\sigma = 10^{-2} \cdot 2v/\epsilon$. At each step of the Markov chain, we randomly pick $k = 200$ records from the data set to compute the gradient. We set the parameters $\tau = 0.51$, $\lambda = 10^{-5}$ and $\mu = 0.99$ in Algorithm 3. In Figure 3, we

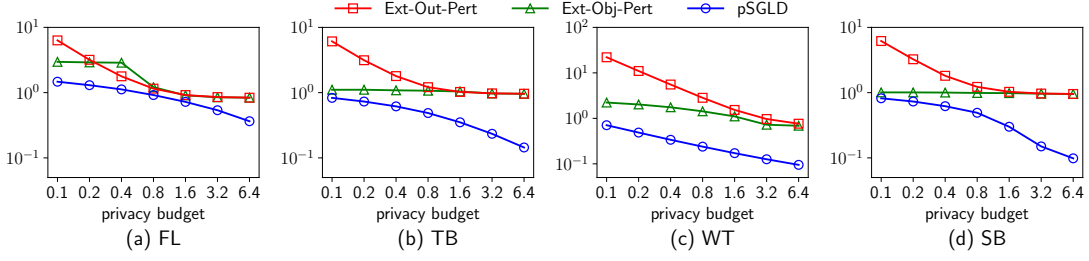


Fig. 5. The performance of our proposed approaches in MRE with privacy budget ϵ from 0.1 to 6.4.

plot the estimated probability densities of the two first parameters ($\beta^{(1)}$ and $\beta^{(2)}$) after 250 epochs from the sampling process. We remove the first 10^4 steps as the Markov chain does not reach the stationary distribution at the beginning. We can observe that the probability densities of the samples are very similar to the normal distributions which are actually what we expect when sampling from the posterior distributions.

For a more formal test, we use the mean relative error (MRE) as a statistical test of convergence. MRE is defined as follows:

$$\text{MRE} = \frac{1}{t} \sum_{i=1}^t \frac{\|f_i - f^*\|}{\|f^*\|} \quad (9)$$

where f_i is the parameter vector from the sampling process, f^* is the optimal parameter vector which maximizes the likelihood in non-private setting and t is the number of samples. We plot the MRE as the function of epochs with three different privacy budgets in Figure 4. Each epoch is a bundle of n steps. We observe that after 250 epochs, MRE becomes stable which indicates that the sampling procedure converges to its stationary distribution.

6.3 Trade-off between Privacy and Accuracy

Table 2. The performance in MRE of Ext-Out-Pert approach for different regularization parameters with privacy budget $\epsilon = 6.4$. The best performance results are in bold.

Λ	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^0
FL	981.635	98.135	9.828	1.195	0.837	0.882
TB	90.994	9.113	1.273	0.964	0.975	0.983
WT	342.939	34.297	3.456	0.763	0.765	0.81
SB	9.59	1.224	0.957	0.993	0.998	0.999

In this section, we investigate the trade-off between privacy and accuracy in our proposed approaches. We first need to pick the value of regularization terms for the perturbation approaches (Ext-Obj-Pert and Ext-Out-Pert) as the accuracy of these approaches are very much depend on the regularization parameter Λ . We report in Table 2 the MREs of Ext-Out-Pert with different values of Λ and privacy budget $\epsilon = 6.4$. For consistency in performance comparison, we will use the best values of Λ , which lead to the smallest relative error per data set.

To measure the accuracy of the proposed approaches at different privacy levels, the privacy budget is varied from 0.1 to 6.4. We also use MRE for the measurement. The results are shown in Figure 5. Overall, pSGLD outperforms both

Ext-Out-Pert and Ext-Obj-Pert approaches. Moreover, we observe that the accuracy of Ext-Out-Pert and Ext-Obj-Pert does not improve much at high privacy budgets. It is due to the large regularization parameter that causes the output parameter moving towards the zero vector instead of the optimal parameter as the regularization term is the dominant factor of the objective function. Meanwhile, our proposed sampling approach (pSGLD) does not suffer from this effect which leads to much better results at high privacy budgets.

6.4 Comparison with Cox regression

Table 3. Relative error of the discrete-time survival regression as compared to the Cox regression.

Data set	Relative error (%)
FL	2.589%
TB	9.039%
WT	3.617%
SB	2.618%

Here, we want to confirm that the discrete-time regression models are good alternatives to the Cox regression model. We compare the results obtained from the non-private discrete-time regression models without regularization term to the results obtained from Cox regression. We use the relative error (RE) which is defined as:

$$RE = \frac{\|\beta - \beta^*\|}{\|\beta^*\|}$$

where β is from the discrete-time regression with logit link and β^* is from Cox regression. The results are shown in Table 3. We observe that the results obtained from the discrete-time regressions are very similar to the results obtained from the Cox regression with relative errors ranging from 2% – 9%. At the worse case of the data set TB, the parameter obtained from the discrete-time model $\beta = [0.0122443, -0.849823, -0.239539]'$ is still a good approximation of the parameter obtained from the Cox model $\beta^* = [0.0585478, -0.790977, -0.23906]'$. As such, these results confirm that the discrete-time regression models are good alternatives to the Cox regression in practice.

7 CONCLUSION

In this work, we propose solutions for the problem of protecting differential privacy for discrete-time regression models used in survival analysis. In particular, we extend the perturbation approaches to a generalized form in which the loss function is a sum of logistic loss functions. In addition, we propose a sampling approach to practically protect differential privacy by sampling a scaled posterior distribution with the pSGLD sampling algorithm. Even though we focus our work on discrete-time survival regression, our proposed approaches can be applied to other problems with similar loss functions as well. Moreover, our proposed approaches can be easily extended to discrete-time regression models in which the explanatory variables are changed over time. For further work, a differentially private version of Cox regression would be a good complement to our work.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Xiaokui Xiao for his insightful comments on the privacy problem of Cox regression.

REFERENCES

- Sungjin Ahn, Anoop Korattikara Balan, and Max Welling. 2012. Bayesian Posterior Sampling via Stochastic Gradient Fisher Scoring. In *ICML*.
- Paul D Allison. 1982. Discrete-time methods for the analysis of event histories. *Sociological methodology* 13 (1982), 61–98.
- Per Kragh Andersen and Richard David Gill. 1982. Cox’s regression model for counting processes: a large sample study. *The annals of statistics* (1982), 1100–1120.
- Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Foundations of Computer Science (FOCS), 2014 IEEE 55th Annual Symposium on*. IEEE, 464–473.
- Patrick Billingsley. 2008. *Probability and measure*. John Wiley & Sons.
- David Blumenthal and Marilyn Tavenner. 2010. The “meaningful use” regulation for electronic health records. *N Engl J Med* 2010, 363 (2010), 501–504.
- Christian Borgs, Jennifer Chayes, and Adam Smith. 2015. Private graphon estimation for sparse graphs. In *Advances in Neural Information Processing Systems*. 1369–1377.
- Kamalika Chaudhuri and Claire Monteleoni. 2009. Privacy-preserving logistic regression. In *Advances in Neural Information Processing Systems*. 289–296.
- Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12, Mar (2011), 1069–1109.
- Tianqi Chen, Emily Fox, and Carlos Guestrin. 2014. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*. 1683–1691.
- David R Cox. 1992. Regression models and life-tables. In *Breakthroughs in statistics*. Springer, 527–541.
- David Roxbee Cox and David Oakes. 1984. *Analysis of survival data*. Vol. 21. CRC Press.
- Catherine M DesRoches, Eric G Campbell, Sowmya R Rao, Karen Donelan, Timothy G Ferris, Ashish Jha, Rainu Kaushal, Douglas E Levy, Sara Rosenbaum, Alexandra E Shields, et al. 2008. Electronic health records in ambulatory care—a national survey of physicians. *New England Journal of Medicine* 359, 1 (2008), 50–60.
- Nan Ding, Youhan Fang, Ryan Babbush, Changyou Chen, Robert D Skeel, and Hartmut Neven. 2014. Bayesian sampling using stochastic gradient thermostats. In *Advances in neural information processing systems*. 3203–3211.
- Angela Dispenzieri, Jerry A Katzmman, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton, et al. 2012. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*, Vol. 87. Elsevier, 517–523.
- Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*. Springer, 1–19.
- Cynthia Dwork. 2009. The differential privacy frontier. In *Theory of Cryptography Conference*. Springer, 496–502.
- Cynthia Dwork. 2011. A firm foundation for private data analysis. *Commun. ACM* 54, 1 (2011), 86–95.
- Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- Roger Fletcher. 2013. *Practical methods of optimization*. John Wiley & Sons.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.
- Wolfgang Hess and Maria Persson. 2012. The duration of trade revisited. *Empirical Economics* (2012), 1–25.
- Lorentz M Irgens. 2000. The Medical Birth Registry of Norway. Epidemiological research and surveillance throughout 30 years. *Acta obstetricia et gynecologica Scandinavica* 79, 6 (2000), 435–439.
- Ashish K Jha, Catherine M DesRoches, Eric G Campbell, Karen Donelan, Sowmya R Rao, Timothy G Ferris, Alexandra Shields, Sara Rosenbaum, and David Blumenthal. 2009. Use of electronic health records in US hospitals. *New England Journal of Medicine* 360, 16 (2009), 1628–1638.
- Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2013. Analyzing graphs with node differential privacy. In *Theory of Cryptography*. Springer, 457–476.
- Daniel Kifer, Adam Smith, and Abhradeep Thakurta. 2012. Private convex empirical risk minimization and high-dimensional regression. *Journal of Machine Learning Research* 1, 41 (2012), 3–1.
- Robert A Kyle, Terry M Therneau, S Vincent Rajkumar, Dirk R Larson, Matthew F Plevak, Janice R Offord, Angela Dispenzieri, Jerry A Katzmman, and L Joseph Melton III. 2006. Prevalence of monoclonal gammopathy of undetermined significance. *New England Journal of Medicine* 354, 13 (2006), 1362–1369.
- Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. 2015. Preconditioned stochastic gradient Langevin dynamics for deep neural networks. *arXiv preprint arXiv:1512.07666* (2015).
- Chao Li, Michael Hay, Vibhor Rastogi, Gerome Miklau, and Andrew McGregor. 2010. Optimizing linear counting queries under differential privacy. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. ACM, 123–134.
- Wentian Lu and Gerome Miklau. 2014. Exponential random graph estimation under differential privacy. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 921–930.
- Yi-An Ma, Tianqi Chen, and Emily Fox. 2015. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*. 2917–2925.

- Ashwin Machanavajjhala, Aleksandra Korolova, and Atish Das Sarma. 2011. Personalized social recommendations: accurate or private. *Proceedings of the VLDB Endowment* 4, 7 (2011), 440–450.
- Frank McSherry and Ilya Mironov. 2009. Differentially private recommender systems: building privacy into the net. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 627–636.
- Frank McSherry and Kunal Talwar. 2007. Mechanism Design via Differential Privacy, In Annual IEEE Symposium on Foundations of Computer Science (FOCS). <https://www.microsoft.com/en-us/research/publication/mechanism-design-via-differential-privacy/>
- Bengt Muthén and Katherine Masyn. 2005. Discrete-time survival mixture analysis. *Journal of Educational and Behavioral statistics* 30, 1 (2005), 27–58.
- E Seiler and B Simon. 1975. An inequality among determinants. *Proceedings of the National Academy of Sciences of the United States of America* 72, 9 (1975), 3277.
- Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. 2015. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. 2493–2502.
- Laura Wichert and Ralf A Wilke. 2008. Simple non-parametric estimators for unemployment duration analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57, 1 (2008), 117–126.
- Shipeng Yu, Glenn Fung, Romer Rosales, Sriram Krishnan, R Bharat Rao, Cary Dehing-Oberije, and Philippe Lambin. 2008. Privacy-preserving cox regression for survival analysis. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1034–1042.
- Xiaojuan Zhang, Rui Chen, Jianliang Xu, Xiaofeng Meng, and Yingtao Xie. 2014. Towards accurate histogram publication under differential privacy. In *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM, 587–595.