Kulkarni, Tejas; Jälkö, Joonas; Koskela, Antti ; Kaski, Samuel; Honkela, Antti

Differentially Private Bayesian Inference for Generalized Linear Models

Published in:
Proceedings of the 38th International Conference on Machine Learning

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Kulkarni, T., Jälkö, J., Koskela, A., Kaski, S., & Honkela, A. (2021). Differentially Private Bayesian Inference for Generalized Linear Models. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th International Conference on Machine Learning* (pp. 5838-5849). (Proceedings of Machine Learning Research ; Vol. 139). JMLR.
http://proceedings.mlr.press/v139/kulkarni21a.html

Differentially Private Bayesian Inference for Generalized Linear Models

Tejas Kulkarni [1] **Joonas Jälkö** [1] **Antti Koskela** [2] **Samuel Kaski** [1 3] **Antti Honkela** [2]

Abstract

Generalized linear models (GLMs) such as logistic regression are among the most widely used arms in data analyst's repertoire and often used on sensitive datasets. A large body of prior works that investigate GLMs under differential privacy (DP) constraints provide only private point estimates of the regression coefficients, and are not able to quantify parameter uncertainty.

In this work, with logistic and Poisson regression as running examples, we introduce a generic noise-aware DP Bayesian inference method for a GLM at hand, given a noisy sum of summary statistics. Quantifying uncertainty allows us to determine which of the regression coefficients are statistically significantly different from zero. We provide a tight privacy analysis and experimentally demonstrate that the posteriors obtained from our model, while adhering to strong privacy guarantees, are close to the non-private posteriors.

1. Introduction

Differential privacy (DP) (Dwork et al., 2006) provides a strong framework for protecting the privacy of data subjects against privacy violations via models trained on their personal data. DP protection requires injecting noise to the learning process. Bayesian inference is a natural complement to DP, because it seeks to quantify the impact of noise to inference result in terms of quantifying the uncertainty of the result. In our work we seek to develop a Bayesian method to perform inference under DP and quantify the uncertainty caused by the injected noise for the widely used class of regression models, generalised linear models (GLMs). This method allows statistical inference on the regression coefficients, such as determining which coefficients can be confidently inferred to be different from zero.

Using Bayesian inference to counter the noise injected to ensure DP was first proposed by Williams and McSherry (2010). The process is fairly straightforward for models where the joint distribution $\Pr[\boldsymbol{D}, \boldsymbol{\theta}, \boldsymbol{Z}]$ over the data $\boldsymbol{D}$, all parameters $\boldsymbol{\theta}$ and possible latent variables $\boldsymbol{Z}$ of interest is specified as part of the model. This was demonstrated by Bernstein and Sheldon (2018), who presented an efficient inference method for exponential family models of this form.

The problem becomes significantly more difficult for discriminative models such as regression models, where the model does not specify a distribution over the input data $\boldsymbol{X}$ but only target outputs $\boldsymbol{Y}$ and parameters $\Pr[\boldsymbol{Y}, \boldsymbol{\theta} \mid \boldsymbol{X}]$. This is because we also consider input $\boldsymbol{X}$ as private information, and thus we cannot observe it directly. On the other hand, the model does not specify a natural prior for $\boldsymbol{X}$ either. In order to solve this issue, we need to augment the model to include a prior for $\boldsymbol{X}$. This was first demonstrated by Bernstein and Sheldon (2019), who developed a method for linear regression by placing a Gaussian prior on $\boldsymbol{X}$. The method relies on the ability to express the model via a sufficient statistic of fixed size, and hence only applies to linear regression.

In our work we extend this sufficient statistics based noise-aware DP Bayesian inference framework for a much broader class of regression models, GLMs. These include many of the most widely used statistical models, such as logistic regression and Poisson regression. As GLMs typically have no sufficient statistics, this is achieved by approximating the joint distribution of the inputs and outputs under the GLM using a finite number of moments as approximate sufficient statistics and fitting the model parameters to match these moments.

Related work. Linear models have received a huge amount of attention under DP since its proposal. The techniques for point estimates for regression parameters fall to five main categories, a) *objective perturbation* (Chaudhuri et al., 2011; Kifer and Machanavajjhala, 2011; Zhang et al., 2012; Iyengar et al., 2019), b) *output perturbation* (Wu et al.,

[1]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland [2]Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland [3]Department of Computer Science, University of Manchester, United Kingdom. Correspondence to: Tejas Kulkarni <tejasvijaykulkarni@gmail.com>, Antti Honkela <antti.honkela@helsinki.fi>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

2017; Zhang et al., 2017), c) *gradient perturbation* (Bassily et al., 2014; Abadi et al., 2016), d) *subsample and aggregate* (Smith, 2008; Dwork and Smith, 2010; Barrientos et al., 2019), and e) *sufficient statistic perturbation (SSP)* (McSherry and Mironov, 2009; Vu and Slavkovic, 2009; Sheffet, 2017; Wang, 2018). Other representative works that specifically study *generalized linear models* (GLMs) more generally under various DP models include (Kifer et al., 2012; Jain and Thakurta, 2014; Pihur et al., 2018; Wang et al., 2019; 2021). Only a handful of these works, e.g. (Sheffet, 2017; Barrientos et al., 2019), quantify the uncertainty in the model parameter estimates via frequentist tools such as confidence intervals and hypothesis testing.

To quantify the uncertainty beyond the frequentist tools, many Bayesian inference techniques under DP setting have been proposed, starting from the seminal work of Williams and McSherry (2010). The field can be roughly clustered into three broad categories:

1. *sufficient statistics perturbation* based inference (Foulds et al., 2016; Zhang et al., 2016; Honkela et al., 2018; Bernstein and Sheldon, 2018; 2019; Park et al., 2020)
2. *gradient perturbation* based Markov chain Monte Carlo (MCMC) (Wang et al., 2015; Zhang et al., 2016; Li et al., 2019a) and variational inference (VI) (Jälkö et al., 2017)
3. *DP posterior sampling* (Dimitrakakis et al., 2014; Foulds et al., 2016; Zhang et al., 2016; Heikkilä et al., 2019; Yıldırım and Ermiş, 2019).

Categories 2 and 3 aim to provide a general-purpose solution for differentially private Bayesian inference. However, the output of these mechanisms is an approximation of the posterior distribution where the impact of added uncertainty from privacy is not quantified. Category 1 is most closely related to our work. In these works, posterior distribution of the model parameters is captured through perturbed sufficient statistics, but also here many approaches fail to quantify the impact of the added uncertainty from privacy.

Among the several approaches proposed, inference techniques based on *sufficient statistics perturbation* stand out due to their computational efficiency and accuracy (Wang, 2018). Unlike DP variants of general purpose MCMC methods, the privacy cost of training in a sufficient statistics based model is typically invariant to the number of iterations/posterior samples. We only pay once to perturbs the sum of sufficient statistics and then rely on the post-processing property of DP to run iterative inference without additional privacy cost.

Main contributions. The main contributions of this work are as follows:

- We derive noise-aware DP Bayesian inference for GLMs based on approximate sufficient statistics from low-order polynomials.
- We prove tight (ϵ, δ)-DP bounds for releasing the approximate sufficient statistics.
- We demonstrate accurate privacy-preserving inference of which regression coefficients are significantly different from zero.
- We demonstrate high degree of similarity between the privacy-preserving and non-private posterior distributions for GLMs even under strong privacy for moderately sized data.

2. Background

2.1. Differential Privacy (DP)

Assume a generic dataset $\boldsymbol{D} \in \mathbb{R}^{N \times d}$ containing d-dimensional records of N individuals. We define neighbourhood relation $\boldsymbol{D} \sim \boldsymbol{D}'$ when $\boldsymbol{D}'$ can be obtained from $\boldsymbol{D}$ by replacing a single record. Dwork et al. (2006) proposed the following notion:

Definition 2.1. *For $\epsilon \geq 0, \delta \geq 0$, a randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if for all neighbouring datasets $\boldsymbol{D} \sim \boldsymbol{D}'$, and for all outputs $O \subseteq Range(\mathcal{M})$, the following constraint holds:*

$$\Pr[\mathcal{M}(\boldsymbol{D}) \in O] \leq \exp(\epsilon) \times \Pr[\mathcal{M}(\boldsymbol{D}') \in O] + \delta. \quad (1)$$

Lower values of ϵ and δ provide stronger privacy.

The condition (1) can often be satisfied, for example, by adding Gaussian noise to a function of the dataset such that every individual's contribution is masked. A key property of DP is its robustness to post-processing: the privacy loss of $\mathcal{M}$ cannot be increased by applying any randomized function independent of the data to $\mathcal{M}$'s output.

The concept of *sensitivity* measures the worst-case impact of an individual's record on the output of a function.

Definition 2.2. *The L_2-sensitivity $\Delta_{\boldsymbol{t}}$ of a function $\boldsymbol{t} : \mathbb{R}^{N \times d} \to \mathbb{R}^m$ is defined as $\Delta_{\boldsymbol{t}} = \max_{\boldsymbol{D} \sim \boldsymbol{D}'} ||\boldsymbol{t}(\boldsymbol{D}) - \boldsymbol{t}(\boldsymbol{D}')||_2$.*

Analytic Gaussian Mechanism (Balle and Wang, 2018). Balle and Wang (2018) proposed an algorithmic noise calibration strategy based on the Gaussian cumulative density function (CDF) to obtain a mechanism that adds the least amount of Gaussian noise needed for (ϵ, δ)-DP.

Definition 2.3. *(Analytic Gaussian Mechanism) For any $\epsilon \geq 0, \delta \in [0, 1]$, a mechanism $\mathcal{M}(\boldsymbol{D}) = \boldsymbol{t}(\boldsymbol{D}) + \zeta$ with sensitivity $\Delta_{\boldsymbol{t}}$ satisfies (ϵ, δ)-DP with $\zeta \sim \mathcal{N}(0, \sigma^2 I)$ iff*

$$\Phi\left(\frac{\Delta_{\boldsymbol{t}}}{2\sigma} - \frac{\epsilon\sigma}{\Delta_{\boldsymbol{t}}}\right) - \exp(\epsilon)\Phi\left(-\frac{\Delta_{\boldsymbol{t}}}{2\sigma} - \frac{\epsilon\sigma}{\Delta_{\boldsymbol{t}}}\right) \leq \delta. \quad (2)$$

We use the implementation based on Algorithm 1 of Balle and Wang (2018) to find a minimal σ that satisfies the condition (2).

2.2. Bayesian inference based on sufficient statistics

For certain statistical models, the information about data needed for parameter inference can be captured by a limited number of *sufficient statistics*. Sufficient statistics are available for exponential family models, such as linear regression. For Bayesian linear regression, the sufficient statistics are $\boldsymbol{s} = \sum_{i=1}^{N} \boldsymbol{t}(\boldsymbol{x}_i, y_i) = [\boldsymbol{X}^T\boldsymbol{X}, \boldsymbol{X}^T\boldsymbol{y}, \boldsymbol{y}^T\boldsymbol{y}]$, where $\boldsymbol{X} \in \mathbb{R}^{N\times d}$. With access to $\boldsymbol{s}$ ($\mathcal{O}(N)$ operation), we can evaluate the total log-likelihood $\log(\prod_{i=1}^{N} \Pr[y_i|\boldsymbol{x}_i, \boldsymbol{\theta}]) = \sum_{i=1}^{N} \log(\Pr[y_i|\boldsymbol{x}_i, \boldsymbol{\theta}])$ or its gradients in nearly a constant time. As a consequence, running time of a training algorithm taking K passes over data reduces substantially to $\mathcal{O}(N + K)$ from $\mathcal{O}(NK)$.

2.3. Privacy-preserving posterior inference with sufficient statistics

Consider a conditional probabilistic model $\Pr[\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{X}]$ where the information in $(\boldsymbol{X}, \boldsymbol{y})$ needed for inference of $\boldsymbol{\theta}$ can be represented by sufficient statistics $\boldsymbol{s} \in \mathbb{R}^m$. Let $\boldsymbol{\theta} \in \mathbb{R}^d$ denote the regression parameters. In order to adapt the model for DP, we need to make additional assumptions on $\boldsymbol{X}$. Following Bernstein and Sheldon (2019), we assume $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is an additional parameter denoting the input data covariance matrix. We guarantee privacy by operating solely on the perturbed sufficient statistics, denoted by $\boldsymbol{z} \in \mathbb{R}^m$. The noise-aware joint posterior distribution of the model parameters $\boldsymbol{\theta}, \boldsymbol{\Sigma}$, given noisy sufficient statistics $\boldsymbol{z}$ is

$$\begin{aligned}\Pr[\boldsymbol{\theta},\boldsymbol{\Sigma}|\boldsymbol{z}] \propto \Pr[\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{z}] &= \int_{\boldsymbol{s}} \Pr[\boldsymbol{\theta}, \boldsymbol{\Sigma}, \boldsymbol{s}, \boldsymbol{z}]\,\mathrm{d}\boldsymbol{s} \\ &= \int_{\boldsymbol{s}} \Pr[\boldsymbol{\theta}]\Pr[\boldsymbol{\Sigma}]\Pr[\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}]\Pr[\boldsymbol{z} \mid \boldsymbol{s}]\,\mathrm{d}\boldsymbol{s},\end{aligned} \tag{3}$$

where $\Pr[\boldsymbol{\theta}]$ and $\Pr[\boldsymbol{\Sigma}]$ are the priors for model parameters and the privacy inducing noise is quantified by the term $\Pr[\boldsymbol{z} \mid \boldsymbol{s}]$. Figure 1 depicts above model. The remaining challenge is to define the probabilistic model for the latent sufficient statistics $\boldsymbol{s}$.

Normal approximation of $\boldsymbol{s}$. We obtain the sufficient statistic distribution $\Pr[\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}]$ by marginalizing over the data: $\Pr[\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}] = \int_{\boldsymbol{X},\boldsymbol{y}:\boldsymbol{t}(\boldsymbol{X},\boldsymbol{y})=\boldsymbol{s}} \Pr[\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}]\,\mathrm{d}\boldsymbol{X}\,\mathrm{d}\boldsymbol{y}$. However, this integral is in general intractable due to (possibly infinite) number of combinations of $\boldsymbol{X} \in \mathbb{R}^{N\times d}, \boldsymbol{y} \in \mathbb{R}^N$ that produce the sufficient statistic $\boldsymbol{s}$. As $\boldsymbol{s}$ is a sum of individual sufficient statistics $\boldsymbol{t}(\boldsymbol{x}_i, y_i)$, Bernstein and Sheldon (2019) proposed to approximate $\Pr[\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}]$ as a multivariate normal distribution $\mathcal{N}(\boldsymbol{s} \mid N\boldsymbol{\mu}_{\boldsymbol{s}}, N\boldsymbol{\Sigma}_{\boldsymbol{s}})$, according to the central limit theorem,[1] with mean $\boldsymbol{\mu}_{\boldsymbol{s}} = \mathbb{E}[\boldsymbol{s}]$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{s}} = \mathrm{Cov}[\boldsymbol{s}]$.

2.4. Generalized linear models

Generalized linear models (GLMs, Nelder and Wedderburn, 1972) include some of the most commonly used statistical models. GLMs extend linear regression by allowing for the possibility of more general outcome distributions such as binary, count, and heavy-tailed observations, and using a linear model for the mean parameter of the outcome distribution.

Denoting the input $\boldsymbol{x} \in \mathbb{R}^d$ and unknown regression parameter $\boldsymbol{\theta} \in \mathbb{R}^d$, GLMs use a link function g to associate the linear model $\boldsymbol{x}^T\boldsymbol{\theta}$ to the mean of a response variable $y \in \mathbb{R}$ as $\mathbb{E}[y] = \mu = g^{-1}(\boldsymbol{x}^T\boldsymbol{\theta})$, where $g^{-1} : \mathbb{R} \to \mathbb{R}$ is the inverse link function. Typical examples of GLMs include logistic regression with binomial model for y and logistic link $g(\mu) = \log(\frac{\mu}{1-\mu})$, as well as Poisson regression with Poisson distribution for y, usually combined with the log-link $g(\mu) = \log(\mu)$.

GLMs do not generally admit finite sufficient statistics. Huggins et al. (2017) propose the PASS-GLM framework to develop polynomial approximations of degree m to GLMs that admit sufficient statistics. Sufficient statistics for such a polynomial approximation can be seen as summary statistics or approximate sufficient statistics for the original GLM.

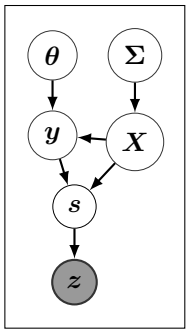

Figure 1: Differentially Private Bayesian GLM

3. Privacy-preserving Bayesian inference for GLMs

3.1. Model and problem formulation

We consider the usual centralised DP setting where a dataset $\boldsymbol{D} = \{\boldsymbol{X}, \boldsymbol{y}\} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ is a multiset of N observations. Motivated by the PASS-GLM approach, we summarise the data using low-order moments of $\boldsymbol{D}$ that are useful for inference of the particular GLM.

[1] The central limit theorem ensures the asymptotic accuracy of this approximation.

The data holder computes the sum of summary statistics $\boldsymbol{s}$ for $\boldsymbol{D}$ and releases a perturbed version $\boldsymbol{z} = \boldsymbol{s} + \zeta = \sum_{i=1}^{N} \boldsymbol{t}(\boldsymbol{x}_i, y_i) + \zeta$ with noise ζ drawn from the analytic Gaussian mechanism defined in Section 2.1. The data holder also releases the details of the target GLM and DP mechanism. With access to $\boldsymbol{z}$ and the noise mechanism, our goal is to design a noise-aware method for inferring the posterior distributions of the model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ for representative GLMs — logistic and Poisson regression.

3.2. Normal approximation of the summary statistics

Following Bernstein and Sheldon (2019), we model $\boldsymbol{x}_i \in \mathbb{R}^d$ as $\boldsymbol{x}_i \sim \mathcal{N}^d(\mathbf{0}, \boldsymbol{\Sigma})$, with an unknown covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$. This assumption can also be justified from the maximum entropy principle.

Recall that we approximate the distribution of the sum of summary statistics $\Pr[\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}]$ with a normal approximation $\Pr[\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma}] \approx \mathcal{N}(\boldsymbol{s} \mid N\boldsymbol{\mu}_{\boldsymbol{s}}, N\boldsymbol{\Sigma}_{\boldsymbol{s}})$. Next we show how to obtain the mean and covariance of the summary statistics analytically under our model. Note that we could also estimate these moments numerically, which might be easier for certain models, but creates an extra computational cost.

Closed forms for the entries of $\boldsymbol{\mu}_{\boldsymbol{s}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}}$ for logistic regression.

Assuming $y_i \in \{-1, 1\}$, the logistic regression model can be written as

$$\Pr[y_i|\boldsymbol{x}_i, \boldsymbol{\theta}] = \sigma(\boldsymbol{x}_i^T\boldsymbol{\theta})^{\frac{1+y_i}{2}}(1 - \sigma(\boldsymbol{x}_i^T\boldsymbol{\theta}))^{\frac{1-y_i}{2}}, \quad (4)$$

where $\sigma(x) = \frac{1}{1+\exp(-x)}$ denotes the sigmoid function. Motivated by the success of a second order PASS-GLM approximation for logistic regression, we use the same summary statistic, presented in Figure 2. We begin our calculations by computing the a-th moment of y:

$$\begin{aligned}\mathbb{E}[y^a \mid \boldsymbol{x}] &= (-1)^a\left(\tfrac{\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}\right) + (1)^a\tfrac{1}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})} \\ &= \begin{cases} 1, & \text{for even a's} \\ \frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}, & \text{for odd a's.}\end{cases} \end{aligned} \quad (5)$$

The entries in $\boldsymbol{\mu}_{\boldsymbol{s}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}}$ are indexed by the non-negative integer exponents a, b, c, d such that $a + b = m, c + d = m, m \le 2$. Therefore, for all i, jth co-ordinates in $\boldsymbol{x}$, the corresponding entries in $\boldsymbol{\mu}_{\boldsymbol{s}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}}$ can be populated using Equation 5 as below:

1. $\mathbb{E}[x_i^a y^a x_j^b y^b] = \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b \mathbb{E}_{y|\boldsymbol{x}}[y^{a+b}]]$
$= \begin{cases}\mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b], & \text{for even a+b's} \\ \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b \frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}] & \text{for odd a+b's}\end{cases}$
2. $\mathrm{Cov}[x_i^a x_j^b y^{a+b}, x_k^c x_l^d y^{c+d}] = \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d y^{a+b+c+d}] - \mathbb{E}[x_i^a x_j^b y^{a+b}]\mathbb{E}[x_k^c x_l^d y^{c+d}]$

We can further simplify the covariance entries based on the parity of $a + b + c + d$.

(a) When both $a + b$ and $c + d$ are even.

$$\begin{aligned}&\mathrm{Cov}[x_i^a x_j^b y^{a+b}, x_k^c x_l^d y^{c+d}] \\ &= \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d] - \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b]\mathbb{E}_{\boldsymbol{x}}[x_k^c x_l^d]\end{aligned}$$

(b) When both $a + b$ and $c + d$ are odd.

$$\begin{aligned}&\mathrm{Cov}[x_i^a x_j^b y^{a+b}, x_k^c x_l^d y^{c+d}] \\ &= \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d\Big(\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}\Big)] \\ &- \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b\Big(\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}\Big)] \\ &\mathbb{E}_{\boldsymbol{x}}[x_k^c x_l^d\Big(\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}\Big)]\end{aligned}$$

(c) When $a + b$ is even and $c + d$ is odd.

$$\begin{aligned}&\mathrm{Cov}[x_i^a x_j^b y^{a+b}, x_k^c x_l^d y^{c+d}] \\ &= \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d\Big(\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}\Big)] \\ &- \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b]\mathbb{E}_{\boldsymbol{x}}\Big[x_k^c x_l^d\Big(\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}\Big)\Big]\end{aligned}$$

(d) Case $a + b$ is odd and $c + d$ is even follows identically from the previous case.

Taylor series expansion. The non-linear term $\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}$ makes the expectation in previous formulas intractable. We approximate this using a truncated Taylor series. The first two terms of the Taylor series approximation for $\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}$ are $-\frac{\boldsymbol{x}^T\boldsymbol{\theta}}{2} + \frac{(\boldsymbol{x}^T\boldsymbol{\theta})^3}{24}$. This approximation is reasonably accurate as long as $\boldsymbol{x}^T\boldsymbol{\theta} \in [-1, 1]$. We now approximate one of the expectations from the cases above:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d\Big(\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}\Big)] \\ &\approx \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d\Big[-\frac{\boldsymbol{x}^T\boldsymbol{\theta}}{2} + \frac{(\boldsymbol{x}^T\boldsymbol{\theta})^3}{24}\Big]] \\ &= -\frac{\mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d(\boldsymbol{x}^T\boldsymbol{\theta})]}{2} + \frac{\mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d(\boldsymbol{x}^T\boldsymbol{\theta})^3]}{24} \\ &= \frac{-\sum_{n=1}^d \theta_n \mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d x_n]}{2} + \\ &\frac{\sum_{\boldsymbol{e}:\sum_{o=1}^d e_o=3}\binom{d}{\boldsymbol{e}}(\prod_{n=1}^d \theta_n^{e_n})(\mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b x_k^c x_l^d \prod_{n=1}^d x_n^{e_n}])}{24}.\end{aligned}$$

In the derivation above, the term $(\boldsymbol{x}^T\boldsymbol{\theta})^3$ is expanded using the multinomial theorem. We can approximate

$$\boldsymbol{t}(\boldsymbol{x}, y) = \left[1, x_1 y, x_2 y, x_3 y, x_4 y, x_1^2 y^2, x_2^2 y^2, x_3^2 y^2, x_4^2 y^2, x_1 x_2 y^2, x_1 x_3 y^2, x_1 x_4 y^2, x_2 x_3 y^2, x_2 x_4 y^2, x_3 x_4 y^2\right]$$

Figure 2: An example of a second order (i.e. $m = 2$) approximate sufficient statistic $\boldsymbol{t}(\boldsymbol{x}, y)$ for logistic regression when $d = 4$.

$\mathbb{E}_{\boldsymbol{x}}[x_i^a x_j^b \left(\frac{1-\exp(\boldsymbol{x}^T\boldsymbol{\theta})}{1+\exp(\boldsymbol{x}^T\boldsymbol{\theta})}\right)]$ using similar calculations. The corresponding calculations for Poisson regression are found in the Supplement.

Evaluation of higher-order Gaussian moments. Now that our closed forms for the entries of $\boldsymbol{\mu_s}$ and $\boldsymbol{\Sigma_s}$ only include the sum of monomials, what remains is the actual evaluation of these Gaussian moments. We use the *Isserlis' theorem* (Wick, 1950) to compute these moments. This theorem presents even-degree moments of a zero-centered multivariate Gaussian variable as a sum of products of $\boldsymbol{\Sigma}$ entries.

Theorem 3.1. *Isserlis' theorem (Wick, 1950). Let $\boldsymbol{x} \sim \mathcal{N}^d(\boldsymbol{0}, \boldsymbol{\Sigma})$ be a $d-$dimensional random variable. Then*

$$\mathbb{E}[x_1, \cdots, x_d] = \sum_{p \in P_d^2} \prod_{\{i,j\} \in p} \mathbb{E}[x_i x_j] = \sum_{p \in P_d^2} \prod_{\{i,j\} \in p} \Sigma_{ij},$$

where d is assumed to be an even number and P is the set of all possible ways of partitioning $\{1, \cdots, d\}$ in to pairs $\{i, j\}$. For odd d's, $\mathbb{E}[x_1, \cdots, x_d] = 0$.

A few examples of moment calculations using Isserlis' theorem are found in the Supplement.

Computational complexity for moment population. For a normal approximation of a second order summary statistics, we require moments of degree $2, 4$, and 6 to populate $\frac{(d+1)^2(d+2)^2}{8} + \frac{(d+1)(d+2)}{2} \in \mathcal{O}(\frac{d^4}{8})$ unique entries in $\boldsymbol{\mu_s}, \boldsymbol{\Sigma_s}$. Under a normal model for $\boldsymbol{X}$, these can all be computed from the covariance of $\boldsymbol{X}$.

Proposition 3.2. *To evaluate a moment with degree $2k$, Theorem 3.1 generates $\frac{(2k-1)!}{2^{k-1}(k-1)!}$ partitions (number of summands) each containing k entries.*

Applying Proposition 3.2, we calculate that for a single degree $2, 4$ and 6 moment, we need to perform at-most 1, 6, and 45 unique multiplications. However, modern hardware can compute each moment in nearly a constant time with clever caching and indexing tricks. So a very loose upper bound on the order of operations performed in each iteration is $\mathcal{O}(d^4)$.

3.3. Satisfying DP

The last step in our model is to define $\Pr[\boldsymbol{z}|\boldsymbol{s}]$. In order to bound the global sensitivity, we assume that each input instance has a bounded L_2-norm, i.e. $||\boldsymbol{x}||_2 \leq R$.

3.3.1. Logistic regression

Sensitivity analysis. Recall that the approximate sufficient statistics for logistic regression contain both linear and quadratic terms. To this end, we define the functions $t_1 : \mathbb{R}^d \to \mathbb{R}^d$ and $t_2 : \mathbb{R}^d \to \mathbb{R}^{\binom{d+2}{2}}$ as

$$\begin{aligned} t_1(\boldsymbol{x}) &= \boldsymbol{x}, \\ t_2(\boldsymbol{x}) &= \left[x_1^2 \ \ldots \ x_d^2 \ \ \sqrt{2}x_1x_2 \ \ldots \ \sqrt{2}x_{d-1}x_d\right]^T. \end{aligned} \tag{6}$$

Using this notation, the approximate sufficient statistics are given as $[1, yt_1(\boldsymbol{x}), y^2 t_2(\boldsymbol{x})]$, which due to $y \in \{-1, 1\}$ yields $t(\boldsymbol{x}, y) = [1, yt_1(\boldsymbol{x}), t_2(\boldsymbol{x})]$. We consider a Gaussian mechanism where we release the linear and quadratic terms simultaneously. When compared to individual releases of $yt_1(\boldsymbol{x})$ and $t_2(\boldsymbol{x})$, this leads to a better utility.

Lemma 3.3. *Let t_1 and t_2 be defined as in* (6) *and let $\sigma_1, \sigma_2 > 0$. Let $\boldsymbol{s}_1 = \sum_{i=1}^N y_i t_1(\boldsymbol{x}_i)$ and $\boldsymbol{s}_2 = \sum_{i=1}^N t_2(\boldsymbol{x}_i)$. Consider the mechanism*

$$\mathcal{M}(\boldsymbol{s}) = \left[\begin{smallmatrix}\boldsymbol{s}_1 \\ \boldsymbol{s}_2\end{smallmatrix}\right] + \mathcal{N}\left(0, \begin{bmatrix}\sigma_1^2 I_d & 0 \\ 0 & \sigma_2^2 I_{d_2}\end{bmatrix}\right),$$

where $d_2 = \binom{d+2}{2}$. Assuming $||\boldsymbol{x}||_2 \leq R$, the tight (ϵ, δ)-DP for $\mathcal{M}$ is obtained by considering a Gaussian mechanism with noise variance σ_1^2 and sensitivity

$$\Delta = \sqrt{\frac{\sigma_2^2}{2\sigma_1^2} + 2R^2 + 2\frac{\sigma_1^2}{\sigma_2^2}R^4}.$$

Proof. Let $(\boldsymbol{x}, y), (\boldsymbol{x}', y')$ be the neighboring inputs. For the first order terms we have

$$\begin{aligned} ||yt_1(\boldsymbol{x}) - y't_1(\boldsymbol{x}')||_2^2 &= ||y\boldsymbol{x}||^2 + ||y'\boldsymbol{x}'||^2 - 2\langle y\boldsymbol{x}, y'\boldsymbol{x}'\rangle \\ &= ||\boldsymbol{x}||^2 + ||\boldsymbol{x}'||^2 - 2yy'\langle \boldsymbol{x}, \boldsymbol{x}'\rangle \\ &\leq 2R^2 - 2yy'\langle \boldsymbol{x}, \boldsymbol{x}'\rangle, \end{aligned} \tag{7}$$

and for the second order terms (see the Supplements)

$$\begin{aligned} ||t_2(\boldsymbol{x}) - t_2(\boldsymbol{x}')||_2^2 &= ||\boldsymbol{x}||^4 + ||\boldsymbol{x}'||^4 - 2\langle \boldsymbol{x}, \boldsymbol{x}'\rangle^2 \\ &\leq 2R^4 - 2\langle \boldsymbol{x}, \boldsymbol{x}'\rangle^2. \end{aligned} \tag{8}$$

For a single input $(\boldsymbol{x}, y)$, we have that

$$\begin{aligned} \mathcal{M}(\boldsymbol{x}) &\sim \begin{bmatrix} yt_1(\boldsymbol{x}) \\ t_2(\boldsymbol{x})\end{bmatrix} + \mathcal{N}\left(0, \begin{bmatrix}\sigma_1^2 I_d & 0 \\ 0 & \sigma_2^2 I_{d_2}\end{bmatrix}\right) \\ &\sim \begin{bmatrix} I_d & 0 \\ 0 & \frac{\sigma_2}{\sigma_1} I_{d_2}\end{bmatrix} \left(\begin{bmatrix} yt_1(\boldsymbol{x}) \\ \frac{\sigma_1}{\sigma_2} t_2(\boldsymbol{x})\end{bmatrix} + I_{d+d_2}\mathcal{N}(0, \sigma_1^2)\right). \end{aligned}$$

The matrix $\begin{bmatrix} I_d & 0 \\ 0 & \frac{\sigma_2}{\sigma_1} I_{d_2} \end{bmatrix}$ is a constant scaling and does not impact the sensitivity of $\mathcal{M}$. After discarding it, we see that the equivalent mechanism is $\mathcal{M}(\boldsymbol{x}) = \begin{bmatrix} yt_1(\boldsymbol{x}) \\ \frac{\sigma_1}{\sigma_2} t_2(\boldsymbol{x}) \end{bmatrix} + I_{d+d_2}\mathcal{N}(0, \sigma_1^2)$. Define

$$F(\boldsymbol{x}, y) = \begin{bmatrix} yt_1(\boldsymbol{x}) \\ \frac{\sigma_1}{\sigma_2} t_2(\boldsymbol{x}) \end{bmatrix}.$$

From (7) and (8) we see that for the sensitivity of the mechanism is given by,

$$\begin{aligned} \Delta &= \sqrt{\|F(\boldsymbol{x}, y) - F(\boldsymbol{x}', y')\|_2^2} \\ &\leq \sqrt{-2ct^2 - 2yy't + 2cR^4 + 2R^2}, \end{aligned}$$

where we denote $c = \frac{\sigma_1^2}{\sigma_2^2}$, $t = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$. The bound has its maximum at $t = -\frac{yy'}{2c}$, which leads to the claim. □

Corollary 3.4. *In the special case $R = 1$ and $\sigma_1 = \sigma_2 = \sigma$, by Lemma 3.3, the optimal (ϵ, δ) is obtained by considering the Gaussian mechanism with noise variance σ^2 and sensitivity $\Delta = \sqrt{4\frac{1}{2}}$.*

The general case. When using a higher order polynomial (i.e. $m > 2$), each $t_m(\boldsymbol{x})$ has to include all monomials of the form

$$x_{i_1}^{k_1} \cdots x_{i_{m'}}^{k_{m'}}$$

for all *combinations* of positive integers $(k_1, \ldots, k_{m'})$ such that $k_1 + \ldots k_{m'} = m$. Multiplying each monomial with the multinomial coefficient $\sqrt{\binom{m}{k_1, \ldots, k_{m'}}}$ and assuming $t_m(\boldsymbol{x})$ contains the monomials of order m, we find that (see the Supplements)

$$||t_m(\boldsymbol{x}) - t_m(\boldsymbol{x}')||_2^2 = ||\boldsymbol{x}||^{2m} + ||\boldsymbol{x}'||^{2m} - 2\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^m.$$

Finding the sensitivity upper bounds can then be carried out as in the case $m = 2$. For example, adding Gaussian noise with covariance $\sigma^2 I$ to all terms, we need to bound the sensitivity of the function $\begin{bmatrix} t_1(\boldsymbol{x})^T & \ldots & t_m(\boldsymbol{x})^T \end{bmatrix}^T$ for which we have

$$\Delta^2 \leq \sum\nolimits_{i=1}^{m} 2R^{2i} - 2\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^i.$$

Maximum of the right hand side is found by minimising the polynomial $\sum_{i=1}^m t^i$. For example, for $m = 6$, we find that the upper bound is attained for $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \approx -0.67$ and then $\Delta \approx \sqrt{12.72}$.

3.3.2. Poisson regression

In case of Poisson regression, in addition to $t_1(\boldsymbol{x})$ and $t_2(\boldsymbol{x})$, sufficient statistics requires releasing $y \in \mathbb{N}^{\geq 0}$. Let $\boldsymbol{s}_1 = \sum_{i=1}^N t_1(\boldsymbol{x}), \boldsymbol{s}_2 = \sum_{i=1}^N t_2(\boldsymbol{x}), \boldsymbol{s}_3 = \sum_{i=1}^N y_i t_1(\boldsymbol{x}_i), \boldsymbol{s}_4 = \sum_{i=1}^N y_i t_2(\boldsymbol{x}_i)$. Similarly to Lemma 3.3, we can show the following:

Lemma 3.5. *Let t_1, t_2 and $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4$ be defined as above and in (6) and let $\sigma_1, \sigma_2, \sigma_3, \sigma_4 > 0$. Suppose $||\boldsymbol{x}||_2 \leq R_x$ and $y \leq R_y$. Consider the mechanism*

$$\mathcal{M}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \\ \boldsymbol{s}_3 \\ \boldsymbol{s}_4 \end{bmatrix} + \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 I_d & 0 & 0 & 0 \\ 0 & \sigma_2^2 I_{d_2} & 0 & 0 \\ 0 & 0 & \sigma_3^2 I_{d_1} & 0 \\ 0 & 0 & 0 & \sigma_4^2 I_{d_2} \end{bmatrix}\right).$$

Then, the tight (ϵ, δ)-DP for $\mathcal{M}$ is obtained by considering a Gaussian mechanism with noise variance σ_1^2 and sensitivity

$$\Delta = \frac{2(c_2 + c_4)R_x^2 + c_3 + 1}{\sqrt{2(c_2 + c_4)}},$$

where $c_2 = \frac{\sigma_1^2}{\sigma_2^2}$, $c_3 = \frac{\sigma_1^2}{\sigma_3^2} R_y^2$ and $c_4 = \frac{\sigma_1^2}{\sigma_4^2} R_y^2$.

Proof. Similar to Lemma 3.3, we consider mechanism $\mathcal{M}$ for a single input $\{\boldsymbol{x}, y\}$.

$$\begin{aligned} \mathcal{M}(\boldsymbol{x}) &\sim \begin{bmatrix} t_1(\boldsymbol{x}) \\ t_2(\boldsymbol{x}) \\ yt_1(\boldsymbol{x}) \\ yt_2(\boldsymbol{x}) \end{bmatrix} + \mathcal{N}\left(0, \begin{bmatrix} \sigma_1^2 I_d & 0 & 0 & 0 \\ 0 & \sigma_2^2 I_{d_2} & 0 & 0 \\ 0 & 0 & \sigma_3^2 I_{d_1} & 0 \\ 0 & 0 & 0 & \sigma_4^2 I_{d_2} \end{bmatrix}\right) \\ &\sim \begin{bmatrix} I_d & 0 & 0 & 0 \\ 0 & \frac{\sigma_2 I_{d_2}}{\sigma_1} & 0 & 0 \\ 0 & 0 & \frac{\sigma_3 I_d}{\sigma_1} & 0 \\ 0 & 0 & 0 & \frac{\sigma_4 I_{d_2}}{\sigma_1} \end{bmatrix} \left(\begin{bmatrix} t_1(\boldsymbol{x}) \\ \frac{\sigma_1 t_2(\boldsymbol{x})}{\sigma_2} \\ \frac{\sigma_1 yt_1(\boldsymbol{x})}{\sigma_3} \\ \frac{\sigma_1 yt_2(\boldsymbol{x})}{\sigma_4} \end{bmatrix} + \mathcal{N}(0, \sigma_1^2 I) \right). \end{aligned}$$

Removing the constant scaling, we see that it is equivalent to consider the mechanism

$$\mathcal{M}(\boldsymbol{x}) = \begin{bmatrix} t_1(\boldsymbol{x}) \\ \frac{\sigma_2}{\sigma_1} t_2(\boldsymbol{x}) \\ \frac{\sigma_1}{\sigma_3} yt_1(\boldsymbol{x}) \\ \frac{\sigma_1}{\sigma_4} yt_2(\boldsymbol{x}) \end{bmatrix} + \mathcal{N}(0, \sigma_1^2 I).$$

Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ such that $||\boldsymbol{x}||_2 \leq R_x$ and $||\boldsymbol{x}'||_2 \leq R_x$ be the neighboring inputs. Define

$$F(\boldsymbol{x}) = \begin{bmatrix} t_1(\boldsymbol{x}) \\ \frac{\sigma_1}{\sigma_2} t_2(\boldsymbol{x}) \\ \frac{\sigma_1}{\sigma_3} yt_1(\boldsymbol{x}) \\ \frac{\sigma_1}{\sigma_4} yt_2(\boldsymbol{x}) \end{bmatrix}.$$

Since

$$\begin{aligned} ||t_1(\boldsymbol{x}) - t_1(\boldsymbol{x}')||_2^2 &= ||\boldsymbol{x}||^2 + ||\boldsymbol{x}'||^2 - 2\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \\ &\leq 2R_x^2 - 2\langle \boldsymbol{x}, \boldsymbol{x}' \rangle, \\ ||t_2(\boldsymbol{x}) - t_2(\boldsymbol{x}')||_2^2 &= ||\boldsymbol{x}||^4 + ||\boldsymbol{x}'||^4 - 2\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^2 \\ &\leq 2R_x^4 - 2\langle \boldsymbol{x}, \boldsymbol{x}' \rangle^2, \end{aligned}$$

we have that

$$\begin{aligned}\|F(\boldsymbol{x}) - F(\boldsymbol{x}')\|^2 &= \|t_1(\boldsymbol{x}) - t_1(\boldsymbol{x}')\|^2 \\ &+ \frac{\sigma_1^2}{\sigma_2^2}\|t_2(\boldsymbol{x}) - t_2(\boldsymbol{x}')\|^2 \\ &+ \frac{\sigma_1^2}{\sigma_3^2}\|yt_1(\boldsymbol{x}) - y't_1(\boldsymbol{x}')\|^2 \\ &+ \frac{\sigma_1^2}{\sigma_4^2}\|yt_2(\boldsymbol{x}) - y't_2(\boldsymbol{x}')\|^2 \\ &\leq -2t + 2R_x^2 - 2c_2t^2 + 2c_2R_x^4 \\ &- 2c_3t + 2c_3R_x^2 - 2c_4t^2 + 2c_4R_x^4,\end{aligned}$$

where $c_2 = \frac{\sigma_1^2}{\sigma_2^2}$, $c_3 = \frac{\sigma_1^2}{\sigma_3^2}R_y^2$, $c_4 = \frac{\sigma_1^2}{\sigma_4^2}R_y^2$ and $t = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$. The right hand side of this inequality has its maximum at

$$t = -\frac{1 + c_3}{2(c_2 + c_4)}$$

which shows that

$$\|F(\boldsymbol{x}) - F(\boldsymbol{x}')\|^2 \leq \frac{(2(c_2 + c_4)R_x^2 + c_3 + 1)^2}{2(c_2 + c_4)}.$$

This gives sensitivity bound of the claim. □

Corollary 3.6. *In the special case $R_x = 1$ and $\sigma_1 = \ldots = \sigma_4 = \sigma$, the optimal (ϵ, δ) is obtained by considering the Gaussian mechanism with noise variance σ^2 and sensitivity $\Delta = \sqrt{4\frac{1}{2}(1 + R_y^2)}$.*

4. Experiments

In this Section we present experiments on logistic regression. Additional experiments on Poisson regression are included in the Supplement.

4.1. Default settings and implementation

Throughout our experiments, we use the first two central moments of joint $(\boldsymbol{X}, y)$ as the summary statistics. Both noise-aware and non-private baseline models for logistic regression are specified in Stan (Carpenter et al., 2017) using its Python interface. We use Stan's default *No-U-Turn* (Homan and Gelman, 2014) sampler, which is a variant of Hamiltonian Monte Carlo. We run 4 Markov chains in parallel and discard the first 50% as warm-up samples. We fix $R = 1$ to use Corollary 3.4. The Gelman-Rubin convergence statistic (Brooks and Gelman, 1998) was consistently below 1.1 for all Stan experiments. For brevity we provide comparisons only for the $\boldsymbol{\theta}$'s in our figures.

Datasets. We use Adult (Blake and Merz, 1998) and Diabetes (Kahn, 1994) datasets from UCI repository as these are standard and easy to explain. To reduce the training time to be more manageable, we chose 6/13 and 14/20 features from Adult and Diabetes datasets. The selected features had a significant effect on the target variable.

Since error in the centralized model is proportional to $\mathcal{O}(\frac{1}{N})$, we do not want to attribute higher accuracy from our model to larger sample size (N). Therefore, for the Adult dataset, we trained our model on randomly sampled 8000/40,000 records for a fair evaluation.

Setting priors for model parameters $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$. For the data covariance matrix $\boldsymbol{\Sigma}$ we gave a scaled LKJ (Lewandowski et al., 2009) prior. We scale a positive definite correlation matrix from the LKJ correlation distribution of shape $\eta = 2$ from both sides with a diagonal matrix with $\mathcal{N}(0, 2.5)$ distributed diagonal entries. The probabilistic model is:

$$\begin{aligned}&\Omega \sim \text{LKJ}(2), \quad \boldsymbol{\tau} \sim \mathcal{N}(\mathbf{0}, 2.5 \cdot \mathbf{I}), \\ &\boldsymbol{\Sigma} = \text{diag}(\boldsymbol{\tau})\, \Omega\, \text{diag}(\boldsymbol{\tau}).\end{aligned}$$

In order to prevent the inference from sampling $\boldsymbol{\theta}$'s with large magnitudes, we gave the regression coefficients' orientation a uniform prior, and the squared norm a truncated Chi-square prior. We treat the upper-bound for the truncation as a hyper-parameter, which was set to 2 or 3 times the square of non-private $\boldsymbol{\theta}'s$ norm. The exact probabilistic model is:

$$\begin{aligned}&\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \rho \sim \chi^2(d) \\ &\boldsymbol{\theta} = \sqrt{\max(\rho, s)}\frac{\mathbf{p}}{||\mathbf{p}||_2}.\end{aligned}$$

In reality, the upper bound s could be obtained in a DP way or its approximate value could be known from domain expertise. The question of designing a better prior that does not require such truncation bound is left as a future exercise.

Private Baseline. We compare our method with a Python implementation of general-purpose DP posterior inference DP-SGLD (Wang et al., 2015; Li et al., 2019b), using the Fourier accountant (Koskela et al., 2020) for tight DP accounting. In each iteration, DP-SGLD samples a mini-batch of records and perturb the aggregated (and clipped) gradients of the posterior distribution. Updated weights at the end of each iteration can be treated as posterior samples. The total privacy budget in DP-SGLD scales proportional to the number of iterations, whereas, we only pay a small upfront privacy cost for perturbing the sum of sufficient statistics and enjoy posterior samples for free. DP-SGLD may provide unsatisfactory utility for strong privacy parameters regimes because the noise (which is already quite large for smaller privacy budgets) is further amplified due to uncertainty induced by batch sub-sampling, destroying signal in the gradients. We run DP-SGLD with batch-size $\sqrt{N}$ (as suggested by Abadi et al., 2016) for 10,000 iterations and discard the first 6000 samples as burn-in.

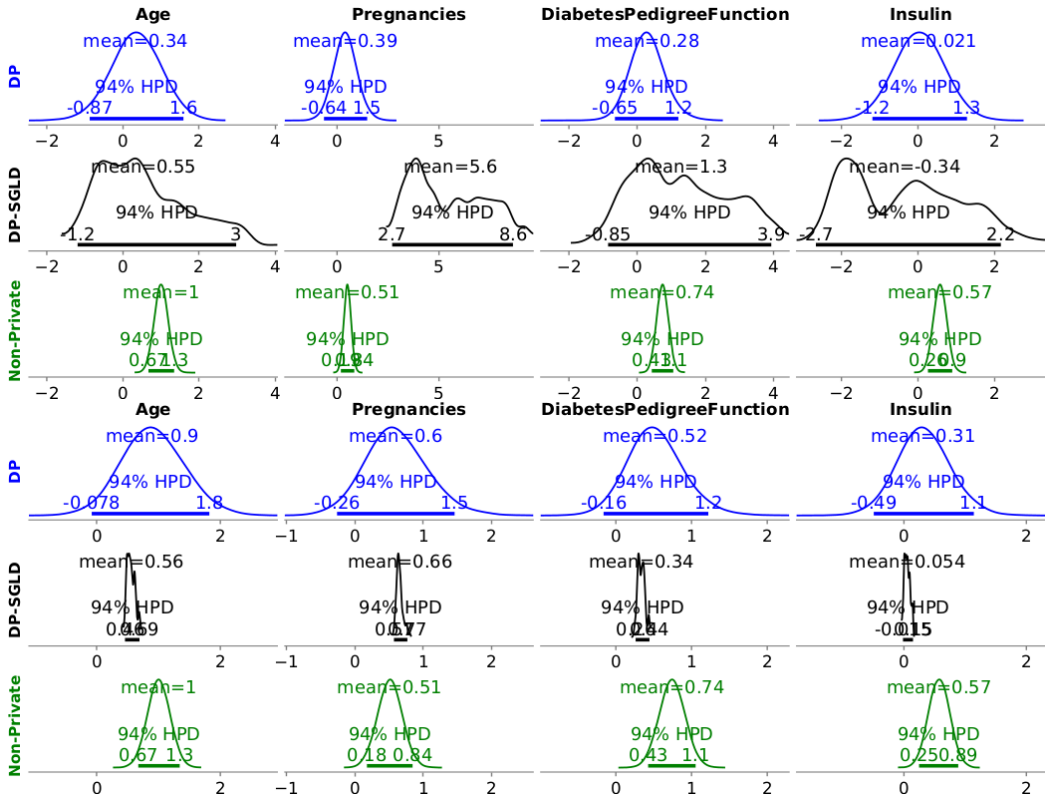

Figure 3: Comparison of differentially private posteriors from our model (blue) and DP-SGLD (black) with non-private posteriors for $\boldsymbol{\theta}$ for the UCI Diabetes (Kahn, 1994) dataset ($N = 758$) for $\epsilon = 0.1, \delta = 10^{-5}$ (top) and $\epsilon = 0.3, \delta = 10^{-5}$ (bottom) after $10,000$ iterations. The batch size and the learning rate chosen for DP-SGLD were 28 and 10^{-1}. The posteriors from DP-SGLD are more biased and either exhibit a much higher variance or fail to quantify the expected uncertainty.

4.2. Posterior recovery

We trained our logistic regression model on 8000 pre-processed random records from the UCI Adult dataset with features age, workclass, education, martial-status, occupation, relationship, and gender to predict whether a person's income exceeds \$50k. Furthermore, we also train on abridged version of a much smaller pre-processed UCI Diabetes dataset with dimensions age, number of pregnancies, diabetes pedigree function, and insulin levels to predict whether a person has diabetes or not. Figures 3 and 4 illustrate the outcome of these experiments for DP parameters in a strong-privacy regime. We verify that for both datasets, private posteriors from our model are close to the non-private posteriors. The posteriors obtained from DP-SGLD are highly variable, ranging from highly biased to hugely underconfident to massively overconfident.

Looking at inference of statistical significance of regression coefficients based on zero not being included in the high probability intervals, extra noise from DP causes DP-GLM to just miss significance at $\epsilon = 0.3$ with Diabetes (Figure 3) but perfectly match non-private results at $\epsilon = 0.1$ with Adult (Figure 4). DP-SGLD results are highly unstable.

4.3. Varying privacy requirements

We now test the accuracy of our methods against a verity of privacy settings. Figure 5 compares the posterior empirical cumulative distribution functions (CDFs) of private and non-private $\boldsymbol{\theta}$'s for a synthetic dataset with a randomly sampled positive definite non-identity co-variance matrix with true $\boldsymbol{\theta}$ as $[-0.9, -0.5, 0.3]$. We see that the private and non-private CDFs are almost overlapping for $\epsilon > 0.1$. In the right-most column, we additionally plot the Kolmogorov-Smirnov scores (maximum absolute difference between two CDFs) for 10 equally spaced ϵ values in the range [0.001,1.1]. Once again, we note a general non-increasing trend.

These results demonstrate that our model is accurate for datasets with small true $\boldsymbol{\theta}$'s for moderate to large sample sizes even when privacy requirements are strict.

5. Limitations

In additional internal experiments on synthetic datasets, we studied accuracy as a function of L_2-norm of true $\boldsymbol{\theta}$, fixing other parameters, and found that the accuracy decreases sharply when $||\boldsymbol{\theta}||_2 \geq 3$, when using the proposed approxi-

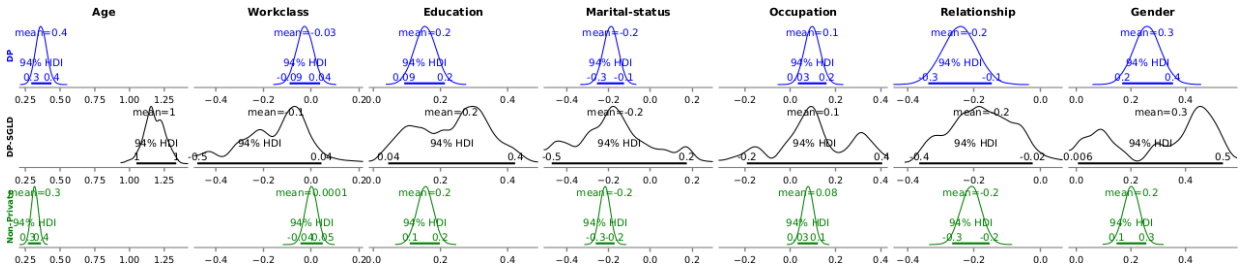

Figure 4: Comparison of differentially private posteriors from our model (DP) and DP-SGLD with non-private posteriors for $\boldsymbol{\theta}$ for randomly sampled 8000 records in the UCI Adult dataset (Blake and Merz, 1998) for $\epsilon = 0.1, \delta = 10^{-5}$ after $10,000$ iterations. The batch size and the learning rate chosen for DP-SGLD were 89 and 10^{-2}. DP-SGLD posteriors are biased and overestimate the uncertainty even at such large sample size.

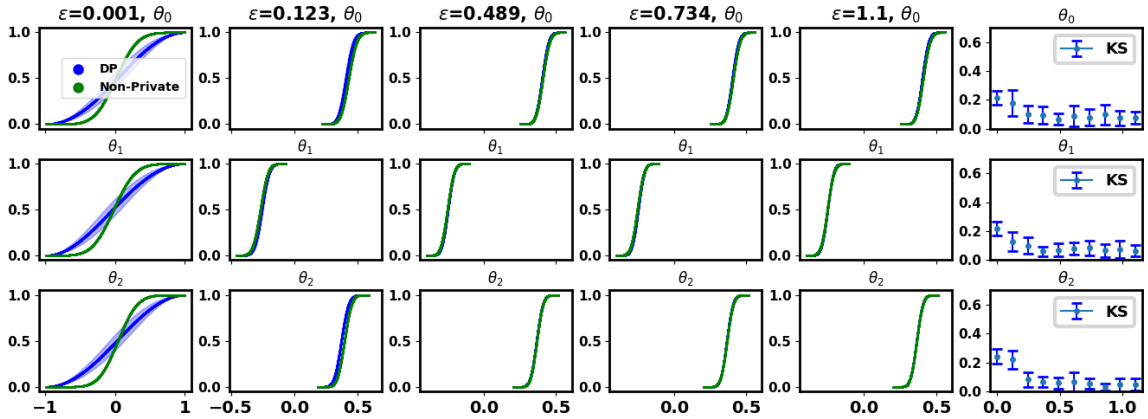

Figure 5: Comparison of differentially private and non-private empirical CDFs for $\boldsymbol{\theta}$'s posteriors for a synthetic dataset with $N = 1000$ after $30,000$ iterations and 20 repetitions while fixing $\delta = 10^{-5}$ for various ϵ values. The rightmost column shows the Kolmogorov-Smirnov scores between non-private and private empirical CDFs for various ϵ values.

mations. We suspect this is due the truncation in the Taylor series, used to reduce the computational complexity. The bound on theta norm essentially means that the predicted probabilities can differ by at most a modest number of logistic units. While this may sound limiting, it may not be a serious issue for real datasets, where the signal is not very strong. With a proper prior, the model would most likely simply underfit in such cases. The implications on inference of the signs of the regression coefficients would also likely be limited.

6. Concluding Remarks

This work formulates a noise-aware model for GLMs for performing DP Bayesian inference and demonstrates its efficacy for datasets with regression coefficients of small magnitudes. Our method combines a normal approximation based on the central limit theorem with moment matching for perturbed low order data moments. We carry out a sensitivity analysis for the DP mechanisms which gives tight bounds and leads to high utility. This is also reflected in the experimental results on the logistic regression. Our sensitivity analysis also shows that we can increase utility by simultaneously releasing the linear and quadratic terms. Since computation of approximate sufficient statistics is a transformation of data, it seems possible to develop similar noise-aware models in distributed learning scenarios such as federated learning (Kairouz et al., 2021) and local differential privacy (Cormode et al., 2018).

Acknowledgements

This work was sponsored by the Academy of Finland; Grants 325573, 325572, 313124, 335516, and Flagship programme: Finnish Center for Artificial Intelligence, FCAI, as well as by the Strategic Research Council at the Academy of Finland (Grant 336032). We are grateful to the Aalto Science-IT project for their computational resources. We also wish to thank Daniel Sheldon for useful discussions and for the suggestion to use the PASS-GLM framework.

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016*, pages 308–318. ACM.

Balle, B. and Wang, Y. (2018). Improving the Gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 403–412. PMLR.

Barrientos, A. F., Reiter, J. P., Machanavajjhala, A., and Chen, Y. (2019). Differentially private significance tests for regression coefficients. *Journal of Computational and Graphical Statistics*, 28(2):440–453.

Bassily, R., Smith, A. D., and Thakurta, A. (2014). Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014*, pages 464–473. IEEE Computer Society.

Bernstein, G. and Sheldon, D. R. (2018). Differentially private Bayesian inference for exponential families. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018,*, pages 2924–2934.

Bernstein, G. and Sheldon, D. R. (2019). Differentially private Bayesian linear regression. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 523–533.

Blake, C. and Merz, C. (1998). UCI machine learning repository.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, 76(1):1–32.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. (2011). Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109.

Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., and Wang, T. (2018). Privacy at scale: Local differential privacy in practice. Tutorial at SIGMOD and KDD.

Dimitrakakis, C., Nelson, B., Mitrokotsa, A., and Rubinstein, B. I. (2014). Robust and private Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 291–305. Springer.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. D. (2006). Calibrating noise to sensitivity in private data analysis. In Halevi, S. and Rabin, T., editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer.

Dwork, C. and Smith, A. D. (2010). Differential privacy for statistics: What we know and what we want to learn. *J. Priv. Confidentiality*, 1(2).

Foulds, J., Geumlek, J., Welling, M., and Chaudhuri, K. (2016). On the theory and practice of privacy-preserving Bayesian data analysis. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 192–201.

Heikkilä, M. A., Jälkö, J., Dikmen, O., and Honkela, A. (2019). Differentially private Markov chain Monte Carlo. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019*, pages 4115–4125.

Homan, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623.

Honkela, A., Das, M., Nieminen, A., Dikmen, O., and Kaski, S. (2018). Efficient differentially private learning improves drug sensitivity prediction. *Biology direct*, 13(1):1.

Huggins, J. H., Adams, R. P., and Broderick, T. (2017). PASS-GLM: polynomial approximate sufficient statistics for scalable Bayesian GLM inference. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017*, pages 3611–3621.

Iyengar, R., Near, J. P., Song, D., Thakkar, O., Thakurta, A., and Wang, L. (2019). Towards practical differentially private convex optimization. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 299–316.

Jain, P. and Thakurta, A. G. (2014). (near) dimension independent risk bounds for differentially private learning. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 476–484. JMLR.org.

Jälkö, J., Dikmen, O., and Honkela, A. (2017). Differentially private variational inference for non-conjugate models. In *Uncertainty in Artificial Intelligence 2017, Proceedings of the 33rd Conference (UAI),*.

Kahn, M. (1994). UCI machine learning repository.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. (2021). Advances and open problems in federated learning. In *Foundations and Trends in Machine Learning*.

Kifer, D. and Machanavajjhala, A. (2011). No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD '11, page 193–204. Association for Computing Machinery.

Kifer, D., Smith, A. D., and Thakurta, A. (2012). Private convex optimization for empirical risk minimization with applications to high-dimensional regression. In *COLT 2012 - The 25th Annual Conference on Learning Theory*, volume 23 of *JMLR Proceedings*, pages 25.1–25.40. JMLR.org.

Koskela, A., Jälkö, J., and Honkela, A. (2020). Computing tight differential privacy guarantees using FFT. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020,*, volume 108 of *Proceedings of Machine Learning Research*, pages 2560–2569. PMLR.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989 – 2001.

Li, B., Chen, C., Liu, H., and Carin, L. (2019a). On connecting stochastic gradient MCMC and differential privacy. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS, 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 557–566. PMLR.

Li, B., Chen, C., Liu, H., and Carin, L. (2019b). On connecting stochastic gradient MCMC and differential privacy. In *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 557–566. PMLR.

McSherry, F. and Mironov, I. (2009). Differentially private recommender systems: Building privacy into the Netflix prize contenders. In IV, J. F. E., Fogelman-Soulié, F., Flach, P. A., and Zaki, M. J., editors, *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2009*, pages 627–636. ACM.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Park, M., Foulds, J., Chaudhuri, K., and Welling, M. (2020). Variational Bayes in private settings (VIPS). *Journal of Artificial Intelligence Research*, 68:109–157.

Pihur, V., Korolova, A., Liu, F., Sankuratripati, S., Yung, M., Huang, D., and Zeng, R. (2018). Differentially-private "draw and discard" machine learning. *CoRR*, abs/1807.04369.

Sheffet, O. (2017). Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning*.

Smith, A. (2008). Efficient, differentially private point estimators. *CoRR*, abs/0809.4794.

Vu, D. and Slavkovic, A. (2009). Differential privacy for clinical trial data: Preliminary evaluations. In *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 138–143. IEEE.

Wang, D., Chen, C., and Xu, J. (2019). Differentially private empirical risk minimization with non-convex loss functions. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019*, volume 97 of *Proceedings of Machine Learning Research*, pages 6526–6535. PMLR.

Wang, D., Zhang, H., Gaboardi, M., and Xu, J. (2021). Estimating smooth GLM in non-interactive local differential privacy model with public unlabeled data. In *Algorithmic Learning Theory, 16-19 March 2021, Virtual Conference, Worldwide*, volume 132 of *Proceedings of Machine Learning Research*, pages 1207–1213. PMLR.

Wang, Y.-X. (2018). Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Wang, Y.-X., Fienberg, S., and Smola, A. (2015). Privacy for free: Posterior sampling and stochastic gradient Monte

Carlo. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 2493–2502.

Wick, G. C. (1950). The evaluation of the collision matrix. *Phys. Rev.*, 80:268–272.

Williams, O. and McSherry, F. (2010). Probabilistic inference and differential privacy. In *Advances in Neural Information Processing Systems*, pages 2451–2459.

Wu, X., Li, F., Kumar, A., Chaudhuri, K., Jha, S., and Naughton, J. F. (2017). Bolt-on differential privacy for scalable stochastic gradient descent-based analytics. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017*, pages 1307–1322. ACM.

Yıldırım, S. and Ermiş, B. (2019). Exact MCMC with differentially private moves. *Statistics and Computing*, 29(5):947–963.

Zhang, J., Zhang, Z., Xiao, X., Yang, Y., and Winslett, M. (2012). Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 5(11):1364–1375.

Zhang, J., Zheng, K., Mou, W., and Wang, L. (2017). Efficient private ERM for smooth objectives. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3922–3928. ijcai.org.

Zhang, Z., Rubinstein, B. I. P., and Dimitrakakis, C. (2016). On the differential privacy of Bayesian inference. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 2365–2371. AAAI Press.