A differential privacy workflow for inference of parameters in the Rasch model*

Teresa Anna Steiner, David Enslev Nyrnberg, and Lars Kai Hansen

Department of Applied Mathematics and Computer Science,
Technical University of Denmark B324, DK-2800 Kongens Lyngby, Denmark
s170063@student.dtu.dk, s123997@student.dtu.dk, lkai@dtu.dk

Abstract. The Rasch model is used to estimate student performance and task difficulty in simple test scenarios. We design a workflow for enhancing student feedback by release of difficulty parameters in the Rasch model with privacy protection using differential privacy. We provide a first proof of differential privacy in Rasch models and derive the minimum noise level in objective perturbation to guarantee a given privacy budget. We test the workflow in simulations and in two real data sets.

Keywords: Rasch model · Differential privacy · Student feedback.

1 Introduction

Protection of private information is a key democratic value and so-called 'privacy by design' is core to the new European General Data Protection Regulatory (GDPR) [6].

> *Privacy by design as a concept has existed for years now, but it is only just becoming part of a legal requirement with the GDPR. At its core, privacy by design calls for the inclusion of data protection from the onset of the designing of systems, rather than an addition. More specifically - 'The controller shall..implement appropriate technical and organisational measures..in an effective way.. in order to meet the requirements of this Regulation and protect the rights of data subjects'.*

Differential privacy is one such tool allowing a 'controller' to train a machine learning model on inherently private data, but with mathematical bounds on the actual loss of privacy when results are released [4,8]. Differential privacy is based on randomized algorithms using noise to reduce the probability of breach of privacy. The key idea is to secure that the randomized output does not in a significant way depend on any of the possible data subjects' data.

Educational technology is important to serve the increasing needs for lifelong learning [10]. Learning processes and tests are typically highly personal, yet, significant gains are conceivable from integrating and sharing such information.

* This work was supported by the Danish Innovation Foundation through the Danish Center for Big Data Analytics and Innovation (DABAI).

Sharing could, e.g., be used to provide more detailed feedback on tests, hence, enhance the learning process. The basic question addressed in the present work is if differentially private machine learning methods can be used to provide more detailed feedback on students' tests, while still respecting the privacy of the individual students.

The concept is illustrated in Figure 1. The use case concerns a class of students each answering a set of tasks. The teacher ('the controller') can by conventional means estimate each students performance and release this information in private to the given student. Here our aim is in addition to share a difficulty score for each task and investigate whether it is feasible to compute this score in a differentially private manner, hence, with mathematical bounds on the amount of individual information leaked by releasing the difficulty scores. Given the privatized difficulty scores, every student can then use their sensitive data to estimate their own ability scores and probabilities of passing a subject. The paper is organized as follows. We first present the differential privacy model

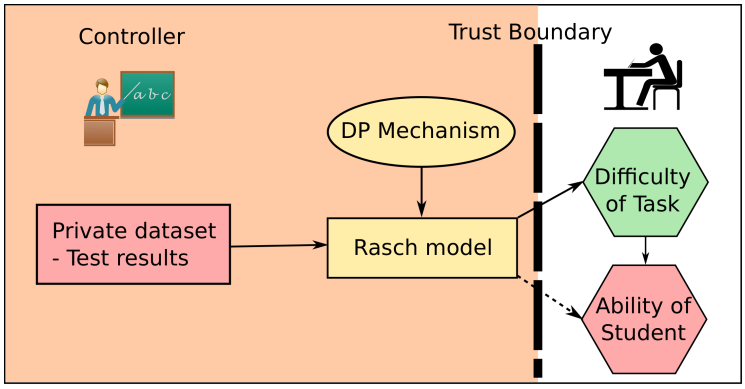


Fig. 1: Concept of the differentially private Rasch model and its use in enhanced feedback in teaching.

in the educational technology context. Student performance and test scores are inferred using item response theory ('Rasch model'). Next, we investigate the loss of accuracy when privacy is enforced at various privacy budgets. Finally, we demonstrate viability in a real world data set. The proof of the differential privacy mechanism (so-called objective perturbation) is provided in an appendix. The original contributions can be summarized as follows: *1) We define a workflow and model for privacy preserving machine learning of student performance and task difficulty. 2) We show by simulation that the student performance is well estimated for each student separately. 3) We give the first proof of differential privacy for the Rasch model based on so-called objective perturbation. 4) We derive the minimum noise level that allows us to release the task difficulty at a given privacy budget.*

All code can be found at the following github repository.

2 Preliminaries

The concept of differential privacy is based on a privacy parameter or 'budget' ϵ. The algorithm $\mathcal{A}$ is ϵ-differentially private if for all data sets D_1 and D_2 that differ by a single entry (data subject)

$$P[\mathcal{A}(D_1) = w] \leq P[\mathcal{A}(D_2) = w] \exp(\epsilon), \tag{1}$$

where P is the probability taken over the randomness used by the algorithm $\mathcal{A}$, and w is the output of the algorithm. The privacy budget quantifies how likely it is that a well-informed adversarial can determine whether a specific data subject participated or not. The randomness is added to the algorithm to hinder this identification. This randomness is achieved e.g. through the addition of noise. This noise is scaled as $\Delta f/\epsilon$ where Δf is the sensitivity of a function f, defined as

$$\Delta f = \max \|f(D_1) - f(D_2)\|_1, \tag{2}$$

where again D_1 and D_2 differ in a single entry [8].

The data sets we work with are arranged as a (number of students N) $\times$ (number of test items I) matrix X, and every entry stands for a right or wrong answer of a student to an item. In this work, we consider differential privacy in the sense that the output from our model should not depend much on whether a particular student is in the set or not. That is, for X and $\tilde{X}$ with $X_{n,i} = \tilde{X}_{n,i}$ for all $n = 1, \ldots, N-1$ and $i = 1, \ldots, I$, so two data sets that can differ in at most one row (corresponding to one student), we want to achieve

$$\frac{P(w|X)}{P(w|\tilde{X})} \leq e^{\epsilon}, \tag{3}$$

where w is the output of the algorithm.

The Rasch model is a simple example of item response theory (IRT). IRT concerns performance testing quantifying the probability that students can answer a specific test task in terms of the difficulty of the task and their general ability. The model is similar to the logistic regression and used to estimate the probability of passing a task

$$P(X_{n,i} = 1|\beta_n, \delta_i) = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)}, \tag{4}$$

where β_n models the ability of student n and δ_i is the difficulty of task i. X_{ni} is a dichotomous observation of a student's (n) correct or incorrect answer to a task (i), where 1 is a correct answer, and 0 is an incorrect answer. The model is generated by estimating δ_i and β_n from the results of a particular test. The parameters are estimated by maximizing the likelihood

$$\Lambda = \prod_n \prod_i \frac{e^{x_{ni}(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}}. \tag{5}$$

In our work, we will introduce differentially private methods of estimating β and δ, but only the δ-values will be released to the public. Every student can then, based on the public δ and their personal results, estimate their own parameter β_n.

3 Methods

We will implement this workflow, i.e., release differentially private δ parameters and then re-estimate the parameters β_n on $X_{n,:}$, by first calculating both parameters with differentially private algorithms, assuring a private δ, then re-estimating each β_n given that δ. The re-estimation was also proposed in Choppin [3]. We will investigate the impact of the re-estimation compared to the global parameter estimation in section 4.

We consider two different methods for constructing a differentially private Rasch Model. The first one is the objective perturbation, first introduced by Chaudhuri and Monteleoni for logistic regression [1], and analyzed in more detail by Chaudhuri et al. [2]. For the differentially private Rasch model, we use a slightly modified version of the objective perturbation, and prove that it is ϵ-differentially private as defined in equation (3).

We will also consider a simple reference method based on perturbing sufficient statistics as discussed in [7]. By adding enough noise to the sufficient statistics, we may release them as differentially private. Then any algorithm based on these statistics will be differentially private. The latter observation follows from the post-processing theorem [5]. For the Rasch model, the sufficient statistics are $r_n = \sum_i^I X_{n,i}$ and $s_i = \sum_n^N X_{n,i}$, since those are all that is needed for minimizing the regularized objective function. We add noise to the vectors r and s, scaled with their sensibilities: if student n in the data set is changed, r_n will change by at most I, and so the L_1 norm of r will change by at most I. Similarly, s_i can change by at most 1 for every i, so again, the L_1 norm of s changes by at most I. So the noise we add to both vectors is scaled with I/ϵ as in [7]. Since making the sufficient statistics differentially private is more general than objective perturbation, and does not use the specific structure of the learning algorithm, we expect it to be weaker (adding more noise).

As we notice below, the objective perturbation approach effectively perturbs the sufficient statistics with noise scaling as $\sqrt{I}/\epsilon$, while direct perturbation of the sufficient statistics requires noise scaling as I/ϵ [7]. In the following we will show the costs of the less favorable scaling.

An important aspect of learning the Rasch model parameters is to quantify the available prior information. In the educational technology context we could imagine substantial prior information to be present from earlier exams etc. Here we for simplicity assume that the test difficulties and student abilities both follow normal distributions, hence, we add a regularization term $\frac{\lambda}{2} w^T w$, where w is the entire parameter vector $w = [\beta\ \delta]$, to the Log likelihood function in equation (5). A discussion on how to estimate parameter λ while preserving differential privacy can be found below.

Algorithm 1 describes the details of the modified objective perturbation algorithm for the differentially private Rasch model.

Algorithm 1:

- Draw vector b with dimension I from density function $h(b) \propto \exp\left(-\frac{\epsilon}{\sqrt{I}}\right)$. To do that, draw direction uniformly at random from the I dimensional unit sphere, and draw the norm from $\Gamma(I, \frac{\sqrt{I}}{\epsilon})$.
- Minimize

$$F(\beta, \delta) = \sum_{n}^{N} \sum_{i}^{I} \log(1 + \exp(\beta_n - \delta_i)) + \sum_{i}^{I} \delta_i \sum_{n}^{N} X_{n,i} - \sum_{n}^{N} \beta_n \sum_{i}^{I} X_{n,i} + \frac{\lambda}{2}(\beta^T \beta + \delta^T \delta) + \sum_{i}^{I} b_i \delta_i \tag{6}$$

with respect to β, δ for $\lambda > 0$.

Theorem 1. *Algorithm 1 is ϵ-differentially private.*

The proof follows the strategy developed in Chaudhuri et al. [1,2], details are found in the appendix.

Naive approaches of estimating the regularization parameters by f.e. cross validation can lead to a loss of privacy, since information is leaked by every evaluation of the model, and this information accumulates. Chaudhuri et al. [2] propose two different methods on how to handle this, which we can also apply. The first one is to use a small publicly available data set that follows roughly the same distribution for the estimation of λ. The second one is an algorithm which splits the data set into $(m + 1)$ subsets, calculates the model for m different guesses of λ on respective subsets, and evaluates the model on the last one. Then, based on the number z_i of errors made by the i^{th} estimate, λ_i is chosen with probability

$$\frac{e^{-\epsilon z_i/2}}{\sum_{j=1}^{m} e^{-\epsilon z_j/2}}. \tag{7}$$

For our model, we split the data into subsets of students. For the m different estimates of δ, we first compute the β values for the last subset and the corresponding probabilities. We then compare the rounded probabilities to the actual 0 or 1 entries in the $(m+1)^{\text{st}}$ subset in the data.

4 Experiments

Experiments were run for simulated data, real data from M. Vahdat et al. [9], as well as a data set from a course at the Technical University of Denmark.

The first experiment is run on simulated data and compares the results of calculating both β and δ globally to re-estimating β.

The next experiments compare the performance of the two methods for introducing privacy, the objective perturbation and sufficient statistics. We use correlation coefficients between the estimated probabilities and the true probabilities (i.e., the ones used to simulate the data) resp. the non-private estimates with 95% confidence intervals. Further, we show test misclassification rates (how well do we predict if a student passes a test) on a new data set drawn from the same distribution as the training data. Experiments were run for the two real data sets. The first one had to be modified to fit the Rasch model, so the answers were rounded to 1 or 0 depending on whether half of the points for a question were scored. The DTU data set are the results of a multiple choice test, so the original data can be used. For the real data sets, we use bootstrapping with 1000 samples to calculate confidence intervals.

We used MATLAB's *fminunc* function for minimizing the objective function in all experiments - with the following settings:

```
options = optimoptions('fminunc','Algorithm','quasi-newton',
'off','SpecifyObjectiveGradient',true,'MaxIter',10^5,
'MaxFunEvals',10^5,'TolX',10^-5);
```

The experiments on simulated data were run with 50 repetitions, used a regularization parameter of 0.01, and privacy budgets of 1, 5 and 10. The amount of students (N) vary from 40 to 200 in steps of 40 and amount of questions (I) is fixed to 20. The parameters β_n and δ_i were drawn from normal distributions with mean 0, and standard deviation 1 and 2, respectively. The Rasch probabilities were then calculated with the given β_n and δ_i and used to simulate a data set by drawing from a Bernoulli distribution with the given probabilities.

The M. Vahdat et al. data set has 62 students and 16 questions. The DTU data set has 212 students and 27 questions.

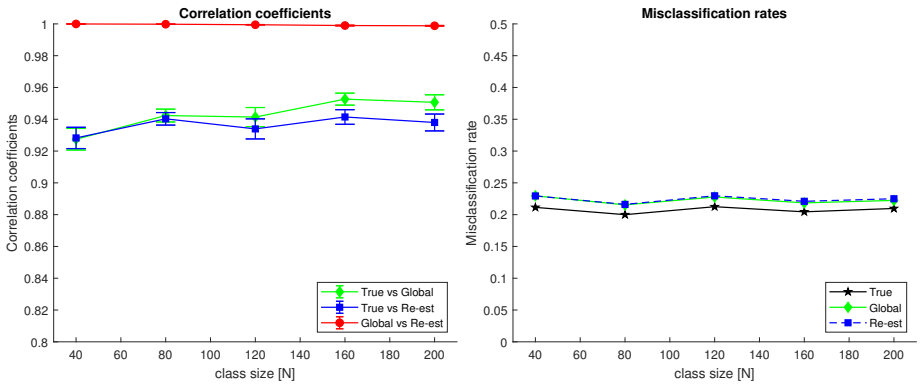
In pilot experiments we found that fine-tuning of the regularization parameter λ was not essential. So for simplicity reasons we use a common regularization paraameter in all experiments.

4.1 Results

To test the impact of introducing differential privacy to the Rasch model, we ran several experiments. In the first one we test the retraining of β in the non-private setting, to show the students can calculate their own abilities from a given, private δ. The second experiment show the performance of the two methods on simulated data, while the third and fourth show the performance on the M. Vahdat real data and DTU data, respectively.

Experiment 1: Global vs. re-estimated Rasch parameters In Figure 2 we compare the Rasch model with global parameter estimation with the results

obtained by re-estimating the student abilities. We plot correlation coefficients with 95% confidence intervals between the probabilities, also with respect to the true model parameters, as well as misclassification rates on a new data set drawn from the same distribution as the train data.



(a) (b)

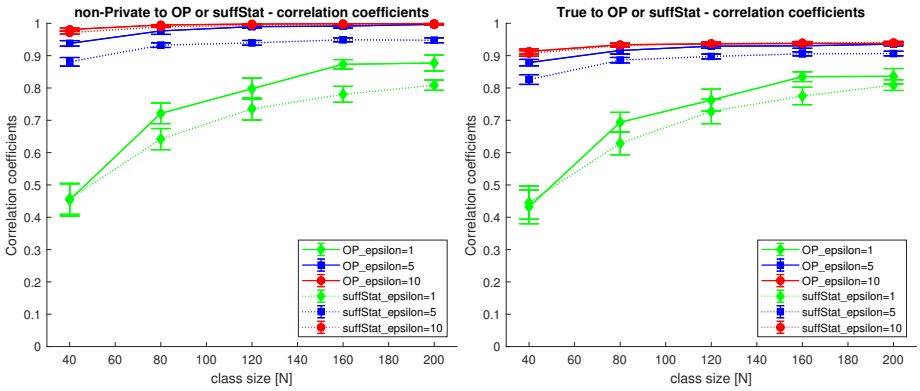
Fig. 2: The plots show: (a) Correlation coefficient for non-private global and re-estimation method compared to the true Rasch model. (b) Misclassification for non-private global and retrain estimation method compared to the true Rasch model.

From the correlation coefficients between the estimated probabilities and the ground truth probabilities used to simulate the data, we can see that the confidence intervals overlap for smaller data set sizes, while the global estimates provide better results for larger data set sizes (number of students). From now on, we will only consider the re-estimation method, since this is the one defined by our workflow.

Experiment 2: Differential Privacy on simulated data In Figure 3 we show a comparison of the objective perturbation and sufficient statistics method with three different values of epsilon: 1, 5 and 10. We compare their performance by calculating the correlation coefficients between the private estimation to the non-private resp. true estimation. Again, we show 95% confidence intervals.

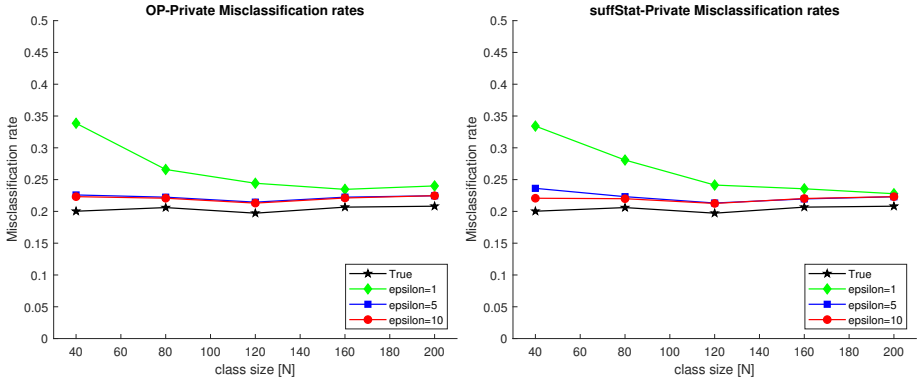
In Figure 4, we show the missclassification rates on a simulated data set of the same distribution as the training set.

We make several observations. First, the objective perturbation performs better in general. This is due to the smaller amount of noise added to the objective perturbation, as mentioned in Section 3. Next, we see that for lower epsilon values, i.e. higher privacy, the model generally performs worse, but converges



(a) (b)

Fig. 3: Correlation coefficient of objective perturbation and sufficient statistics methods with $\epsilon = (1, 5 \text{ and } 10)$ to: (a) the non-Private estimates. (b) the true model.



(a) (b)

Fig. 4: Misclassification rates of objective perturbation and sufficient statistics method with $\epsilon = (1, 5 \text{ and } 10)$: (a) objective perturbation estimates. (b) sufficient statistics estimates.

with larger class size. This is what we would expect and consistent with what is broadly observed in applications of differential privacy.

Experiment 3 and 4: Differential Privacy on real data Experiment 2 illustrated the impact of privacy so for experiments 3 and 4, the privacy budget was fixed to $\epsilon = 5$ with changing data sizes. In experiment 3 we use the data set from Vahdat et al. [9]. Experiment 4 is run on the DTU data set.

Figure 5 shows the objective perturbation and sufficient statistics performance on real data with $\epsilon = 5$. The misclassification rate here is calculated on the original data set (so corresponds to a train, not a test error).

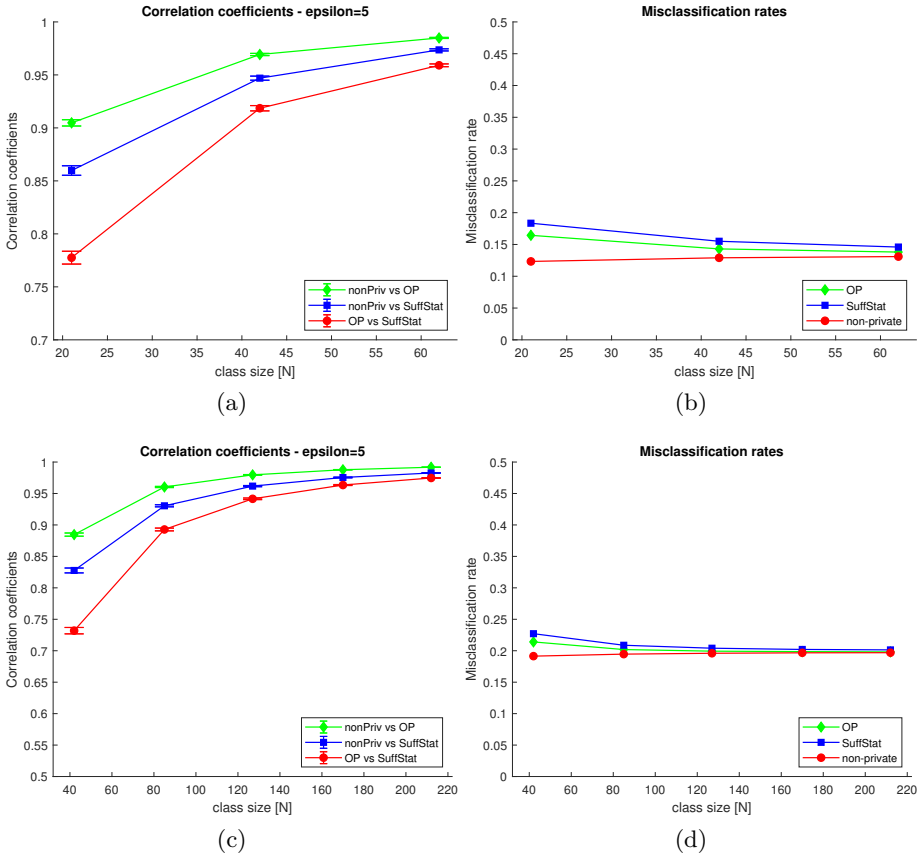


Fig. 5: Objective perturbation and sufficient statistics methods on real data: (a) Correlation coefficients on Vahdat et al.'s data [9]. (b) misclassification on Vahdat et al.'s data [9]. (c) Correlation coefficients on DTU data. (d) misclassification on DTU data.

In experiment 3, Figure 5 (a) and (b), we show that the impact of introducing privacy on real data sets is limited, even in relatively small data sets. For both the Vahdat et al. data and the DTU data we find useful correlation between the probabilities of passing the test as inferred in non-private and private models ($\epsilon = 5$).

In experiment 4, Figure 5 (c) and (d), our results are comparable to those on the simulated data. In general, we see that the objective perturbation method performs better than the sufficient statistics method.

In Figure 6, we illustrate the impact of data set size by showing comparisons of the probability estimates of the non-private model to the two private methods on the data set from Vahdat et al. [9], again with fixed $\epsilon = 5$.

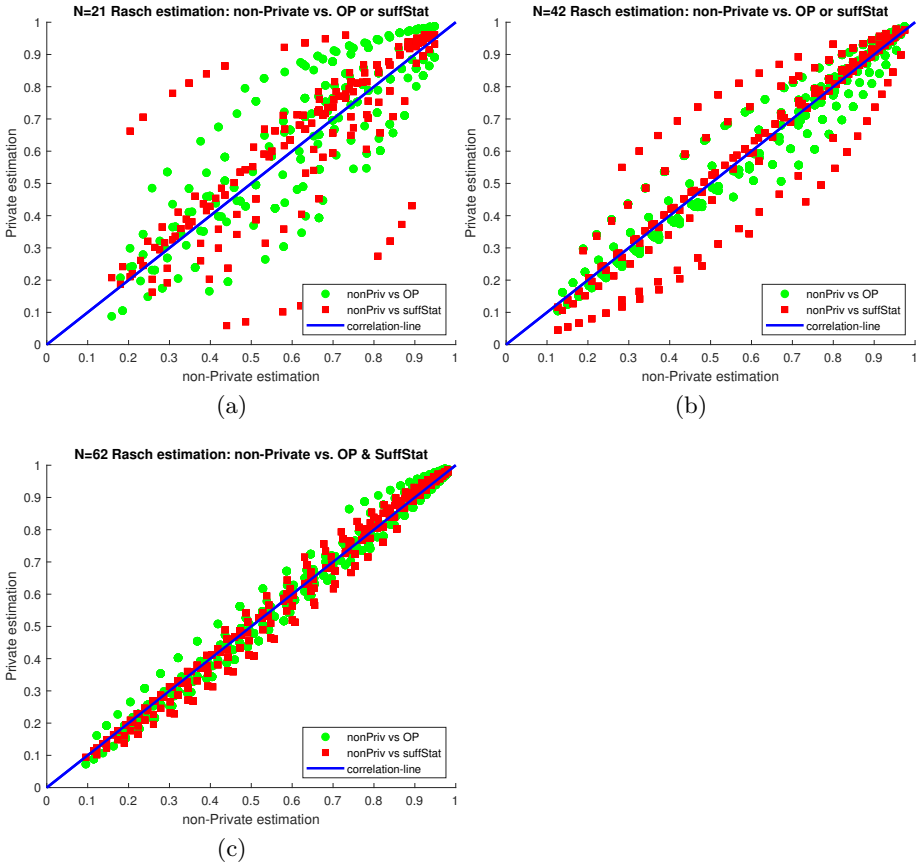


Fig. 6: Rasch estimates of objective perturbation and sufficient statistics methods on Vahdat et al.'s data [9]: (a) nStudent = 21. (b) nStudent = 42. (c) nStudent = 62

.

We see how the estimates are very noisy for small data set sizes, but correlate strongly with the non private for a data set size of 62. Again, the objective perturbatoin yields more accurate results.

5 Conclusion

We have demonstrated viability of the proposed workflow for more detailed, yet differentially private, feedback for students. We proved analytically that objective perturbation for this model satisfies differential privacy and give the minimum noise level necessary. Our experiments based on simulated data suggest that the workflow provides estimates of similar quality as the non-private for medium sized classes and industry standard privacy budgets[1]. These findings were confirmed in two real data sets. As expected, the objective perturbation mechanism performs better than the sufficient statistic method as less noise is added.

Acknowledgements

We would like to thank Martin Søren Engmann Djurhuus, who worked with us on the project in its early stages during the course "Advanced Machine Learning" at DTU. Further, we would like to thank Morten Mørup for access to the DTU data.

Appendix

Proof (of Theorem 1). Since the objective function is differentiable everywhere and a minimizing pair (β^*, δ^*) satisfies $\nabla F(\beta^*, \delta^*) = 0$, for every output (β^*, δ^*) there exists exactly one b which maps the input to the output. On the other hand, since the objective function (6) for $\lambda > 0$ is strongly convex (which can be seen by computing the Hessian matrix H and realizing that $H - \lambda I$ is positive semidefinite), for any fixed b and X, there is exactly one pair $\beta^*,\ \delta^*$ which minimizes the function. As such there is a bijection between (β^*, δ^*) and b.

Now consider two data sets X and $\tilde{X}$ that differ in exactly one student (w.l.o.g., the last one). For $\beta^*,\ \delta^*$ minimizing (6) for both X and $\tilde{X}$ denote the corresponding noise vectors b and $\tilde{b}$. By the transformation property of probability density functions, we get

$$\frac{P(\delta^*, \beta^* | X)}{P(\delta^*, \beta^* | \tilde{X})} = \frac{h(b) \left| \det \left(J_{(\beta^*, \delta^*, \tilde{X})}(\tilde{b}) \right) \right|}{h(\tilde{b}) \left| \det \left(J_{(\beta^*, \delta^*, X)}(b) \right) \right|}, \tag{8}$$

where $J_{(\beta^*, \delta^*, X)}(b)$ denotes the Jacobian matrix of b as a function of (β^*, δ^*), given input set X.

[1] See e.g., https://www.wired.com/story/apple-differential-privacy-shortcomings/

We get b_i as a function of (β^*, δ^*) by setting the gradient of the objective to zero:

$$b_i = \sum_{n=1}^{N} \frac{e^{\beta_n^*}}{e^{\beta_n^*} + e^{\delta_i^*}} - \sum_{n=1}^{N} X_{ni} - \lambda \delta_i^*. \tag{9}$$

Since the sum over X will disappear in any derivative for δ_i^* and β_n^*, the Jacobian matrices in (8) are identical and the determinants cancel.

Furthermore, since $X_{ni} = \tilde{X}_{ni}$ for all $i = 1, \ldots, I$ and all $n = 1, \ldots, (N-1)$, equation (9) also gives

$$X_N + b = \tilde{X}_N + \tilde{b}.$$

By the reverse triangle inequality we get

$$\left| \|b\| - \|\tilde{b}\| \right| \leq \|b - \tilde{b}\| = \|X_N - \tilde{X}_N\| \leq \sqrt{I}$$

and thus

$$\frac{P(\delta^*, \beta^* | X)}{P(\delta^*, \beta^* | \tilde{X})} = \frac{h(b)}{h(\tilde{b})} = e^{-\frac{\epsilon}{\sqrt{I}}(\|b\| - \|\tilde{b}\|)} \leq e^{\epsilon} \tag{10}$$

□

References

1. Chaudhuri, K., Monteleoni, C.: Privacy-preserving logistic regression. In: Proceedings of the 21st International Conference on Neural Information Processing Systems. pp. 289–296. NIPS'08, Curran Associates Inc., USA (2008), http://dl.acm.org/citation.cfm?id=2981780.2981817
2. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. Journal of Machine Learning Research **12**(Mar), 1069–1109 (2011)
3. Choppin, B.: A fully conditional estimation procedure for rasch model parameters (cse report 196): University of california. Center for the Study of Evaluation (1983)
4. Dwork, C.: Differential privacy: A survey of results. In: International Conference on Theory and Applications of Models of Computation. pp. 1–19. Springer (2008)
5. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science **9**(3–4), 211–407 (2014)
6. EU GDPR Portal: Gdpr key changes - an overview of the main changes under gdpr and how they differ from the previous directive. https://www.eugdpr.org/key-changes.html (2018), [Online; accessed 19-May-2018]
7. Foulds, J., Geumlek, J., Welling, M., Chaudhuri, K.: On the theory and practice of privacy-preserving bayesian data analysis. In: Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence. pp. 192–201. AUAI Press (2016)
8. Ji, Z., Lipton, Z.C., Elkan, C.: Differential privacy and machine learning: a survey and review. arXiv preprint arXiv:1412.7584 (2014)

9. M. Vahdat, L. Oneto, D. Anguita, M. Funk, M. Rauterberg.: A learning analytics approach to correlate the academic achievements of students with interaction data from an educational simulator. in: G. conole et al. https://archive.ics.uci.edu/ml/machine-learning-databases/00346/. https://doi.org/10.1007/978-3-319-24258-3 26, eC-TEL 2015, LNCS 9307, pp. 352-366. Springer (2015).
10. Uden, L., Liberona, D., Welzer, T.: Learning technology for education in cloud. In: Third International Workshop, LTEC 2014. Springer (2014)