Differentially Private Automatic Differentiation Variational Inference (DP-ADVI) (Extended Abstract)

Joonas Jälkö, Onur Dikmen, Antti Honkela[*]
Helsinki Institute for Information Technology (HIIT), Department of Computer Science,
University of Helsinki, Finland
{joonas.jalko, onur.dikmen, antti.honkela}@helsinki.fi

Abstract

Many machine learning applications are based on data collected from people, such as their tastes and behaviour as well as biological traits and genetic data. Regardless of how important the application might be, one has to make sure individuals' identities or the privacy of the data are not compromised in the analysis. Differential privacy constitutes a powerful framework that prevents breaching of data subject privacy from the output of a computation. Differentially private versions of many important machine learning methods have been proposed, but there is a lack of an efficient unified approach applicable to arbitrary models. In this contribution, we propose a differentially private variational inference method with a very wide applicability. It is built on top of automatic differentiation variational inference (ADVI), a recent advancement which provides a variational solution to a large class of models. We add differential privacy into ADVI by clipping and perturbing the gradients. The algorithm is made more efficient through privacy amplification from subsampling. We explore the effect of different parameter combinations in logistic regression problems where the method can reach an accuracy close to non-private level under reasonably strong privacy guarantees.

1 Introduction

Using more data usually leads to better generalisation and accuracy in machine learning. With more people getting more tightly involved in the ubiquitous data collection, privacy concerns related to the data are becoming more important. People will be much more willing to contribute their data if they can be sure that the privacy of their data can be protected.

Differential privacy (DP) [4, 5] is a strong framework with strict privacy guarantees against attacks from adversaries with arbitrary side information. The main principle is that output of an algorithm (such as a query or an estimator) should not change much if the data for one individual is modified or deleted. This can be accomplished through adding stochasticity at different levels of the estimation process, such as adding noise to data itself (input perturbation), changing the objective function to be optimised or how it is optimised (objective perturbation), releasing the estimates after adding noise (output perturbation) or by sampling from a distribution based on utility or goodness of estimates (exponential mechanism).

A lot of ground-breaking work has been done on privacy-preserving versions of standard machine learning approaches, such as objective-perturbation-based logistic regression [2], regression using

[*] AH is also with Department of Mathematics and Statistics as well as Department of Public Health, University of Helsinki, Finland. This work was funded by the Academy of Finland (Centre of Excellence COIN; and grants 278300, 259440 and 283107).

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

functional mechanism [19] to name a few. Privacy-preserving Bayesian inference (e.g. [17, 18]) has only recently started attracting more interest. One posterior sample [16] and posterior perturbation [3, 20] have been important steps towards private Bayesian learning, however they suffer from lacking asymptotic efficiency, i.e., learning does not optimally benefit from having larger number of data points. Foulds *et al.* [7] proposed an asymptotically efficient private Gibbs sampling method based on perturbing sufficient statistics of the data. This approach was recently also applied to variational inference [13]. These are applicable to models where non-private inference can be performed by accessing sufficient statistics.

Our goal in this work is to devise a generic privacy-aware variational inference method. Although ideas like perturbing sufficient statistics may find some applicability, a good private method should not depend on specifics of models such as conditional conjugacy or whether data are accessed individually or not. Recently proposed automatic differentiation variational inference (ADVI) method [10] is such a generic approach for non-private use. ADVI applies a series of transformations and approximations so that the variational distributions are Gaussian and can be optimized by stochastic gradient ascent. Here, we propose a differentially private version of it (DP-ADVI) based on gradient clipping and perturbation. We make a thorough case study on the Bayesian logistic regression model with comparisons to the non-private case under different design decisions for DP-ADVI.

2 Background

2.1 Differential privacy

Differential privacy (DP) [4, 5] is a framework that provides mathematical formulation for privacy that enables proving strong privacy guarantees.

Definition 1 ((ϵ, δ)-Differential privacy)**.** A randomised algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for all pairs of adjacent data sets x, x' differing by single sample and for every $S \subset \mathrm{im}(\mathcal{A})$

$$\Pr(\mathcal{A}(x) \in S) \leq e^{\epsilon} \Pr(\mathcal{A}(x') \in S) + \delta.$$

The privacy parameter ϵ measures the strength of the guarantee with smaller values corresponding to stronger privacy. The probabilities of obtaining a specific output with adjacent datasets should be similar with probability $1 - \delta$. When $\delta = 0$, we get ϵ-DP, also known as *pure DP*.

Composition theorems One of the very useful features of DP compared to many other privacy formulations is that it provides a very natural way to study the privacy loss incurred by repeated use of the same data set. Using an algorithm on a data set multiple times will weaken our privacy guarantee because of the potential of each application to leak more information. In fact in worst case if our algorithm is (ϵ, δ)-DP, then k-fold composition of that algorithm provides $(k\epsilon, k\delta)$-DP. Moving from the pure ϵ-DP to general (ϵ, δ)-DP allows a stronger result with a smaller ϵ at the expense of having a larger total δ on the composition. This trade-off is characterised by the Advanced composition theorem of Dwork and Roth [5, Theorem 3.20], which becomes very useful when we need to use data multiple times.

Theorem 1 (Advanced composition theorem)**.** *Given algorithm $\mathcal{A}$ that is (ϵ, δ)-DP and $\delta' > 0$, k-fold composition of algorithm $\mathcal{A}$ is $(\epsilon_{tot}, \delta_{tot})$-DP with*

$$\epsilon_{tot} = \sqrt{2k \ln(1/\delta')}\epsilon + k\epsilon(e^{\epsilon} - 1), \qquad \delta_{tot} = k\delta + \delta'. \tag{1}$$

The theorem states that with small loss in δ_{tot} and with small enough ϵ, we can provide more strict ϵ_{tot} than just summing the ϵ. This is obvious by looking at the first order expansion for small ϵ of

$$\epsilon_{tot} \approx \sqrt{2k \ln(1/\delta')}\epsilon + k\epsilon^2.$$

Potentially even stronger composition and privacy results can be obtained under a different recent relaxation of DP, concentrated DP [6].

Gaussian mechanism There are many possibilities how to make algorithm differentially private. In this paper we use *objective perturbation.* We use the *Gaussian mechanism* as our method for perturbation. Theorem 3.22 of Dwork and Roth [5] states that given query f with ℓ_2-sensitivity of $\Delta_2(f)$, releasing $f(x) + \eta$, where $\eta \sim N(0, \sigma^2)$, is (ϵ, δ)-DP when $\sigma^2 > 2\ln(1.25/\delta)\Delta_2^2(f)/\epsilon^2$, where $\Delta_2(f) = \sup_{||x-x'||=1} ||f(x) - f(x')||_2$.

Privacy amplification We use a stochastic gradient algorithm that uses subsampled data while learning, so we can make use of the amplifying effect of the subsampling on privacy. This *Privacy amplification theorem* [11] states that if we run (ϵ, δ)-DP algorithm $\mathcal{A}$ on randomly sampled subset of data with uniform sampling probability $q > \delta$, the subsampled algorithm is $(\epsilon_{amp}, \delta_{amp})$-DP with

$$\epsilon_{amp} = \log(1 + q(e^{\epsilon} - 1)), \qquad \delta_{amp} = q\delta, \tag{2}$$

assuming $\log(1 + q(e^{\epsilon} - 1)) < \epsilon$.

2.2 Variational Bayes

Variational Bayes (VB) methods [9] provide a way to approximate the posterior distribution of latent variables in a model when the true posterior is intractable. True posterior $p(\boldsymbol{\theta}|\mathbf{x})$ is approximated with a variational distribution $q_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ that has a simpler form than the posterior, obtained generally by removing some dependencies from the graphical model such as the fully-factorised form $q_{\boldsymbol{\xi}}(\boldsymbol{\theta}) = \prod_d q_{\boldsymbol{\xi}_d}(\theta_d)$. $\boldsymbol{\xi}$ are the variational parameters and their optimal values $\boldsymbol{\xi}^*$ are obtained through minimising the Kullback-Leibler (KL) divergence between $q_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{x})$. This is also equivalent to maximising the *evidence lower bound* (ELBO)

$$\mathcal{L}(q_{\boldsymbol{\xi}}) = \int q_{\boldsymbol{\xi}}(\boldsymbol{\theta}) \ln \left(\frac{p(\mathbf{x}, \boldsymbol{\theta})}{q_{\boldsymbol{\xi}}(\boldsymbol{\theta})} \right) = -\mathrm{KL}(q_{\boldsymbol{\xi}}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})) + \sum_i \langle \ln p(x_i|\boldsymbol{\theta}) \rangle_{q_{\boldsymbol{\xi}}(\boldsymbol{\theta})},$$

where $\langle\rangle_{q_{\boldsymbol{\xi}}(\boldsymbol{\theta})}$ is an expectation taken w.r.t $q_{\boldsymbol{\xi}}(\boldsymbol{\theta})$.

When the model is in the conjugate exponential family [8] and $q_{\boldsymbol{\xi}}(\boldsymbol{\theta})$ is factorised, the expectations that constitute $\mathcal{L}(q_{\boldsymbol{\xi}})$ are analytically available and each $\boldsymbol{\xi}_d$ is updated iteratively by fixed point iterations. Most popular applications of VB fall into this category, because handling of the more general case involves more approximations, such as defining another level of lower bound to the ELBO or estimating the expectations using Monte Carlo integration. The recently proposed ADVI framework [10] unifies different classes of models through a transformation of variables and optimizes the ELBO using stochastic gradient ascent (SGA). Constrained variables are transformed into unconstrained ones and their posterior is approximated by Gaussian variational distributions, which can be a product of independent Gaussians (mean-field) or larger multivariate Gaussians. Expectations in the gradients are approximated using Monte Carlo integration and the ELBO is optimized iteratively using SGA. It is also possible to consider mini batches of data at each iteration to handle big datasets, which adds another level of SGA similarly as in [15].

3 Differentially-Private ADVI

Differentially-private ADVI (DP-ADVI) is based on perturbation of contributions of individual data samples to the gradient, $g_t(x_i)$, at each iteration t and could be incorporated into the ADVI implementation in PyMC3 [14] easily. We use ADVI with subsampling to be able to use the privacy amplification. Each $g_t(x_i)$ is clipped in order to calculate gradient sensitivity. Gradient contributions from all data samples in the mini batch are summed and perturbed with Gaussian noise $\mathcal{N}(0, 4c_t^2\sigma_\delta^2\mathbf{I})$. A similar approach was used for deep learning in [1].

The sampling frequency q for subsampling within the data set, total number of iterations T and the clipping threshold c_t are important design parameters that determine the privacy cost. c_t is chosen before learning, and does not need to be constant. After clipping $||g_t(x_i)||_2 \leq c_t$, $\forall i \in U$. Clipping gradients too much will affect accuracy, but on the other hand large clipping threshold will cause large amount of noise to sum of gradients. Parameter q determines how large subsample of the training data we use to for gradient ascent. Small q values enable privacy amplification but may need a need larger T. For a very small q when the mini batches consist of just a few samples, the added noise will dominate over the gradient signal and the optimization will fail. While in our experiments q was fixed, we could also alter the q during iteration.

We have chosen to perturb the gradients at each iteration with zero mean multivariate normal noise with covariance $4c_t^2\sigma_\delta^2\mathbf{I}$. The parameter σ_δ in the noise level determines our total ϵ and depends on the total δ in the privacy budget. We can calculate by the total privacy cost by setting $\sigma = \sqrt{2\ln(1.25/(\delta_{iter}))}/\epsilon_{iter}$. Clipping will make the ℓ_2 sensitivity of the total gradient to be $2c_t$, so perturbing each gradient with the aforementioned noise will lead each iteration to be $(\epsilon_{iter}, \delta_{iter})$-DP w.r.t. the subset. If sampling probability q amplifies privacy i.e $\epsilon_{amp} < \epsilon$, then also δ will be

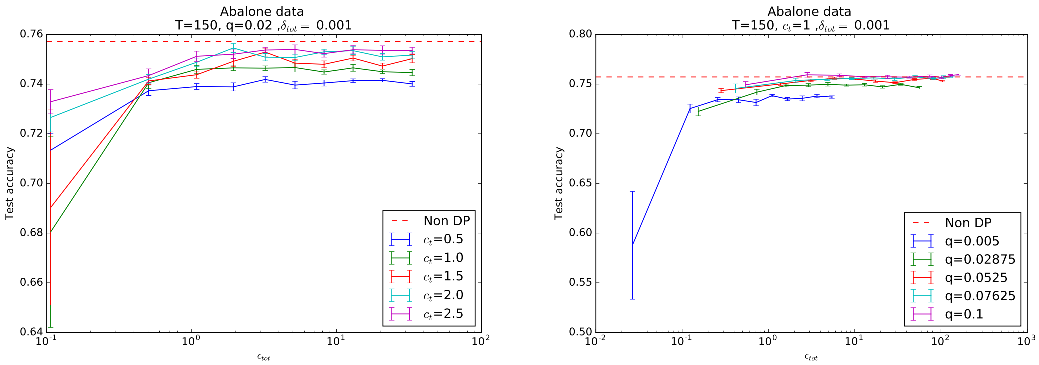

Figure 1: Accuracy vs. total ϵ in Abalone data for (a) different clipping threshold values (b) different SGA sample sizes. The curves show the mean of 10 runs of the DP-ADVI algorithm with error bars denoting the standard error of the mean.

amplified and every iteration and SGA will be $(\log(1 + q(e^{\epsilon_{iter}} - 1)), q\delta_{iter})$-DP w.r.t. the whole data set. Now if we set $\delta_{iter} = (\delta_{tot} - \delta')/Tq$, where δ' comes from the advanced composition, we can provide δ_{tot} as the δ parameter in the total privacy cost. The ϵ parameter in our total privacy cost will be

$$\epsilon_{tot} = \sqrt{2T\ln(1/\delta')}\sigma' + T\sigma'(e^{\sigma'} - 1),$$

where δ_{iter} is chosen as above and

$$\sigma' = \log\Big(1 + q\,\Big(\exp\Big(\sqrt{2\ln(1.25/\delta_{iter})}/\sigma\Big) - 1\Big)\Big).$$

4 Experiments

We tested DP-ADVI with logistic regression using the Abalone data set from the UCI Machine Learning Repository [12] for the binary classification task. Individuals were divided into two classes based on whether individual had less or more than 10 rings. The data set consisted of 4177 samples with 8 attributes. The classifier was trained using 80% of the data using SGA with sampling ratio $q = 0.02$. Before training, features of the data set were normalised by subtracting feature mean and dividing by feature standard deviation.

The plots in Figure 1 shows that the DP classifier performs quite well with relatively small ϵ_{tot} values with the accuracy approaching non-private level as ϵ_{tot} increases. As was mentioned before, there are many parameters that we can change and Figure 1(a) shows the effect of gradient clipping threshold. We can see that aggressive clipping with small c_t values is useful in overcoming the effect of large noise with a tight privacy budget corresponding to a small ϵ, but under a looser privacy budget, clipping will start hurting the learning more. Clipping the gradients too little is also bad because the increase in the level of added noise will be more significant than the increase in the retained information because of less clipping.

The effect of SGA sampling ratio q is shown in Figure 1(b). The figure shows that $q = 0.005$ corresponding to a mini batch size of approximately $q \cdot 0.8 \cdot 4177 \approx 17$ is clearly inferior to the larger values of q. Presumably the level of noise added is too strong relative to the magnitude of the gradient at this sample size. There are no clear trends in the performances of the other sampling ratios, suggesting that mini batch sizes in the range $96, \ldots, 334$ seem to perform reasonably well.

5 Conclusions

We have introduced the DP-ADVI method that can deliver differentially private inference results with accuracy close to the non-private ADVI. The method can effectively harness the power of ADVI to deal with very general models instead of just conjugate exponential models and the option of using multivariate Gaussian posterior approximations for greater accuracy.

References

[1] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. 2016. arXiv:1607.00133 [stat.ML].

[2] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *Adv. Neural Inf. Process. Syst. 21*, pages 289–296, 2008.

[3] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin I. P. Rubinstein. Robust and private Bayesian inference. In *ALT 2014*, volume 8776 of *Lecture Notes in Computer Science*, pages 291–305. Springer Science + Business Media, 2014.

[4] Cynthia Dwork. Differential privacy. In *Proc. 33rd Int. Colloq. on Automata, Languages and Prog. (ICALP 2006), Part II*, pages 1–12, 2006.

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

[6] Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. 2016. arXiv:1603.01887 [cs.DS].

[7] James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving Bayesian data analysis. In *Proc. 32nd Conf. on Uncertainty in Artificial Intelligence (UAI 2016)*, 2016.

[8] Zoubin Ghahramani and Matthew J. Beal. Propagation algorithms for variational Bayesian learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 507–513. MIT Press, 2001.

[9] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999.

[10] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. 2016. arXiv: 1603.00788.

[11] Ninghui Li, Wahbeh Qardaji, and Dong Su. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, pages 32–33, New York, NY, USA, 2012. ACM.

[12] M. Lichman. UCI machine learning repository, 2013.

[13] Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. Variational Bayes in private settings (VIPS). 2016. arXiv:1611.00340 [stat.ML].

[14] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016.

[15] Michalis Titsias and Miguel Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *Proc. 31st Int. Conf. Mach. Learn. (ICML 2014)*, pages 1971–1979, 2014.

[16] Yu-Xiang Wang, Stephen E. Fienberg, and Alexander J. Smola. Privacy for free: Posterior sampling and stochastic gradient Monte Carlo. In *Proc. 32nd Int. Conf. Mach. Learn. (ICML 2015)*, pages 2493–2502, 2015.

[17] O. Williams and F. McSherry. Probabilistic inference and differential privacy. In *Adv. Neural Inf. Process. Syst. 23*, 2010.

[18] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. PrivBayes: Private data release via Bayesian networks. In *SIGMOD'14*, pages 1423–1434, 2014.

[19] Jun Zhang, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, and Marianne Winslett. Functional mechanism: Regression analysis under differential privacy. *PVLDB*, 5(11):1364–1375, 2012.

[20] Zuhe Zhang, Benjamin Rubinstein, and Christos Dimitrakakis. On the differential privacy of Bayesian inference. In *Proc. Conf. AAAI Artif. Intell. 2016*, 2016.