DP-ADVI: Differentially Private Automatic Differentiation Variational Inference

Joonas Jälkö[1], Onur Dikmen[1] and Antti Honkela[1,2,3]

[1] Helsinki Institute for Information Technology HIIT, Department of Computer Science, University of Helsinki, Finland;
[2] Department of Mathematics and Statistics, University of Helsinki, Finland; [3] Department of Public Health, University of Helsinki, Finland

Abstract

Many machine learning applications are based on personal data (e.g. behavioural or health data). When analysing such data, one has to make sure data subjects' identities or the privacy of the data are not compromised.
Differential privacy constitutes a powerful framework to protect the privacy. 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**.
We propose a differentially private variational inference method with a very wide applicability. It is built on top of automatic differentiation variational inference (ADVI). We add differential privacy into ADVI by clipping and perturbing the gradients.

Background: Differentially private learning

(ϵ, δ)-Differential privacy

A randomised algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private (DP) if for all pairs of adjacent data sets x, x' and for every $S \subset \text{im}(\mathcal{A})$

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

When $\delta = 0$, we get ϵ-DP, also known as *pure DP*.

Gaussian mechanism

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

$$\sigma^2 > \frac{2 \ln(1.25/\delta) \Delta_2^2(f)}{\epsilon^2}.$$

ℓ_2-sensitivity of a query is defined as:

$$\Delta_2(f) = \sup_{\substack{x,x' \\ ||x - x'|| = 1}} ||f(x) - f(x')||_2.$$

Composition

- If an algorithm is (ϵ, δ)-DP, then k-fold composition of that algorithm provides $(k\epsilon, k\delta)$-DP
- **Advanced composition theorem [2]:** 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

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

- **Privacy amplification theorem [3]:** If we run (ϵ, δ)-DP algorithm $\mathcal{A}$ on randomly sampled subset of data with uniform sampling probability $q > \delta$, privacy amplification theorem states that the subsampled algorithm is $(\epsilon_{amp}, \delta_{amp})$-DP with

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

  assuming $\log(1 + q(e^{\epsilon} - 1)) < \epsilon$.
- Moments accountant [4] can yield smaller ϵ_{amp} than advanced composition theorem by taking noise distributions into consideration

Differentially Private Automatic Differentiation Variational Inference (DP-ADVI)

Variational inference

- True posterior $p(\boldsymbol{\theta}|\mathbf{x})$ is approximated with a variational distribution $q_{\xi}(\boldsymbol{\theta})$ that has a simpler form (e.g., $q_{\xi}(\boldsymbol{\theta}) = \prod_d q_{\xi_d}(\theta_d)$).
- ξ^* are obtained through minimising the Kullback–Leibler (KL) divergence between $q_{\xi}(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathbf{x})$
- Equivalently, maximising the *evidence lower bound* (ELBO)

$$\mathcal{L}(q_{\xi}) = \int q_{\xi}(\boldsymbol{\theta}) \ln\left(\frac{p(\mathbf{x}, \boldsymbol{\theta})}{q_{\xi}(\boldsymbol{\theta})}\right) = -\text{KL}(q_{\xi}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})) + \sum_{i=1}^{N} \langle \ln p(x_i|\boldsymbol{\theta}) \rangle_{q_{\xi}(\boldsymbol{\theta})}$$

$$= \sum_{i=1}^{N} \left(-\frac{1}{N} \text{KL}(q_{\xi}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})) + \langle \ln p(x_i|\boldsymbol{\theta}) \rangle_{q_{\xi}(\boldsymbol{\theta})} \right) \equiv \sum_{i=1}^{N} \mathcal{L}_i(q_{\xi})$$

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

ADVI [5]

- Constrained variables are transformed into unconstrained ones and their posterior is approximated by Gaussian variational distributions
- Does not need conjugacy, optimizes the ELBO using stochastic gradient ascent (SGA)
- Provides a unified methodology for a broad range of models

DP-ADVI

- Each $g(x_i) = \nabla_{\xi} \mathcal{L}_i(q_{\xi})$ is clipped s.t. $||g(x_i)||_2 \leq c_t$ in order to calculate *gradient sensitivity*
- Subsampling with frequency q in order to use the *privacy amplification theorem*
- Gradient contributions from all data samples in the mini batch are summed and perturbed with Gaussian noise $\mathcal{N}(0, 4c_t^2\sigma_{\delta}^2\mathbf{I})$
- Incorporated into the ADVI implementation in PyMC3.

Calculating the privacy budget [1]

- σ_{δ} determines the total ϵ_{tot} and depends on total δ_{tot}
- Setting $\delta_{subs} = (\delta_{tot} - \delta')/Tq$, define σ_{δ} via iteration-specific ϵ_{subs}

$$\sigma_{\delta} = \sqrt{2 \ln(1.25/(\delta_{subs}))}/\epsilon_{subs}.$$

- Clipping makes ℓ_2 sensitivity of total gradient $2c_t$
- $(\epsilon_{subs}, \delta_{subs})$-DP w.r.t. the subset is $(\log(1 + q(e^{\epsilon_{subs}} - 1)), q\delta_{subs})$-DP w.r.t. whole data set.
- Total privacy cost ϵ_{tot} over T iterations is

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

  with

$$\epsilon_{iter} = \log(1 + q\left(\exp(\sqrt{2 \ln(1.25/\delta_{subs})}/\sigma_{\delta}) - 1\right))$$

  where δ' comes from advanced composition

Experiments: Logistic regression

- Bayesian logistic regression model

$$P(y|\mathbf{x}, \mathbf{w}) = \sigma(y\mathbf{w}^T\mathbf{x})$$
$$p(\mathbf{w}) = N(\mathbf{w}; \mathbf{w}_0, \mathbf{S}_0),$$

  where $\sigma(x) = 1/(1 + \exp(-x))$.
- We take no prior on the covariance matrix $\mathbf{S}_0$ which is fixed to $\mathbf{S}_0 = \mathbf{I}_d$
- Approximate $p(\mathbf{w}|\mathbf{X}, Y)$ with $q(\mathbf{w})$ which is multivariate normal with mean $\mathbf{m}_N$ and covariance $\mathbf{S}_N$
- We use $\mathbf{S}_N = \sigma\mathbf{I}_d$ (mean-field), it is possible to use full covariance, but DP introduces a new accuracy tradeoff
- Abalone data set from the UCI Machine Learning Repository (4177 samples, 8 features, 2 classes)
- The classifier was trained using 80% of the data using SGA with sampling ratio $q = 0.02$.
- Before training, data are z-normalised

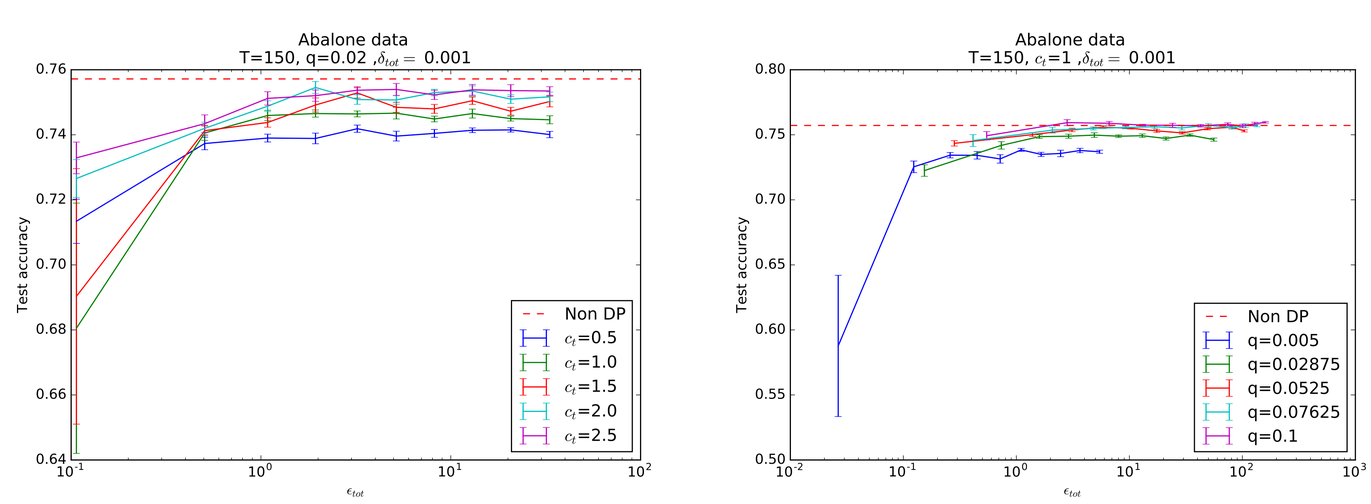

Figure: Accuracy vs. total ϵ in Abalone data for (a) different clipping threshold values (b) different SGA sample sizes. The curves show the mean of 10 runs of the DP-ADVI algorithm with error bars denoting the standard error of the mean.

References

[1] Joonas Jälkö, Onur Dikmen, and Antti Honkela. Differentially Private Variational Inference for Non-conjugate Models. arXiv:1610.08749. 2016.
[2] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(34):211407, 2014.
[3] 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*, 2012.
[4] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. arXiv:1607.00133. 2016.
[5] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M. Blei. Automatic differentiation variational inference. arXiv:1603.00788. 2016.

Acknowledgements

This work was funded by the Academy of Finland (Centre of Excellence COIN; and grants 278300, 259440 and 283107).