Variational Bayes in Private Settings (VIPS) (Extended Abstract)*

James R. Foulds[1†] , **Mijung Park**[2,3†‡] , **Kamalika Chaudhuri**[4] and **Max Welling**[5]
[1]Department of Information Systems, University of Maryland, Baltimore County
[2]Max Planck Institute for Intelligent Systems
[3]Department of Computer Science, University of Tübingen
[4]Department of Computer Science, University of California, San Diego
[5]Amsterdam Machine Learning LAB (AMLAB), Informatics Institute, University of Amsterdam
jfoulds@umbc.edu, mijung.park@tuebingen.mpg.de, kamalika@cs.ucsd.edu, m.welling@uva.nl

Abstract

Many applications of Bayesian data analysis involve sensitive information such as personal documents or medical records, motivating methods which ensure that privacy is protected. We introduce a general privacy-preserving framework for Variational Bayes (VB), a widely used optimization-based Bayesian inference method. Our framework respects differential privacy, the gold-standard privacy criterion. The iterative nature of variational Bayes presents a challenge since iterations increase the amount of noise needed to ensure privacy. We overcome this by combining: (1) an improved composition method, called the *moments accountant*, and (2) the privacy amplification effect of subsampling mini-batches from large-scale data in stochastic learning. We empirically demonstrate the effectiveness of our method on LDA topic models, evaluated on Wikipedia. In the full paper we extend our method to a broad class of models, including Bayesian logistic regression and sigmoid belief networks [Park *et al.*, 2020].

1 Introduction

Bayesian inference, which reasons over the uncertainty in model parameters and latent variables given data and prior knowledge, has found widespread use in data science application domains in which privacy is essential, including text analysis [Blei *et al.*, 2003], medical informatics [Husmeier *et al.*, 2006], and MOOCS [Piech *et al.*, 2013]. In these applications, the goals of the analysis must be carefully balanced against the privacy concerns of the individuals whose data are being studied [Daries *et al.*, 2014]. The *Differential Privacy* (DP) formalism provides a means for analyzing and controlling this trade-off, by quantifying the privacy "cost" of data-driven algorithms [Dwork *et al.*, 2006b]. In this work, we address the challenge of performing Bayesian inference in private settings, by developing an extension of the widely used *Variational Bayes* (VB) algorithm that preserves differential privacy. Variational Bayes provides an optimisation-based alternative to Markov Chain Monte Carlo (MCMC) simulation methods for Bayesian inference, and as such, frequently has faster convergence properties than MCMC.

Iterative algorithms, such as variational Bayes, pose a further challenge when developing a differentially private algorithm: each iteration corresponds to a query to the database which must be privatised, and the number of iterations required to guarantee accurate posterior estimates causes high cumulative privacy loss. To compensate for the loss, one needs to add a significantly higher level of noise to the quantity of interest. We overcome these challenges in the context of variational Bayes by using the following key ideas:

- **perturbation of the expected sufficient statistics** to make effective use of the *per iteration privacy budget*,
- **a refined composition analysis** using the *Moments Accountant* to increase the per-iteration privacy budget,
- **leveraging the privacy amplification effect from subsampling of large-scale data** to scale up the algorithm while improving privacy guarantees, and
- **data augmentation for non-CE family models** to generalize our approach to a broad class of models, which we describe in the full journal paper.

Taken together, these ideas result in an algorithm for privacy-preserving variational Bayesian inference that is both practical and broadly applicable. Our code is available at https://github.com/mijungi/vips_code.

2 Background

Differential Privacy (DP) is a formal definition of the privacy properties of data analysis algorithms [Dwork *et al.*, 2006b; Dwork and Roth, 2014]. A randomized algorithm $\mathcal{M}(\mathcal{D})$ is said to be (ϵ, δ)-differentially private if

$$P(\mathcal{M}(\mathcal{D}) \in \mathcal{S}) \leq \exp(\epsilon) P(\mathcal{M}(\mathcal{D}') \in \mathcal{S}) + \delta \qquad (1)$$

for all measurable subsets $\mathcal{S}$ of the range of $\mathcal{M}$ and for all datasets $\mathcal{D}$, $\mathcal{D}'$ differing by a single entry. In this paper, we assume that the entry difference incurs by replacing an entry with a different value (the "*replace-one*" version of DP).

Intuitively, the definition states that the probability of any event does not change very much when a single individual's

*This paper is an extended abstract of an article in JAIR [Park *et al.*, 2020].

†The first two authors contributed equally.

‡Contact Author.

data is modified, thereby limiting the amount of information that the algorithm reveals about any one individual. There are several standard approaches for designing differentially-private algorithms – see [Dwork and Roth, 2014] and [Sarwate and Chaudhuri, 2013] for surveys. The classical approach is the *global sensitivity method* by [Dwork *et al.*, 2006b]. The global sensitivity of a function F of a dataset $\mathcal{D}$ is defined as the maximum amount (over all datasets $\mathcal{D}$) by which F changes when the private value of a single individual in $\mathcal{D}$ changes: $\Delta F = \max_{\mathcal{D},\mathcal{D}' \colon \mathrm{d}(\mathcal{D},\mathcal{D}') \leq 1} |F(\mathcal{D}) - F(\mathcal{D}')|$, where $\mathcal{D}$ is allowed to vary over the entire data domain, and $|F(\cdot)|$ can correspond to either the L_1 norm $\|F(\cdot)\|_1$ or the L_2 norm $\|F(\cdot)\|_2$. In this paper, we consider a specific form of the global sensitivity method, called the *Gaussian mechanism* [Dwork *et al.*, 2006a], where Gaussian noise calibrated to the global sensitivity in the L_2 norm (the *L2 sensitivity*) is added. For a function F with global sensitivity ΔF, we output:

$$F(\mathcal{D}) + Z, Z \sim N(0, \sigma^2)\,, \sigma^2 \geq 2\log(1.25/\delta)(\Delta F)^2/\epsilon^2\,,$$

where ΔF is computed using the L_2 norm.

2.1 Composition and Subsampling

An important property of differential privacy which makes it conducive to real applications is *composition*, which means that the privacy guarantees decay gracefully as the same private dataset is used in multiple releases. This property allows us to easily design private versions of iterative algorithms by making each iteration private, and then accounting for the privacy loss incurred by a fixed number of iterations.

The key privacy tool in this paper is the composition of multiple iterations of the Subsampled Gaussian Mechanism, originally analyzed by [Abadi *et al.*, 2016], and refined by [Wang *et al.*, 2019] (the Analytical Moments Accountant). The Subsampled Gaussian Mechanism, under the analysis of [Wang *et al.*, 2019], works as follows. Given a dataset $\mathcal{D}$ with N datapoints and a sampling proportion υ, in each iteration, we draw a fresh independent random sample of υN points from the entire dataset $\mathcal{D}$. These υN points are drawn *without replacement* and are hence distinct. We then compute a function F (to be specified later) on these υN points, and privatize F using the Gaussian mechanism. The Moments Accountant keeps track of privacy loss under composition based on a quantity called the log-moment function. Theorem 9 of [Wang *et al.*, 2019] provides an analytical expression for computing the log-moment of the Subsampled Gaussian Mechanism as a function of the sampling proportion, and the log-moments of the Gaussian Mechanism. Since all log-moments of the Gaussian Mechanism can be calculated directly by simple algebra [Abadi *et al.*, 2016; Wang *et al.*, 2019], this gives an analytical way to calculate the log-moment of a single iteration of the Subsampled Gaussian Mechanism. Successive iterations are then composed by adding up the log-moments.

2.2 Variational Bayes

Consider a generative model that produces a dataset $\mathcal{D} = \{\mathcal{D}_n\}_{n=1}^N$ consisting of N (conditionally) independent identically distributed (*i.i.d.*) items ($\mathcal{D}_n$ denotes the nth input/output pair $\{\mathbf{x}_n, y_n\}$ for supervised learning and the nth vector output $\mathbf{y}_n$ for unsupervised learning), generated using a set of latent variables $\boldsymbol{l} = \{\boldsymbol{l}_n\}_{n=1}^N$. The generative model provides $p(\mathcal{D}_n|\boldsymbol{l}_n, \boldsymbol{m})$, where $\boldsymbol{m}$ are the model parameters. Variational Bayes recasts the task of approximating the posterior $p(\boldsymbol{l}, \boldsymbol{m}|\mathcal{D})$ as an optimisation problem: making an approximating distribution $q(\boldsymbol{l}, \boldsymbol{m})$, which is called the *variational distribution*, as similar as possible to the posterior, by minimising some distance (or divergence) between them, typically the KL-divergence. The standard *mean field* assumption is that q is a fully factorized distribution, $q(\boldsymbol{l}, \boldsymbol{m}) = q(\boldsymbol{l})q(\boldsymbol{m}) = q(\boldsymbol{m})\prod_{n=1}^N q(\boldsymbol{l}_n)$. Mean-field VB simplifies to a two-step procedure when the model falls in the Conjugate-Exponential (CE) class of models [Beal, 2003]:

(1) The complete-data likelihood is in the exponential family:

$$p(\mathcal{D}_n, \boldsymbol{l}_n|\boldsymbol{m}) = g(\boldsymbol{m})f(\mathcal{D}_n, \boldsymbol{l}_n)\exp(\mathbf{n}(\boldsymbol{m})^\top \mathbf{s}(\mathcal{D}_n, \boldsymbol{l}_n)),$$

(2) The prior on model parameters $\boldsymbol{m}$ is conjugate to the complete-data likelihood:

$$p(\boldsymbol{m}|\tau, \boldsymbol{\nu}) = h(\tau, \boldsymbol{\nu})g(\boldsymbol{m})^\tau \exp(\boldsymbol{\nu}^\top \mathbf{n}(\boldsymbol{m})),$$

where the natural parameters (to be inferred) and sufficient statistics (a function of the data and the latent variables to be inferred) of the complete-data likelihood are denoted by $\mathbf{n}(\boldsymbol{m})$ and $\mathbf{s}(\mathcal{D}_n, \boldsymbol{l}_n)$, respectively, and g, f, h are some known functions. The hyperparameters (that need to be tuned) are denoted by τ (a scalar) and $\boldsymbol{\nu}$ (a vector).

For more efficient learning, we adopt stochastic variational inference, which uses stochastic optimisation to fit the variational distribution over the parameters. The stochastic variational Bayes algorithm is summarised in Algorithm 1.

Algorithm 1 (Stochastic) Variational Bayes for CE family distributions

```
Input: Data $\mathcal{D}$. Define $\rho_t = (\tau_0 + t)^{-\kappa}$ and mini-batch size $S$.
Output: Expected natural parameters $\bar{\mathbf{n}}$ and expected sufficient statistics $\bar{\mathbf{s}}$.
  for $t = 1, \ldots, J$ do
    Draw a minibatch of $S$ datapoints, without replacement.

    (1) E-step: Given the expected natural parameters $\bar{\mathbf{n}}$, compute $q(\boldsymbol{l}_n)$ for $n = 1, \ldots, S$. Output the expected sufficient statistics $\bar{\mathbf{s}} = \frac{1}{S}\sum_{n=1}^S \langle \mathbf{s}(\mathcal{D}_n, \boldsymbol{l}_n)\rangle_{q(\boldsymbol{l}_n)}$.
    (2) M-step: Given $\bar{\mathbf{s}}$, compute $q(\boldsymbol{m})$ by $\tilde{\boldsymbol{\nu}}^{(t)} = \boldsymbol{\nu} + N\bar{\mathbf{s}}$. Set $\tilde{\boldsymbol{\nu}}^{(t)} \mapsfrom (1-\rho_t)\tilde{\boldsymbol{\nu}}^{(t-1)} + \rho_t\tilde{\boldsymbol{\nu}}^{(t)}$. Output the expected natural parameters $\bar{\mathbf{n}} = \langle \mathbf{n}(\boldsymbol{m})\rangle_{q(\boldsymbol{m})}$.
  end for
```

3 Variational Bayes In Private Settings (VIPS) for the CE Family

A naive way to privatise the VB algorithm is by perturbing both $q(\boldsymbol{l})$ and $q(\boldsymbol{m})$. Unfortunately, this is impractical, due to the excessive amounts of additive noise (we typically have as many latent variables as the number of datapoints). We instead propose to perturb the expected sufficient statistics

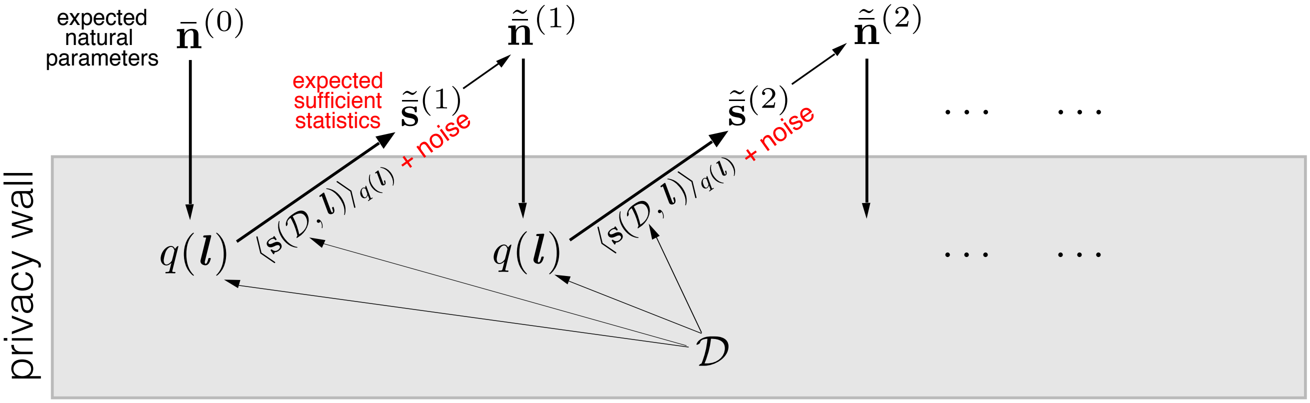
expected natural parameters
$\bar{\mathbf{n}}^{(0)}$
expected sufficient statistics
$\tilde{\bar{\mathbf{s}}}^{(1)}$
$\tilde{\bar{\mathbf{n}}}^{(1)}$
$\tilde{\bar{\mathbf{s}}}^{(2)}$
$\tilde{\bar{\mathbf{n}}}^{(2)}$
+ noise
+ noise
$\langle \mathbf{s}(\mathcal{D}, \boldsymbol{l}) \rangle_{q(\boldsymbol{l})}$
$\langle \mathbf{s}(\mathcal{D}, \boldsymbol{l}) \rangle_{q(\boldsymbol{l})}$
$q(\boldsymbol{l})$
$q(\boldsymbol{l})$
$\mathcal{D}$
. . . . . .
. . . . . .
privacy wall

Figure 1: Schematic of VIPS. Given the initial expected natural parameters $\bar{\mathbf{n}}^{(0)}$, we compute the variational posterior over the latent variables $q(\boldsymbol{l})$. Since $q(\boldsymbol{l})$ is a function of not only the expected natural parameters but also the data $\mathcal{D}$, we compute $q(\boldsymbol{l})$ behind the privacy wall. Using $q(\boldsymbol{l})$, we then compute the expected sufficient statistics. Note that we neither perturb nor output $q(\boldsymbol{l})$ itself. Instead, when we noise up the expected sufficient statistics before outputting, we add noise to each coordinate of the expected sufficient statistics in order to compensate the maximum difference in $\langle \mathbf{s}_l(\mathcal{D}_n, \boldsymbol{l}_n) \rangle_{q(\boldsymbol{l}_n)}$ caused by both $\mathcal{D}_n$ and $q(\boldsymbol{l}_n)$. In the M-step, we compute the variational posterior over the parameters $q(\boldsymbol{m})$ using the perturbed expected sufficient statistics $\tilde{\bar{\mathbf{s}}}^{(1)}$. Using $q(\boldsymbol{m})$, we compute the expected natural parameters $\tilde{\bar{\mathbf{n}}}^{(1)}$, which is already perturbed since it is a function of $\tilde{\bar{\mathbf{s}}}^{(1)}$. We continue performing these two steps until convergence.

```
Algorithm 2 Private VIPS for CE family distributions
Input: Data $\mathcal{D}$. Define $\rho_t = (\tau_0 + t)^{-\kappa}$, noise variance $\sigma^2$,
    mini-batch size $S$, and maximum iterations $J$.
Output: Perturb expected natural parameters $\tilde{\bar{\mathbf{n}}}$ and expected
    sufficient statistics $\tilde{\bar{\mathbf{s}}}$.
    Compute the L2-sensitivity $\Delta$ of the expected sufficient statistics.
    for $t = 1, \ldots, J$ do
        Draw a minibatch of $S$ datapoints, without replacement.

        (1) E-step: Given the expected natural parameters $\bar{\mathbf{n}}$,
        compute $q(\boldsymbol{l}_n)$ for $n = 1, \ldots, S$. Perturb each coordinate
        of $\bar{\mathbf{s}} = \frac{1}{S}\sum_{n=1}^{S} \langle \mathbf{s}(\mathcal{D}_n, \boldsymbol{l}_n) \rangle_{q(\boldsymbol{l}_n)}$ by adding
        $\mathcal{N}(0, \sigma^2\Delta^2 I)$ noise, and output $\tilde{\bar{\mathbf{s}}}$.
        (2) M-step: Given $\tilde{\bar{\mathbf{s}}}$, compute $q(\boldsymbol{m})$ by $\tilde{\boldsymbol{\nu}}^{(t)} = \boldsymbol{\nu} + N\tilde{\bar{\mathbf{s}}}$.
        Set $\tilde{\boldsymbol{\nu}}^{(t)} \leftarrow (1-\rho_t)\tilde{\boldsymbol{\nu}}^{(t-1)} + \rho_t \tilde{\boldsymbol{\nu}}^{(t)}$. Output the expected
        natural parameters $\tilde{\bar{\mathbf{n}}} = \langle \mathbf{n}(\boldsymbol{m}) \rangle_{q(\boldsymbol{m})}$.
    end for
    Compute the privacy loss $(\epsilon_{tot}, \delta_{tot})$ using the analytical
    moments account method [Wang et al., 2019].
```

only. We provide a schematic of our overall procedure for Variational Bayes in private settings (VIPS) in Fig. 1.

Algorithm 2 provides pseudocode of our algorithm for differentially private stochastic variational Bayes for CE family models. In the full paper we develop a method to address the non-CE family case, which we apply to Bayesian logistic regression and sigmoid belief networks.

4 VIPS for Latent Dirichlet Allocation

In the LDA topic model, we observe a corpus of D documents $\mathcal{D}_d$, where each observed word is represented by an indicator vector $\mathbf{w}_{dn}$ (nth word in the dth document) of length V, and where V is the number of terms in a fixed vocabulary set. Given the corpus, the model infers K latent topics $\boldsymbol{\beta}_k$, which are discrete distributions over the vocabulary, and discrete distributions over topics $\boldsymbol{\theta}_d$ for each document $\mathcal{D}_d$. Each word $\mathbf{w}_{dn}$ is given a topic assignment latent variable $\mathbf{z}_{dn}$, represented by an indicator vector of length K. Let $\mathbf{1}_L$ be a vector of ones of length L, for any integer L. The LDA model posits that the generative process for the corpus is:

- Draw topics $\boldsymbol{\beta}_k \sim$ Dirichlet $(\eta \mathbf{1}_V)$, for $k = \{1, \ldots, K\}$, where η is a scalar hyperparameter.
- For each document $\mathcal{D}_d, d \in \{1, \ldots, D\}$
  - Draw topic proportions $\boldsymbol{\theta}_d \sim$ Dirichlet $(\alpha \mathbf{1}_K)$, where α is a scalar hyperparameter.
  - For each word $n \in \{1, \ldots, N\}$
    - ∗ Draw topic assignments $\mathbf{z}_{dn} \sim$ Discrete$(\boldsymbol{\theta}_d)$
    - ∗ Draw word $\mathbf{w}_{dn} \sim$ Discrete$(\boldsymbol{\beta}_{\mathbf{z}_{dn}})$.

The LDA model falls in the CE family, viewing $\mathbf{z}_{d,1:N}$ and $\boldsymbol{\theta}_d$ as two types of latent variables: $\boldsymbol{l}_d = \{\mathbf{z}_{d,1:N}, \boldsymbol{\theta}_d\}$, and $\boldsymbol{\beta}$ as model parameters $\boldsymbol{m} = \boldsymbol{\beta}$. We follow the general framework of VIPS for differentially private LDA, with the addition of several LDA-specific heuristics which are important for good performance. First, while each document originally has a different document length N_d, in order to bound the sensitivity, and to ensure that the signal-to-noise ratio remains reasonable for very short documents, we preprocess all documents to have the same fixed length N. We accomplish this by sampling N words with replacement from each document's bag of words. In our experiments, we use $N = 500$. Note that since privacy is preserved at the level of documents rather than at the level of words, this step does not directly impact the degree of privacy achieved, but it can reduce the relative amount of noise compared to the amount of data.

To perturb the expected sufficient statistics $\bar{\mathbf{s}}$, which is a matrix of size $K \times V$, we add Gaussian noise to each component of this matrix:

$$\tilde{\bar{\mathbf{s}}}_k^v = \bar{\mathbf{s}}_k^v + Y_k^v, \text{ where } Y_k^v \sim \mathcal{N}(0, \sigma^2(\Delta\bar{\mathbf{s}})^2), \qquad (2)$$

$\bar{\mathbf{s}}_k^v = \frac{1}{S}\sum_d \sum_n \phi_{dn}^k \mathbf{w}_{dn}^v$, and $\Delta\bar{\mathbf{s}}$ is the sensitivity. We then map the perturbed components to 0 if they become negative. For LDA, with a mini-batch of S documents we show with simple algebra that the the worst-case sensitivity is given by

$$\Delta\bar{\mathbf{s}} \leq \frac{N}{S} \,. \qquad (3)$$

In our practical implementation, we improve the sensitivity by exploiting the fact that for most typical documents, the document's contribution's norm $|\bar{\mathbf{s}}^d|$ will be smaller than the worst case norm $\frac{N}{S}$. Specifically, inspired by [Abadi *et al.* 2016], we apply a norm clipping strategy, in which the per-document contributions $\bar{\mathbf{s}}^d$ are clipped (or projected) such that $|\bar{\mathbf{s}}^d| \leq a\frac{N}{S}$, for a user-specified $a \in (0, 1]$. Note that when $a = 1$, no clipping is applied. For each document in the mini-batch, if this criterion is not satisfied, we project the expected sufficient statistics $\bar{\mathbf{s}}^d$ down to the required norm via

$$\bar{\mathbf{s}}^d := \frac{aN}{S}\frac{\bar{\mathbf{s}}^d}{|\bar{\mathbf{s}}^d|} \,. \qquad (4)$$

For example, consider a scenario where $N = V = K = 2$, $S = 1$, $a = 0.1$, and $\bar{\mathbf{s}}^d = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$. The worst-case norm is $\frac{N}{S} = 2$, the clipping threshold is calculated as $a\frac{N}{S} = 0.2$, and the norm of $\bar{\mathbf{s}}^d$ is $|\bar{\mathbf{s}}^d| = \sqrt{2} \approx 1.41 > 0.2$. The document's expected sufficient statistics hence are clipped to

$$\bar{\mathbf{s}}^d := \frac{aN}{S}\frac{\bar{\mathbf{s}}^d}{|\bar{\mathbf{s}}^d|} = \frac{0.2}{\sqrt(2)}\begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} \,. \qquad (5)$$

These clipped sufficient statistics have a norm of $|\bar{\mathbf{s}}^d| = 0.2$, as required. Intuitively, for a fixed N and S, the L1 norm of $\bar{\mathbf{s}}^d$ is always N/S, but its L2 norm can become arbitrarily small as the dimensionality of $\bar{\mathbf{s}}^d$ increases. After this procedure, the sensitivity of the clipped expected sufficient statistics matrix becomes $a\Delta\bar{\mathbf{s}}$ (i.e., $a\frac{N}{S}$), and we add noise on this scale to the clipped expected sufficient statistics. We set $a = 0.1$ in our experiments, which empirically resulted in clipping being applied to around $3/4$ of the documents, while improving the sensitivity by an order of magnitude.

4.1 Experiments using Wikipedia Data

We downloaded a random $D = 400,000$ documents from Wikipedia to test our VIPS algorithm. We used 50 topics and a vocabulary set of approximately 8000 terms. The algorithm was run for one epoch in each experiment. We compared our moments accountant approach with a baseline method using the *strong composition* (Theorem 3.20 of [Dwork and Roth, 2014]) and a baseline where the clipping step was not performed. Figure 2 shows the trade-off between ϵ and per-word perplexity on the Wikipedia dataset for the different methods under a variety of conditions. Our proposed method outperformed the baselines. In Table 1 we show the top 10 words in terms of assigned probabilities for an example topic.

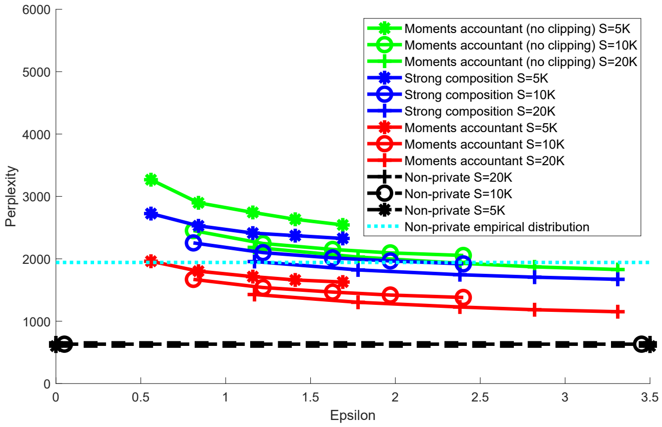

Figure 2: Epsilon versus perplexity, varying σ and S, Wikipedia data, one epoch. Perplexity is approximated using the upper bound of [Hoffman *et al.*, 2010] which is exact for the *non-private empirical distribution* baseline, hence this has a slightly unfair advantage.

Non-private	Moments Accountant	Strong Composition	Moments Acc. (no clipping)
station	station	station	station
line	line	line	line
railway	railway	railway	french
opened	opened	opened	railway
services	services	services	opened
located	closed	stations	services
closed	code	closed	republic
owned	country	section	closed
stations	located	platform	stations
platform	stations	republic	country

Table 1: Example topic from private LDA ($\epsilon = 2.38$)

5 Conclusion

We have developed a practical privacy-preserving VB algorithm and illustrated its performance for topic modeling. In the full paper, we generalize our approach to non-CE models, including Bayesian logistic regression and sigmoid belief networks. Our broader vision is that *practical* private ML algorithms will have a transformative impact on the practice of data science in many real-world applications.

Acknowledgments

M. Park and J. Foulds contributed equally. A significant amount of work of M. Park and M. Welling was supported by the QUVA lab at the University of Amsterdam. The later work of M. Park was supported by the Max Planck Society, as well as the Gibs Schüle Foundation and the Institutional Strategy of the University of Tübingen (ZUK63). K. Chaudhuri was partially supported by ONR under N00014-16-1-261, UC Lab Fees under LFR 18-548554 and a Google Faculty Fellowship.

References

[Abadi *et al.*, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[Beal, 2003] Matthew J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Unit, University College London, 2003.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[Daries *et al.*, 2014] Jon P. Daries, Justin Reich, Jim Waldo, Elise M. Young, Jonathan Whittinghill, Andrew Dean Ho, Daniel Thomas Seaton, and Isaac Chuang. Privacy, anonymity, and big data in the social sciences. *Communications of the ACM*, 57(9):56–63, 2014.

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

[Dwork *et al.*, 2006a] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503. Springer, 2006.

[Dwork *et al.*, 2006b] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[Hoffman *et al.*, 2010] Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc., 2010.

[Husmeier *et al.*, 2006] Dirk Husmeier, Richard Dybowski, and Stephen Roberts. *Probabilistic modeling in bioinformatics and medical informatics*. Springer Science & Business Media, 2006.

[Park *et al.*, 2020] Mijung Park, James R. Foulds, Kamalika Chaudhuri, and Max Welling. Variational Bayes in private settings (VIPS). *Journal of Artificial Intelligence Research (JAIR)*, 68:109–157, 2020.

[Piech *et al.*, 2013] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining*, pages 153–160, 2013.

[Sarwate and Chaudhuri, 2013] Anand D. Sarwate and Kamalika Chaudhuri. Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data. *IEEE Signal Process. Mag.*, 30(5):86–94, 2013.

[Wang *et al.*, 2019] Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled Renyi differential privacy and analytical moments accountant. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019*, volume 89 of *Proceedings of Machine Learning Research*, pages 1226–1235, 2019.