Private Topic Modeling

Park, M.; Foulds, J.; Chaudhuri, K.; Welling, M.

Publication date
2016

Document Version
Final published version

Published in
Private Multi-Party Machine Learning

Private Topic Modeling

Mijung Park
QUVA lab, UvA

James Foulds
USCD

Kamalika Chaudhuri
UCSD

Max Welling
QUVA lab, UvA

Abstract

We develop a privatised stochastic variational inference method for Latent Dirichlet Allocation (LDA). The iterative nature of stochastic variational inference presents challenges: multiple iterations are required to obtain accurate posterior distributions, yet each iteration increases the amount of noise that must be added to achieve a reasonable degree of privacy. We propose a practical algorithm that overcomes this challenge by combining: (1) A relaxed notion of the differential privacy, called *concentrated differential privacy*, which provides high probability bounds for cumulative privacy loss, which is well suited for iterative algorithms, rather than focusing on single-query loss; and (2) *privacy amplification* resulting from subsampling of large-scale data. Focusing on *conjugate exponential* family models, in our private variational inference, all the posterior distributions will be privatised by simply perturbing expected sufficient statistics. Using Wikipedia data, we illustrate the effectiveness of our algorithm for large-scale data.

1 Background

We start by providing background information on the definitions of algorithmic privacy that we use, as well as the general formulation of the variational inference algorithm.

Differential privacy Differential privacy (DP) is a formal definition of the privacy properties of data analysis algorithms [1]. A randomized algorithm $\mathcal{M}(\mathbf{X})$ is said to be (ϵ, δ)-differentially private if $Pr(\mathcal{M}(\mathbf{X}) \in \mathcal{S}) \leq \exp(\epsilon) Pr(\mathcal{M}(\mathbf{X}') \in \mathcal{S}) + \delta$ for all measurable subsets $\mathcal{S}$ of the range of $\mathcal{M}$ and for all datasets $\mathbf{X}$, $\mathbf{X}'$ differing by a single entry. If $\delta = 0$, the algorithm is said to be ϵ-differentially private. Intuitively, the definition states that the output probabilities must not change very much when a single individual's data is modified, thereby limiting the amount of information that the algorithm reveals about any one individual.

Concentrated differential privacy (CDP) [2] is a recently proposed relaxation of DP which aims to make privacy-preserving iterative algorithms more practical than for DP while still providing strong privacy guarantees. The CDP framework treats the *privacy loss* of an outcome, $\log \frac{Pr(\mathcal{M}(\mathbf{X})=o)}{Pr(\mathcal{M}(\mathbf{X}')=o)}$ as a random variable. An algorithm is (μ, τ)-CDP if this privacy loss has mean μ, and after subtracting μ the resulting random variable l is subgaussian with standard deviation τ, i.e. $\forall \lambda \in \mathbb{R} : E[e^{\lambda l}] \leq \exp(\lambda^2 \tau^2 / 2)$. While ϵ-DP guarantees bounded privacy loss, and (ϵ, δ)-DP ensures bounded privacy loss with probability $1 - \delta$, (μ, τ)-CDP requires the privacy loss to be near μ w.h.p.

The general VI algorithm. Consider a generative model that produces a dataset $\mathcal{D} = \{\mathcal{D}_n\}_{n=1}^N$ consisting of N independent identically distributed items, where $\mathcal{D}_n$ is an nth observation, generated using a set of latent variables $\boldsymbol{l} = \{\boldsymbol{l}_n\}_{n=1}^N$. The generative model provides $p(\mathcal{D}_n | \boldsymbol{l}_n, \mathbf{m})$, where $\mathbf{m}$ is the model parameters. We also consider the prior distribution over the model parameters $p(\mathbf{m})$ and the prior distribution over the latent variables $p(\boldsymbol{l})$. Here, we focus on conjugate-exponential (CE) models[1], in which the variational updates are tractable. The CE fam-

[1] A large class of models falls in the CE family including linear dynamical systems and switching models; Gaussian mixtures; factor analysis and probabilistic PCA; hidden Markov models and factorial HMMs; discrete-variable belief networks; and latent Dirichlet allocation (LDA), which we will use in Sec 3.

Privcy workshop (NIPS 2016), Barcelona, Spain.

ily models satisfy the two conditions [3]: (1) Complete-data likelihood is in exponential family: $p(\mathcal{D}_n, \boldsymbol{l}_n|\mathbf{m}) = g(\mathbf{m})f(\mathcal{D}_n, \boldsymbol{l}_n)\exp(\mathbf{n}(\mathbf{m})^\top \mathbf{s}(\mathcal{D}_n, \boldsymbol{l}_n))$, and (2) Prior over $\mathbf{m}$ is conjugate to the complete-data likelihood: $p(\mathbf{m}|\tau, \boldsymbol{\nu}) = h(\tau, \boldsymbol{\nu})g(\mathbf{m})^\tau \exp(\boldsymbol{\nu}^\top \mathbf{n}(\mathbf{m}))$, where natural parameters and sufficient statistics of the complete-data likelihood are denoted by $\mathbf{n}(\mathbf{m})$ and $\mathbf{s}(\mathcal{D}_n, \boldsymbol{l}_n)$, respectively. The hyperparameters are denoted by τ (a scalar) and $\boldsymbol{\nu}$ (a vector).

Variational inference for a CE family model iterates the following two steps in order to optimise the lower bound to the log marginal likelihood, $\mathcal{L}(q(\boldsymbol{l}), q(\mathbf{m})) = \int d\mathbf{m}\, d\boldsymbol{l}\; q(\boldsymbol{l})q(\mathbf{m}) \log \frac{p(\boldsymbol{l},\mathcal{D},\mathbf{m})}{q(\boldsymbol{l})q(\mathbf{m})}$,

(a) given expected natural parameters $\bar{\mathbf{n}}$, the first step computes the approximate posterior over latent variables:

$$q(\boldsymbol{l}) = \prod_{n=1}^{N} q(\boldsymbol{l}_n) \propto \prod_{n=1}^{N} f(\mathcal{D}_n, \boldsymbol{l}_n)\exp(\bar{\mathbf{n}}^\top \mathbf{s}(\mathcal{D}_n, \boldsymbol{l}_n)) = \prod_{n=1}^{N} p(\boldsymbol{l}_n|\mathcal{D}_n, \bar{\mathbf{n}}). \tag{1}$$

Using $q(\boldsymbol{l})$, the first step outputs expected sufficient statistics $\bar{\mathbf{s}}(\mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N} \langle \mathbf{s}(\mathcal{D}_n, \boldsymbol{l}_n)\rangle_{q(\boldsymbol{l}_n)}$.

(b) given expected sufficient statistics $\bar{\mathbf{s}}(\mathcal{D})$, the second step computes the approximate posterior over parameters:

$$q(\mathbf{m}) = h(\tilde{\tau}, \tilde{\boldsymbol{\nu}})g(\mathbf{m})^{\tilde{\tau}} \exp(\tilde{\boldsymbol{\nu}}^\top \mathbf{n}(\mathbf{m})), \text{ where } \tilde{\tau} = \tau + N, \; \tilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + N\bar{\mathbf{s}}(\mathcal{D}). \tag{2}$$

Using $q(\mathbf{m})$, the second step outputs expected natural parameters $\bar{\mathbf{n}} = \langle \mathbf{n}(\mathbf{m})\rangle_{q(\mathbf{m})}$.

2 Privacy Preserving VI algorithm for CE family

The only place where the algorithm looks at the data is when computing the expected sufficient statistics $\bar{\mathbf{s}}(\mathcal{D})$ in the first step. The expected sufficient statistics then dictates the expected natural parameters in the second step. So, perturbing the sufficient statistics leads to perturbing both posterior distributions $q(\boldsymbol{l})$ and $q(\mathbf{m})$. Perturbing sufficient statistics in exponential families is also used in [4]. Existing work focuses on privatising posterior distributions in the context of posterior sampling [5, 6, 7, 8], while our work focuses on privatising approximate posterior distributions for optimisation-based approximate Bayesian inference.

Suppose there are two neighbouring datasets $\mathcal{D}$ and $\tilde{\mathcal{D}}$, where there is only one datapoint difference among them. We also assume that the dataset is pre-processed such that the L_2 norm of any datapoint is less than 1. The maximum difference in the expected sufficient statistics given the datasets, e.g., the L-1 sensitivity of the expected sufficient statistics is given by (assuming $\mathbf{s}$ is a vector of length L) $\Delta\mathbf{s} = \max_{\mathcal{D}_n,\tilde{\mathcal{D}}_n,q(\boldsymbol{l}_n),q(\tilde{\boldsymbol{l}}_n)} \sum_{l=1}^{L} |\frac{1}{N}\mathbb{E}_{q(\boldsymbol{l}_n)}\mathbf{s}_l(\mathcal{D}_n, \boldsymbol{l}_n) - \frac{1}{N}\mathbb{E}_{q(\tilde{\boldsymbol{l}}_n)}\mathbf{s}_l(\tilde{\mathcal{D}}_n, \tilde{\boldsymbol{l}}_n)|$. Under some models like LDA below, the expected sufficient statistic has a limited sensitivity, in which case we add noise to each coordinate of the expected sufficient statistics to compensate the maximum change.

3 Privacy preserving Latent Dirichlet Allocation (LDA)

The most successful topic modeling is based on LDA, where the generative process is given by [9].

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

where each observed word is represented by an indicator vector $\mathbf{w}_{dn}$ (nth word in the dth document) of length V, where V is the number of terms in a fixed vocabulary set. The topic assignment latent variable $\mathbf{z}_{dn}$ is also an indicator vector of length K, where K is the number of topics.

The LDA falls into the CE family, where we think of $\mathbf{z}_{d,1:N}, \boldsymbol{\theta}_d$ as two types of latent variables : $\boldsymbol{l}_d = \{\mathbf{z}_{d,1:N}, \boldsymbol{\theta}_d\}$, and $\boldsymbol{\beta}$ as model parameters $\mathbf{m} = \boldsymbol{\beta}$: (1) Complete-data likelihood per document is in exponential family: $p(\mathbf{w}_{d,1:N}, \mathbf{z}_{d,1:N}, \boldsymbol{\theta}_d|\boldsymbol{\beta}) \propto f(\mathcal{D}_d, \mathbf{z}_{d,1:N}, \boldsymbol{\theta}_d)\exp(\sum_n \sum_k [\log \boldsymbol{\beta}_k]^\top [\mathbf{z}_{dn}^k \mathbf{w}_{dn}])$, where $f(\mathcal{D}_d, \mathbf{z}_{d,1:N}, \boldsymbol{\theta}_d) \propto \exp([\alpha \mathbf{1}_K]^\top [\log \boldsymbol{\theta}_d] + \sum_n \sum_k \mathbf{z}_{dn}^k \log \boldsymbol{\theta}_d^k)$; (2) Conjugate prior over $\boldsymbol{\beta}_k$: $p(\boldsymbol{\beta}_k|\eta \mathbf{1}_V) \propto \exp([\eta \mathbf{1}_V]^\top [\log \boldsymbol{\beta}_k])$, for $k = \{1, \ldots, K\}$. For simplicity, we assume hyperparameters α and η are set manually.

In VI, we assume the posteriors are : (1) Discrete for $q(\mathbf{z}_{dn}) \propto \exp(\mathbf{z}_{dn}^k \log \boldsymbol{\phi}_{dn}^k)$, with variational parameters that capture the posterior probability of topic assignment, $\boldsymbol{\phi}_{dn}^k \propto \exp(\langle \log \boldsymbol{\beta}_k\rangle_{q(\boldsymbol{\beta}_k)}^\top \mathbf{w}_{dn} + \langle \log \boldsymbol{\theta}_d^k\rangle_{q(\boldsymbol{\theta}_d)})$; (2) Dirichlet for $q(\boldsymbol{\theta}_d) \propto$

$\exp({\boldsymbol{\gamma}_d}^\top \log \boldsymbol{\theta}_d)$, with variational parameters $\boldsymbol{\gamma}_d = \alpha \mathbf{1}_K + \sum_{n=1}^N \langle \mathbf{z}_{dn} \rangle_{q(\mathbf{z}_{dn})}$; and (3) Dirichlet for $q(\boldsymbol{\beta}_k) \propto \exp({\boldsymbol{\lambda}_k}^\top \log \boldsymbol{\beta}_k)$, with variational parameters $\boldsymbol{\lambda}_k = \eta \mathbf{1}_V + \sum_d \sum_n \langle \mathbf{z}_{dn}^k \rangle_{q(\mathbf{z}_{dn})} \mathbf{w}_{dn}$. In this case, the expected sufficient statistics is $\bar{\mathbf{s}}_k = \frac{1}{D} \sum_d \sum_n \langle \mathbf{z}_{dn}^k \rangle_{q(\mathbf{z}_{dn})} \mathbf{w}_{dn} = \frac{1}{D} \sum_d \sum_n \boldsymbol{\phi}_{dn}^k \mathbf{w}_{dn}$.

Sensitivity analysis To privatise the variational inference for LDA, we perturb the expected sufficient statistics. While each document has a different document length N_d, we limit the maximum length of any document to N by randomly selecting N words in a document if the number of words in the document is longer than N.

We add Gaussian noise to each coordinate, then map to 0 if the perturbed coordinate becomes negative:

$$\tilde{\bar{\mathbf{s}}}_k^v = \bar{\mathbf{s}}_k^v + Y_k^v, \text{ where } Y_k^v \sim \mathcal{N}(0, \sigma^2), \text{ and } \sigma^2 \geq 2 \log(1.25/\delta_{iter}) (\Delta \bar{\mathbf{s}})^2 / \epsilon_{iter}^2, \tag{3}$$

where $\mathbf{s}_k^v$ is the vth coordinate of a vector of length V: $\bar{\mathbf{s}}_k^v = \frac{1}{D} \sum_d \sum_n \boldsymbol{\phi}_{dn}^k \mathbf{w}_{dn}^v$, and $\Delta \bar{\mathbf{s}}$ is the sensitivity given by

$$\begin{aligned}
\Delta \bar{\mathbf{s}} = & \max_{|\mathcal{D} - \tilde{\mathcal{D}}| = 1} \sqrt{\sum_k \sum_v (\bar{\mathbf{s}}_k^v(\mathcal{D}) - \bar{\mathbf{s}}_k^v(\tilde{\mathcal{D}}))^2}, \\
& = \max_{d, d'} \frac{1}{D} \sqrt{\sum_k \sum_v (\sum_n (\boldsymbol{\phi}_{dn}^k \mathbf{w}_{dn}^v - \boldsymbol{\phi}_{d'n}^k \mathbf{w}_{d'n}^v))^2}, \\
& \leq \max_d \frac{1}{D} \sum_k \sum_v | \sum_n \boldsymbol{\phi}_{dn}^k \mathbf{w}_{dn}^v |, \text{ since L2 norm is less than equal to L1 norm} \\
& \leq \max_d \frac{1}{D} \sum_n (\sum_k \boldsymbol{\phi}_{dn}^k) (\sum_v \mathbf{w}_{dn}^v) \leq \frac{N}{D},
\end{aligned} \tag{4}$$

since $0 \leq \boldsymbol{\phi}_{dn}^k \leq 1$, $\sum_k \boldsymbol{\phi}_{dn}^k = 1$, $\sum_v \mathbf{w}_{dn}^v = 1$, and $\mathbf{w}_{dn}^v \in \{0, 1\}$.

Private stochastic variational learning In a large-scale data setting, it is impossible to handle the entire dataset at once. In such case, stochastic learning using noisy gradients computed on mini-batches of data, a.k.a., stochastic gradient descent (SGD) provides a scalable inference method. While there are a couple of prior work on differentially private SGD (e.g., [10, 11]), privacy amplification due to subsampling combined with CDP composition (which will be described below) has not been used in the the context of variational inference or topic modeling before.

The privacy amplification theorem states the following.

Theorem 1. *(Theorem 1 in [12]) Any $(\epsilon_{iter}, \delta_{iter})$-DP mechanism running on a uniformly sampled subset of data with a sampling ratio ν guarantees (ϵ', δ') differential privacy, where $\epsilon' = \log(1 + \nu(\exp(\epsilon_{iter}) - 1))$ and $\delta' = \nu \delta_{iter}$*

The privacy gain from subsampling allows us to use a much more relaxed privacy budget ϵ_{iter} and the error tolerance δ_{iter} per iteration, to achieve a reasonable level of (ϵ', δ')-DP with a small sampling rate.

Furthermore, the zCDP composition allows a sharper analysis of the per-iteration privacy budget. We first convert DP to zCDP, then use the zCDP composition and finally convert zCDP back to DP (for comparison purposes), for which we use the following lemmas and proposition.

Lemma 1. *(Proposition 1.6 in [13]) The Gaussian mechanism with some noise variance τ and a sensitivity Δ satisfies $\Delta^2/(2\tau)$-zCDP.*

Lemma 2. *(Lemma 1.7 in [13]) If two mechanisms satisfy ρ_1-zCDP and ρ_2-zCDP, respectively, then their composition satisfies $(\rho_1 + \rho_2)$-zCDP.*

Proposition 1. *(Proposition 1.3 in [13]) If $\mathcal{M}$ provides ρ-zCDP, then $\mathcal{M}$ is $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP for any $\rho > 0$.*

So, using Lemma 2 and 3, we obtain $J\Delta^2/(2\tau)$-zCDP after J-composition of the Gaussian mechanism. Using Proposition 4, we convert $J\Delta^2/(2\tau)$-zCDP to $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-DP, where $\rho = J\Delta^2/(2\tau)$.

These seemingly complicated steps can be summarised into two simple steps. First, given a total privacy budget ϵ_{tot} and total tolerance level δ_{tot}, our algorithm calculates an intermediate privacy

Algorithm 1 Private Topic Modeling

```
Require: Data $\mathcal{D}$. Define $D$ (documents), $V$ (vocabulary), $K$ (number of topics).
         Define $\rho_t = (\tau_0 + t)^{-\kappa}$, mini-batch size $S$, and , and hyperparameters $\alpha, \eta$.
Ensure: Privatised expected natural parameters $\langle \log \boldsymbol{\beta}_k \rangle_{q(\boldsymbol{\beta}_k)}$ and sufficient statistics $\tilde{\bar{\mathbf{s}}}$.
  Compute the per-iteration privacy budget $(\epsilon_{iter}, \delta_{iter})$ using eq (5) and eq (6).
  Compute the sensitivity of the expected sufficient statistics given in eq (4).
  for $t = 1, \ldots, J$ do
    (1) E-step: Given expected natural parameters $\langle \log \boldsymbol{\beta}_k \rangle_{q(\boldsymbol{\beta}_k)}$
    for $d = 1, \ldots, S$ do
      Compute $q(\mathbf{z}_{dn}^k)$ parameterised by $\boldsymbol{\phi}_{dn}^k \propto \exp(\langle \log \boldsymbol{\beta}_k \rangle_{q(\boldsymbol{\beta}_k)}{}^\top \mathbf{w}_{dn} + \langle \log \boldsymbol{\theta}_d^k \rangle_{q(\boldsymbol{\theta}_d)})$.
      Compute $q(\boldsymbol{\theta}_d)$ parameterised by $\boldsymbol{\gamma}_d = \alpha \mathbf{1}_K + \sum_{n=1}^N \langle \mathbf{z}_{dn} \rangle_{q(\mathbf{z}_{dn})}$.
    end for
    Output the noised-up expected sufficient statistics $\tilde{\bar{\mathbf{s}}}_k^v = \frac{1}{S} \sum_d \sum_n \boldsymbol{\phi}_{dn}^k \mathbf{w}_{dn}^v + Y_k^v$, where $Y_k^v$ is
    Gaussian noise given in eq (3).
    (2) M-step: Given noised-up expected sufficient statistics $\tilde{\bar{\mathbf{s}}}_k$,
    Compute $q(\boldsymbol{\beta}_k)$ parameterised by $\boldsymbol{\lambda}_k^{(t)} = \eta \mathbf{1}_V + D \tilde{\bar{\mathbf{s}}}_k$.
    Set $\boldsymbol{\lambda}^{(t)} \leftarrow (1 - \rho_t) \boldsymbol{\lambda}^{(t-1)} + \rho_t \boldsymbol{\lambda}^{(t)}$.
    Output expected natural parameters $\langle \log \boldsymbol{\beta}_k \rangle_{q(\boldsymbol{\beta}_k)}$.
  end for
```

budget using the zCDP composition, which maps $(\epsilon_{tot}, \delta_{tot})$ to (ϵ', δ'),

$$\epsilon_{tot} = J\Delta^2/(2\tau) + 2\sqrt{J\Delta^2/(2\tau)\log(1/\delta_{tot})}, \text{ where } \tau \geq 2\log(1.25/\delta')\Delta^2/\epsilon'^2. \tag{5}$$

Second, our algorithm calculates the per-iteration privacy budget using the privacy amplification theorem, which maps (ϵ', δ') to $(\epsilon_{iter}, \delta_{iter})$,

$$\begin{aligned} \epsilon' &= \log(1 + \nu(\exp(\epsilon_{iter}) - 1)), \\ \delta' &= \nu\delta_{iter}. \end{aligned} \tag{6}$$

Algorithm 1 summarizes our private topic modeling algorithm.

4 Experiments using Wikipedia data

We randomly downloaded $D = 400,000$ documents from Wikipedia. We then tested our VIPS algorithm on the Wikipedia dataset with four different values of total privacy budget, using a mini-batch size $S = \{10, 20, 50, 100, 200, 400\}$, until the algorithm sees up to $160,000$ documents. We assumed there are 100 topics, and we used a vocabulary set of approximately 8000 terms.

We compare our method to two baseline methods. First, in *linear* (Lin) composition (Theorem 3.16 of [1]), privacy degrades linearly with the number of iterations. This result is from the Max Divergence of the privacy loss random variable being bounded by a total budget. Hence, the linear composition yields $(J\epsilon', J\delta')$-DP. We use eq (6) to map (ϵ', δ') to $(\epsilon_{iter}, \delta_{iter})$. Second, *advanced* (Adv) composition (Theorem 3.20 of [1]), resulting from the Max Divergence of the privacy loss random variable being bounded by a total budget including a slack variable δ, yields $(J\epsilon'(e^{\epsilon'} - 1) + \sqrt{2J\log(1/\delta'')}\epsilon',\ \delta'' + J\delta')$-DP. Similarly, we use eq (6) to map (ϵ', δ') to $(\epsilon_{iter}, \delta_{iter})$.

As an evaluation metric, we compute the upper bound to the perplexity on held-out documents[2],

$$\text{perplexity}(\mathcal{D}^{test}, \boldsymbol{\lambda}) \leq \exp\left[-\left(\sum_i \langle \log p(\mathbf{n}^{test}, \boldsymbol{\theta}_i, \mathbf{z}_i | \boldsymbol{\lambda}) \rangle_{q(\boldsymbol{\theta}_i, \mathbf{z}_i)} - \langle \log q(\boldsymbol{\theta}, \mathbf{z}) \rangle_{q(\boldsymbol{\theta}, \mathbf{z})}\right) / \sum_{i,n} \mathbf{n}_{i,n}^{test}\right],$$

where $\mathbf{n}_i^{test}$ is a vector of word counts for the ith document, $\mathbf{n}^{test} = \{\mathbf{n}_i^{test}\}_{i=1}^I$. In the above, we use the $\boldsymbol{\lambda}$ that was calculated during training. We compute the posteriors over $\mathbf{z}$ and $\boldsymbol{\theta}$ by performing the first step in our algorithm using the test data and the perturbed sufficient statistics we obtain during training. The per-word-perplexity is shown in Fig. 1. Due to privacy amplification, it is more beneficial to decrease the amount of noise to add when the mini-batch size is small. The zCDP composition results in a better accuracy than the advanced composition.

[2] We used the metric written in the python implementation by authors of [14].

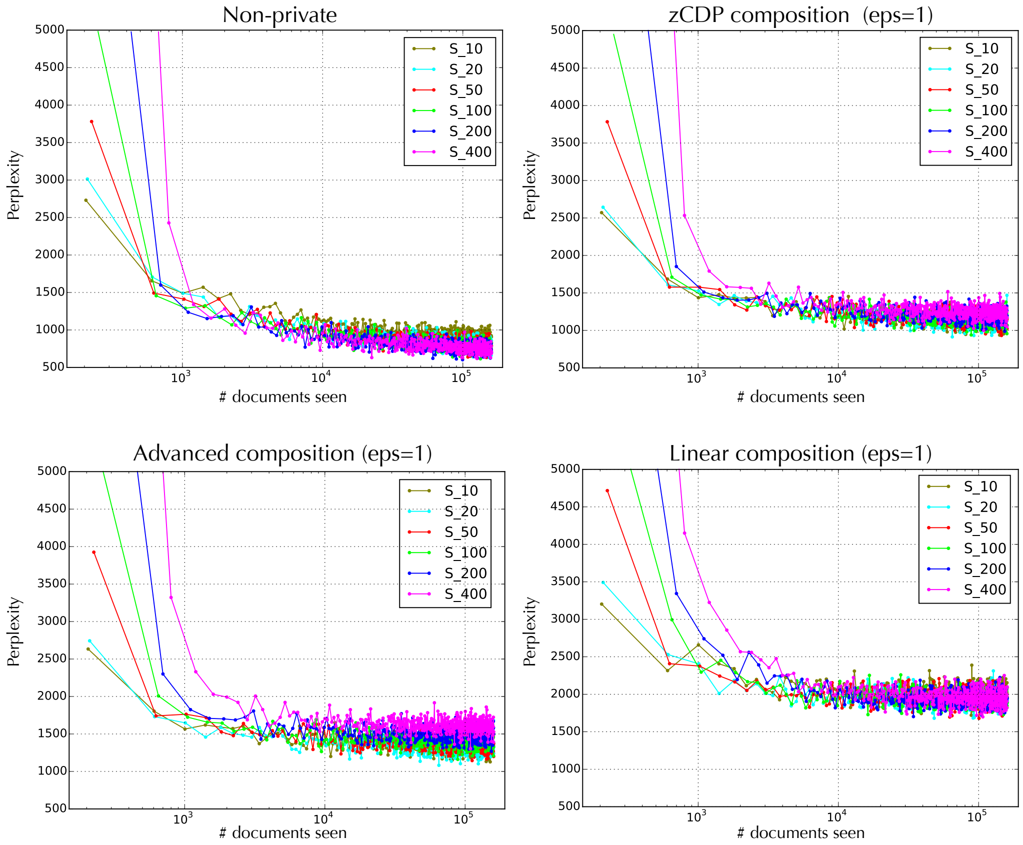

Figure 1: Per-word-perplexity with different mini-batch sizes $S \in \{10, 20, 50, 100, 200, 400\}$. In the private LDA (Top/Right and Bottom), smaller mini-batch size achieves lower perplexity, due to the privacy amplification lemma (See Sec 3). We set the total privacy budget $\epsilon_{tot} = 1$ and the total tolerance $\delta_{tot} = 1e - 4$ in all private methods. Regardless of the mini-batch size, the zCDP composition (Top/Right) achieves a lower perplexity than the Advanced (Bottom/Left) and Linear compositions (Bottom/Right).

In Table 1, we show the top 10 words in terms of assigned probabilities under a chosen topic in each method. We show 4 topics as examples. Non-private LDA results in the most coherent words among all the methods. For the private LDA models with a total privacy budget fixed to 0.5, as we move from zCDP, to advanced, and to linear composition, the amount of noise added gets larger, and therefore more topics have less coherent words.

Table 1: Posterior topics from private and non-private LDA

Non-private		**zCDP (eps=0.5)**		**Adv (eps=0.5)**		**Lin (eps=0.5)**	
topic 3:		**topic 81**:		**topic 72**:		**topic 27**:	
david	0.0667	born	0.0882	fragment	0.0002	horn	0.0002
king	0.0318	american	0.0766	gentleness	0.0001	shone	0.0001
god	0.0304	name	0.0246	soit	0.0001	age	0.0001
son	0.0197	actor	0.0196	render	0.0001	tradition	0.0001
israel	0.0186	english	0.0179	nonproprietary	0.0001	protecting	0.0001
bible	0.0156	charles	0.0165	westminster	0.0001	fils	0.0001
hebrew	0.0123	british	0.0138	proceedings	0.0001	trip	0.0001
story	0.0102	richard	0.0130	clare	0.0001	article	0.0001
book	0.0095	german	0.0119	stronger	0.0001	interests	0.0001
adam	0.0092	character	0.0115	hesitate	0.0001	incidents	0.0001
topic 4:		**topic 82**:		**topic 73**:		**topic 28**:	
university	0.1811	wat	0.0002	mount	0.0034	american	0.0228
press	0.0546	armed	0.0001	display	0.0011	born	0.0154
oxford	0.0413	log	0.0001	animal	0.0011	john	0.0107
italy	0.0372	fierce	0.0001	equipment	0.0011	name	0.0094
jacques	0.0359	infantry	0.0001	cynthia	0.0009	english	0.0062
cambridge	0.0349	sehen	0.0001	position	0.0008	actor	0.0061
barbara	0.0280	selbst	0.0001	systems	0.0008	united	0.0058
research	0.0227	clearly	0.0001	support	0.0008	british	0.0051
murray	0.0184	bull	0.0001	software	0.0008	character	0.0051
scientific	0.0182	recall	0.0001	heavy	0.0008	people	0.0048
topic 5:		**topic 83**:		**topic 74**:		**topic 29**:	
association	0.0896	david	0.0410	david	0.0119	shelter	0.0001
security	0.0781	jonathan	0.0199	king	0.0091	rome	0.0001
money	0.0584	king	0.0188	god	0.0072	thick	0.0001
joint	0.0361	samuel	0.0186	church	0.0061	vous	0.0001
masters	0.0303	israel	0.0112	samuel	0.0054	leg	0.0001
banks	0.0299	saul	0.0075	son	0.0051	considering	0.0001
seal	0.0241	son	0.0068	israel	0.0039	king	0.0001
gilbert	0.0235	dan	0.0067	name	0.0038	object	0.0001
trade	0.0168	god	0.0053	century	0.0038	prayed	0.0001
heads	0.0166	story	0.0048	first	0.0036	pilot	0.0001
topic 6:		**topic 84**:		**topic 75**:		**topic 30**:	
law	0.0997	simon	0.0101	recognise	0.0001	despair	0.0001
court	0.0777	cat	0.0008	comparison	0.0001	ray	0.0001
police	0.0442	maison	0.0005	violates	0.0001	successfully	0.0001
legal	0.0396	breach	0.0005	offices	0.0001	respectable	0.0001
justice	0.0292	says	0.0005	value	0.0001	acute	0.0001
courts	0.0229	dirty	0.0005	neighbor	0.0001	accompany	0.0001
welcome	0.0204	rifle	0.0004	cetait	0.0001	assuming	0.0001
civil	0.0178	door	0.0004	composed	0.0001	florence	0.0001
signal	0.0170	property	0.0004	interests	0.0001	ambition	0.0001
pan	0.0163	genus	0.0004	argue	0.0001	unreasonable	0.0001

Using the same data, we then tested how perplexity changes as we change the mini-batch sizes. As shown in Fig. 1, due to privacy amplification, it is more beneficial to decrease the amount of noise to add when the mini-batch size is small. The zCDP composition results in a better accuracy than the advanced composition.

5 Conclusion

We have developed a practical privacy-preserving topic modeling algorithm which outputs accurate and privatized expected sufficient statistics and expected natural parameters. Our approach uses the zCDP composition analysis combined with the privacy amplification effect due to subsampling of data, which significantly decrease the amount of additive noise for the same expected privacy guarantee compared to the standard analysis.

References

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

[2] C. Dwork and G. N. Rothblum. Concentrated Differential Privacy. *ArXiv e-prints*, March 2016.

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

[4] Mijung Park, Jimmy Foulds, Kamalika Chaudhuri, and Max Welling. Practical privacy for expectation maximization. *CoRR*, abs/1605.06995, 2016.

[5] Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private Bayesian inference. In *Algorithmic Learning Theory (ALT)*, pages 291–305. Springer, 2014.

[6] Zuhe Zhang, Benjamin Rubinstein, and Christos Dimitrakakis. On the differential privacy of Bayesian inference. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[7] James R. Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *CoRR*, abs/1603.07294, 2016.

[8] Gilles Barthe, Gian Pietro Farina, Marco Gaboardi, Emilio Jesús Gallego Arias, Andy Gordon, Justin Hsu, and Pierre-Yves Strub. Differentially private Bayesian programming. *CoRR*, abs/1605.00283, 2016.

[9] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013.

[10] Xi Wu, Arun Kumar, Kamalika Chaudhuri, Somesh Jha, and Jeffrey F. Naughton. Differentially private stochastic gradient descent for in-rdbms analytics. *CoRR*, abs/1606.04722, 2016.

[11] Y.-X. Wang, S. E. Fienberg, and A. Smola. Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo. *ArXiv e-prints*, February 2015.

[12] 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.

[13] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *CoRR*, abs/1605.02065, 2016.

[14] Matthew Hoffman, Francis R. Bach, and David M. Blei. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc., 2010.