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

Citation for published version (APA):
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 \( \mathcal{M}(X) \) is said to be \((\epsilon, \delta)\)-differentially private if \( \Pr(\mathcal{M}(X) \in S) \leq \exp(\epsilon) \Pr(\mathcal{M}(X') \in S) + \delta \) for all measurable subsets \( S \) of the range of \( \mathcal{M} \) and for all datasets \( X, X' \) differing by a single entry. If \( \delta = 0 \), the algorithm is said to be \( \epsilon \)-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, \( \log \frac{Pr(\mathcal{M}(X) = o)}{Pr(\mathcal{M}(X') = o)} \) as a random variable. An algorithm is \((\mu, \tau)\)-CDP if this privacy loss has mean \( \mu \), and after subtracting \( \mu \) the resulting random variable \( l \) is subgaussian with standard deviation \( \tau \), i.e. \( \forall \lambda \in \mathbb{R} : E[e^{\lambda l}] \leq \exp(\lambda^2 \tau^2 / 2) \). While \( \epsilon \)-DP guarantees bounded privacy loss, and \((\epsilon, \delta)\)-DP ensures bounded privacy loss with probability \( 1 - \delta \), \((\mu, \tau)\)-CDP requires the privacy loss to be near \( \mu \) w.h.p.

The general VI algorithm. Consider a generative model that produces a dataset \( D = \{D_n\}_{n=1}^{N} \) consisting of \( N \) independent identically distributed items, where \( D_n \) is an \( n \)th observation, generated using a set of latent variables \( I = \{I_n\}_{n=1}^{N} \). The generative model provides \( p(D_n | I_n, m) \), where \( m \) is the model parameters. We also consider the prior distribution over the model parameters \( p(m) \) and the prior distribution over the latent variables \( p(I) \). Here, we focus on conjugate-exponential (CE) models 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.
The LDA falls into the CE family, where we think of the only place where the algorithm looks at the data is when computing the expected sufficient statistics of the complete-data likelihood are denoted by \( \mathbf{n}(\mathbf{m}) \) and \( \mathbf{s}(\mathbf{D}_n, I_n) \), respectively. The hyperparameters are denoted by \( \tau \) (a scalar) and \( \nu \) (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, \( \mathcal{L}(q(l), q(\mathbf{m})) = \int d\mathbf{m} d\nu q(\mathbf{m}) \log \frac{p(l, \mathbf{D}, \mathbf{m})}{q(l)q(\mathbf{m})} \), (a) given expected natural parameters \( \bar{\mathbf{n}} \), the first step computes the approximate posterior over latent variables:

\[
q(l) = \prod_{n=1}^{N} q(l_n) \propto \prod_{n=1}^{N} f(\mathbf{D}_n, I_n) \exp(\mathbf{n}^\top \mathbf{s}(\mathbf{D}_n, I_n)) = \prod_{n=1}^{N} p(l_n|\mathbf{D}_n, \bar{\mathbf{n}}).
\]

Using \( q(l) \), the first step outputs expected sufficient statistics \( \bar{\mathbf{s}}(\mathbf{D}) = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{s}(\mathbf{D}_n, I_n))_{q(l_n)} \).

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

\[
q(\mathbf{m}) = h(\bar{\tau}, \bar{\nu}) \mathbf{m}^\top \exp(\bar{\nu}^\top \mathbf{n}(\mathbf{m})), \quad \text{where} \quad \bar{\tau} = \tau + N, \quad \bar{\nu} = \nu + N\bar{\mathbf{s}}(\mathbf{D}).
\]

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

## 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 \( \bar{\mathbf{s}}(\mathbf{D}) \) 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 \( q(l) \) and \( q(\mathbf{m}) \). Perturbing sufficient statistics in exponential families is also used in \cite{privacy}. Existing work focuses on privatising posterior distributions in the context of posterior sampling \cite{privacy2, privacy3, privacy4, privacy5}, while our work focuses on privatising approximate posterior distributions for optimisation-based approximate Bayesian inference.

Suppose there are two neighbouring datasets \( \mathbf{D} \) and \( \bar{\mathbf{D}} \), 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 \( L \))

\[
\Delta \mathbf{s} = \max_{\mathbf{D}_n, \mathbf{D}_n^{'}, q(l_n), q(l_n')} \sum_{l=1}^{L} |\frac{1}{N} \mathbf{E}_{q(l_n)} \mathbf{s}(\mathbf{D}_n, I_n) - \frac{1}{N} \mathbf{E}_{q(l_n')} \mathbf{s}(\mathbf{D}_n, I_n')|.
\]

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 \cite{privacy6}.

- Draw topics \( \beta_k \sim \text{Dirichlet}(\eta 1_K) \), for \( k = \{1, \ldots, K\} \), where \( \eta \) is a scalar hyperparameter.
- For each document \( d \in \{1, \ldots, D\} \)
  - Draw topic proportions \( \theta_d \sim \text{Dirichlet}(\alpha 1_K) \), where \( \alpha \) is a scalar hyperparameter.
  - For each word \( n \in \{1, \ldots, N\} \)
    - Draw topic assignments \( z_{dn} \sim \text{Discrete}(\theta_d) \)
    - Draw word \( w_{dn} \sim \text{Discrete}(\beta_{z_{dn}}) \)

where each observed word is represented by an indicator vector \( w_{dn} \) (n-th word in the \( d \)-th document) of length \( V \), where \( V \) is the number of terms in a fixed vocabulary set. The topic assignment latent variable \( z_{dn} \) 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 \( z_{d,1:N}, \theta_d \) as two types of latent variables : \( l_d = \{z_{d,1:N}, \theta_d\} \), and \( \beta \) as model parameters \( \mathbf{m} = \beta \). (1) Complete-data likelihood per document is in exponential family:

\[
p(w_{\mathbf{d}}, z_{d,1:N}, \theta_d | \beta) \propto f(D_d, z_{d,1:N}, \theta_d) \exp(\sum_n \sum_k [\log \beta_k]^{\top} [z_{kn} w_{dn}]),
\]

where \( f(D_d, z_{d,1:N}, \theta_d) \propto \exp([\alpha 1_K]^{\top} [\log \theta_d] + \sum_n \sum_k z_{kn}^{\top} \log \beta_k) \); (2) Conjugate prior over \( \beta_k \): \( p(\beta_k | \eta 1_V) \propto \exp([\eta 1_V]^{\top} [\log \beta_k]) \), for \( k = \{1, \ldots, K\} \). For simplicity, we assume hyperparameters \( \alpha \) and \( \eta \) are set manually.

In VI, we assume the posteriors are : (1) Discrete for \( q(z_{dn}) \propto \exp(z_{dn}^{\top} \log \phi_{z_{dn}}) \), with variational parameters that capture the posterior probability of topic assignment, \( \phi_{z_{dn}} \propto \exp([\log \beta_k]_{q(\beta_k)}^{\top} w_{dn} + \log \theta_d^{\top} [\log \theta_d]) \); (2) Dirichlet for \( q(\theta_d) \propto \exp([\log \beta_k]_{q(\beta_k)}^{\top} \theta_d + \log \theta_d^{\top} [\log \theta_d]) \).
We add Gaussian noise to each coordinate, then map to 0 if the perturbed coordinate becomes

\[ \rho > \frac{1}{\gamma} \]

where \( \rho \) is 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 \( N \).

These seemingly complicated steps can be summarised into two simple steps. First, given a total privacy budget \( \epsilon \), we subsample a subset of data with a sampling ratio \( \nu \). So, using Lemma 2 and 3, we obtain

\[ J \Delta^2 / (2\tau) \]-zCDP for any \( \Delta \) satisfies \( J \Delta^2 / (2\tau) \)-zCDP.

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:

\[ \tilde{s}^k_v = s^k_v + Y^v_k, \]

where \( Y^v_k \sim \mathcal{N}(0, \sigma^2) \), and \( \sigma^2 \geq 2 \log(1.25/\delta_{\text{iter}})/\epsilon_{\text{iter}}^2 \),

where \( s^k_v \) is the \( v \)th coordinate of a vector of length \( V \): \( s^k_v = \frac{1}{D} \sum_d \sum_n \phi^k_{dn} w^n_{dv} \), and \( \Delta s \) is the sensitivity given by

\[
\Delta s = \max_{|D - \bar{D}| = 1} \left| \frac{1}{D} \sum_k \sum_v (s^k_v(\bar{D}) - s^k_v(\bar{D}))^2 \right|,
\]

\[
= \max_{d,\nu} \frac{1}{D} \sum_k \sum_v \left( \phi^k_{dn} w^n_{dv} - \phi^k_{d'n} w^n_{d'v} \right)^2,
\]

\[
\leq \max_{d} \frac{1}{D} \sum_k \sum_v | \phi^k_{dn} | (\sum_v w^n_{dv}) \leq \frac{N}{D},
\]

since \( 0 \leq \phi^k_{dn} \leq 1 \), \( \sum_k \phi^k_{dn} = 1 \), \( \sum_v w^n_{dv} = 1 \), and \( w^n_{dv} \in \{0, 1\} \).

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 \((\epsilon_{\text{iter}}, \delta_{\text{iter}})\)-DP mechanism running on a uniformly sampled subset of data with a sampling ratio \( \nu \) guarantees \((\epsilon', \delta')\) differential privacy, where \( \epsilon' = \log(1 + \nu(\exp(\epsilon_{\text{iter}}) - 1)) \) and \( \delta' = \nu \delta_{\text{iter}} \).

The privacy gain from subsampling allows us to use a much more relaxed privacy budget \( \epsilon_{\text{iter}} \) and the error tolerance \( \delta_{\text{iter}} \) per iteration, to achieve a reasonable level of \((\epsilon', \delta')\)-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 \( \tau \) and a sensitivity \( \Delta \) satisfies \( \Delta^2 / (2\tau) \)-zCDP.

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

Proposition 1. (Proposition 1.3 in \[13\]) If \( M \) provides \( \rho \)-zCDP, then \( M \) 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 \( \epsilon_{\text{tot}} \) and total tolerance level \( \delta_{\text{tot}} \), our algorithm calculates an intermediate privacy
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 [11]), 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 \((\epsilon', \delta')\) to \((\epsilon_{iter}, \delta_{iter})\).

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

\[
\text{perplexity}(D^\text{test}, \lambda) \leq \exp \left[ - \left( \sum_i \langle \log p(n^\text{test}_i, \theta, z, \lambda) \rangle_{q(\theta, z)} - \langle \log q(\theta, z) \rangle_{q(\theta, z)} \right) / \sum_i n^\text{test}_i \right],
\]

where \( n^\text{test}_i \) is a vector of word counts for the \( i \)th document, \( n^\text{test} = \{ n^\text{test}_i \}_{i=1}^I \). In the above, we use the \( \lambda \) that was calculated during training. We compute the posteriors over \( z \) and \( \theta \) 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.

---

<sup>2</sup> 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 $\epsilon_{\text{tot}} = 1$ and the total tolerance $\delta_{\text{tot}} = 1e^{-4}$ 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.
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.

5 Conclusion

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.
References


