Improving Privacy Guarantee and Efficiency of Latent Dirichlet Allocation Model Training Under Differential Privacy

Tao Huang[1,2], **Hong Chen**[1,2]
[1] Key Laboratory of Data Engineering and Knowledge Engineering of Ministry of Education
[2] School of Information, Renmin University of China
huang-tao@ruc.edu.cn
chong@ruc.edu.cn

Abstract

Latent Dirichlet allocation (LDA), a widely used topic model, is often employed as a fundamental tool for text analysis in various applications. However, the training process of the LDA model typically requires massive text corpus data. On one hand, such massive data may expose private information in the training data, thereby incurring significant privacy concerns. On the other hand, the efficiency of the LDA model training may be impacted, since LDA training often needs to handle these massive text corpus data. To address the privacy issues in LDA model training, some recent works have combined LDA training algorithms that are based on collapsed Gibbs sampling (CGS) with differential privacy. Nevertheless, these works usually have a high accumulative privacy budget due to vast iterations in CGS. Moreover, these works always have low efficiency due to handling massive text corpus data. To improve the privacy guarantee and efficiency, we combine a subsampling method with CGS and propose a novel LDA training algorithm with differential privacy, SUB-LDA. We find that subsampling in CGS naturally improves efficiency while amplifying privacy. We propose a novel metric, the efficiency–privacy function, to evaluate improvements of the privacy guarantee and efficiency. Based on a conventional subsampling method, we propose an adaptive subsampling method to improve the model's utility produced by SUB-LDA when the subsampling ratio is small. We provide a comprehensive analysis of SUB-LDA, and the experiment results validate its efficiency and privacy guarantee improvements.

1 Introduction

Latent Dirichlet allocation (LDA)(Blei et al., 2003) is a widely used topic model to discover the latent semantic of text data. High-dimensional text data can be mapped to low-dimensional latent topic space via LDA. Thus, LDA simplifies subsequent text analysis tasks, such as similarity judgment. Platforms based on LDA for analyzing various text data have been established by many enterprises, such as Tencent (Wang et al., 2014)(Yut et al., 2017) and Microsoft (Yuan et al., 2015).

Differential privacy (DP) is a de-facto standard of privacy protection definition with a rigorous mathematical proof and is widely used for quantifying the privacy risks of random algorithms. To address privacy issues when touching datasets containing sensitive information in the training process of LDA, some works (Park et al., 2016)(Zhu et al., 2016)(Wang et al., 2020)(Zhao et al., 2019)(Zhao et al., 2020) combine DP with LDA. In this study, we focus on LDA training algorithms based on collapsed Gibbs sampling (CGS).

HDP-LDA, proposed by Zhao et al (Zhao et al., 2020), has been demonstrated to be effective and outperforms other relevant works (Park et al., 2016)(Zhu et al., 2016)(Zhao et al., 2019) when protecting sensitive word-count information in CGS training. HDP-LDA injects noise into word counts in each training iteration. However, this method suffers from worse efficiency when dealing with massive text corpus data. Moreover, even when HDP-LDA chooses a small privacy budget in each iteration, the accumulative privacy budget during the whole training may be very large due to a mass of iterations.

Subsampling is a widely used method to achieve privacy amplification in differentially private algorithms (Dwork et al., 2014)(Balle et al., 2020)(Zhu and Wang, 2019)(Wang et al., 2019)(Mironov et al., 2019). A subsampled randomized algorithm takes a subsample of the original dataset generated by some subsampling procedure, and then applies a known randomized mechanism to the subsampled data. When introducing a subsampling operation in CGS, we discover that subsampling naturally improves the efficiency of CGS while amplifying privacy.

Moreover, a natural question is whether we can

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 143–152
November 7–11, 2021.

amplify privacy while improving the efficiency of CGS simultaneously. We need a metric to evaluate the efficiency–privacy improvement.

In this study, we propose a subsampling solution to improve the privacy guarantee and efficiency of HDP-LDA. We call our novel LDA training algorithm with differential privacy SUB-LDA. Then, we propose a novel metric, the efficiency–privacy function, to evaluate the privacy guarantee and efficiency improvements of SUB-LDA. When the subsampling ratio is small, the model always suffers from heavy utility loss. We propose an adaptive subsampling (AS) method to mitigate the dilemma. Our contributions are summarized as follows.

- We combine subsampling with HDP-LDA, a general differentially private CGS algorithm, and propose our SUB-LDA algorithm, which provides a better privacy guarantee and efficiency than existing methods.
- We propose a novel metric, called the efficiency–privacy function, and provide a comprehensive analysis of SUB-LDA and how the metric behaves when we change the subsampling ratio. We find that we can improve efficiency and privacy guarantee simultaneously only under a certain range of the subsampling ratio.
- We propose an AS method that can be used to improve the model's utility. We conduct extensive experiments on several real-world datasets to validate the effectiveness of SUB-LDA. The experiments show that SUB-LDA achieves better efficiency and amplifies privacy.

2 Related works

We divide related works into the following three categories.

(a)LDA training with differential privacy

As a widely used machine learning model, LDA with DP has attracted the interest of researchers. Zhu et al.(Zhu et al., 2016) propose a privacy-preserving tag release algorithm. To protect intermediate private weight information, they add Laplace noise to the weights in the last iteration of CGS. Zhao et al. (Zhao et al., 2019) propose a locally private LDA training algorithm on crowdsourced data to provide local DP for individual data contributors. (Zhao et al., 2020) propose a centralized privacy-preserving algorithm that can prevent data inference from the intermediate statistics in CGS training. Variational Bayes for parameter estimation of LDA is the focus of (Park et al., 2016). In this study, we aim to provide an LDA model trained via CGS with a DP guarantee under a centralized situation.

(b)Subsampled differential privacy

Since machine learning algorithms always handle massive sensitive data and perform many iterations before finding the optimal solution, limitations arise when we want to bound the privacy budget of iterative machine learning algorithms. Privacy amplification by subsampling has gradually attracted the interest of researchers. Wang et al. (Wang et al., 2019) propose a general "RDP-amplification" bound that applies to any randomized mechanism equipped with subsampling without replacements. However, this bound is a constant factor away from being optimal. Zhu and Wang (Zhu and Wang, 2019) provide a more general result of tighter RDP-amplification bound under Poisson subsampling. Mironov (Mironov, 2017) discuss the special sampled Gaussian mechanism, which is successfully used in several machine learning applications. They describe a numerically stable procedure for precise computation of sampled Gaussian Mechanism's Rényi Differential Privacy (RDP) and prove a nearly tight closed-form bound. Dwork et al. (Dwork et al., 2014) give a general bound of privacy loss of the subsampled mechanism in terms of (ε, δ)-DP. Balle et al. (Balle et al., 2020) improve the bound and propose a general framework to derive tight bound of privacy loss of the subsampled mechanism in terms of (ε, δ)-DP.

(c)Efficient collapsed Gibbs sampling

CGS is a widely used method to train the LDA model. However, the complexity of traditional CGS is $O(NZ)$, which is a large number, where N and Z are the total number of words and latent topics in text corpus. To improve the efficiency of traditional CGS, some efficient CGS algorithms (Porteous et al., 2008)(Yao et al., 2009)(Li et al., 2014)(Yuan et al., 2015)(Hu et al., 2017) have been proposed recently. *FastLDA* (Porteous et al., 2008) reduces operations per sample to improve the efficiency of CGS. Yao et al. (Yao et al., 2009) obtain better efficiency of CGS by reducing the complexity $O(NZ)$ of traditional CGS to $O(N(Z_w + Z_d))$, where Z_w and Z_d are the numbers of distinct topics that are assigned to a word w and a document d,

respectively. Usually, $Z_w + Z_d$ is much smaller than Z. Li et al. (Li et al., 2014) utilize the sparsity in the topic model and reduce the complexity from $O(NZ)$ to $O(NZ_d)$ by combining Metropolis–Hasting sampling and the alias table method (Walker, 1977). Yuan et al. (Yuan et al., 2015) propose a compute-and-memory efficient distributed LDA implementation, called *LightLDA*. The complexity of LightLDA is $O(N)$. Hu et al. (Hu et al., 2017) observe that topic distributions of words are skewed, and only a subset of documents can approximately represent the semantics of the whole corpus. They reduce N via *approximate semantics* and reduce Z via *skewed topic distribution*.

3 Preliminaries

3.1 Latent Dirichlet Allocation and Collapsed Gibbs Sampling

The LDA model is widely used to discover the latent structures of text corpus datasets. The latent structures are depicted as probability distributions with prior and obtained posterior distributions after training via Bayes rules. Text corpus is considered a mixture of K different latent topics, and each document m in text corpus is represented by a K-dimensional document-topic distribution θ_m. Moreover, each latent topic k is represented by a V-dimensional topic-word distribution ϕ_k where V is the total number of unique words in text corpus.

The CGS training process aims to discover topic-word distribution ϕ_k. For each word w_i, CGS samples a new topic z_i based on the following full conditional distribution:

$$p\left(z_i = k \mid \vec{z}_{\neg i}, \vec{w}\right) \propto \frac{n_k^t + \beta}{\sum_{t=1}^{V}\left(n_k^t + \beta\right)} \cdot \frac{n_m^k + \tau}{\sum_{k=1}^{K}\left(n_m^k + \tau\right)} \quad (1)$$

where $\neg i$ denotes the whole words in text corpus without the absence of word w_i, n_m^k is the count of topic k that appeared in document m, and n_k^t is the count of topic k assigned to word t. τ is the document-topic prior hyper-parameter and β is the topic-word prior hyper-parameter. CGS runs over three periods: initialization, burn-in, and estimation. During initialization, each word w in text corpus is randomly assigned to a topic $k \in K$. Then, the document-topic count n_m^k and topic-word count n_k^t are obtained. In the subsequent burn-in process, the topic assignment for each word w is updated via sampling from a multinomial distribution $\mathbf{P} = [p_1, \ldots, p_k, \ldots, p_K]$, where p_k is calculated according to equation (1). After a series of iterations, the burn-in process ends and we can estimate ϕ_k^t by

$$\mathbb{E}\left[\phi_k^t \mid \mathbf{z}, \mathbf{w}\right] = \frac{n_k^t + \beta}{\sum_{t=1}^{V}\left(n_k^t + \beta\right)} \quad (2)$$

More details about LDA and CGS can be found in (Porteous et al., 2008)(Xiao and Stibor, 2010)(MacKay and Mac Kay, 2003)(Carlo, 2004)(Liu, 2008).

Since counting n_k^t needs to touch original dataset, n_k^t is considered as sensitive information and thus needs to be protected. HDP-LDA (Zhao et al., 2020) suggests adding noise, for example, Laplace noise, to each n_k^t independently in each iteration of CGS. Thus, even the adversary can monitor the whole training process of CGS, n_k^t in each iteration could be protected.

3.2 Poisson Subsampled Rényi Differential Privacy

In this subsection, we introduce background on DP, RDP, Poisson subsampling, and its privacy-amplification effects.

DP has been embraced by multiple research communities as a standard principle of privacy for algorithms. DP bounds a shift in the output distribution of a randomized algorithm when a small change is induced in its input.

Definition 3.1((ε, δ)-DP) (Dwork et al., 2014). A randomized mechanism $f : \mathcal{G} \mapsto \mathcal{R}$ offers (ε, δ)-DP if for any adjacent $G, G' \in \mathcal{G}$ and $R \in \mathcal{R}$ $\Pr[f(G) \in R] \leq e^{\varepsilon} \Pr\left[f\left(G'\right) \in R\right] + \delta$.

This definition restrains an adversary's ability to infer whether the input dataset is G or G'. RDP is a refinement of DP. RDP utilizes Rényi-divergence as a distance metric instead of sup-divergence in DP.

Definition 3.2((α, ε)-RDP) (Mironov, 2017). A randomized mechanism $f : \mathcal{G} \mapsto \mathcal{R}$ is said to have ε-RDP of order α, abbreviated as (α, ε)-RDP, if for any adjacent $G, G' \in \mathcal{G}$ it holds that Rényi-divergence $D_\alpha\left(f(G) \| f\left(G'\right)\right) \leq \varepsilon$.

Recent works have often adopted privacy amplification by subsampling in differentially private machine learning. Applying a randomized mechanism to a subsampled dataset always produces a lower bound on privacy loss, that is, privacy is am-

plified. RDP is a useful technique for analyzing how much the privacy loss is improved by the subsampling operation (Zhu and Wang, 2019)(Wang et al., 2019)(Mironov et al., 2019). Before introducing the subsampled RDP privacy amplification theorem, we first introduce Poisson subsampling.

Definition 3.3(Poisson subsampling). Given a dataset G, the procedure Poisson subsampling outputs a subset $\{g_i|\sigma_i = 1, i \in [n]\}$ of the original dataset G by sampling $\sigma_i \sim \mathrm{Ber}(\gamma)$ independently for $i = 1, 2, ..., n$.

Zhu and Wang (Zhu and Wang, 2019) give a tight bound to the privacy loss of the Poisson subsampling mechanism when noise is drawn independently from Gaussian or Laplace distribution.

Theorem 3.1(Privacy amplification theorem for subsampled RDP). Let M be a randomized algorithm that obeys $(\alpha, \varepsilon(\alpha))$-RDP whose randomness comes from Gaussian or Laplace noise. Let γ be the Poisson subsampling probability. $M \circ$ PoissonSubsample (G) denotes the composition function $M($ PoissonSubsample $(G))$ and $\varepsilon_{M\circ\text{PoissonSubsample}}(\alpha)$ is the privacy loss of $M \circ$ PoissonSubsample (G). Then,

$$\begin{aligned}&\varepsilon_{MoPoissonSubsample}(\alpha)\\&= \tfrac{1}{\alpha-1}\log\Big\{(1-\gamma)^{\alpha-1}(\alpha\gamma-\gamma+1)\\&+\textstyle\sum_{\ell=2}^{\alpha}\binom{\alpha}{\ell}(1-\gamma)^{\alpha-\ell}\gamma^{\ell}e^{(\ell-1)\varepsilon(\ell)}\Big\}.\end{aligned} \tag{3}$$

$\varepsilon_{MoPoissonSubsample}(\alpha)$ is simplified to $\varepsilon_{subsample}(\alpha)$ in the following sections.

4 Framework of SUB-LDA

In this section, we introduce our algorithm SUB-LDA, which achieves better privacy guarantee and efficiency than HDP-LDA. SUB-LDA is presented in Algorithm 1. Given document corpus G, SUB-LDA first preprocesses the corpus and randomly allocates topics to each word in the corpus. Then, SUB-LDA conducts CGS on a subset of words in each document produced by the Poisson subsampling process. When the convergence condition is satisfied or the amount of accumulative iterations reach maximum value $ITER$, then the burn-in period of CGS is stopped. Since SUB-CGS touches only a subset of sensitive words in each document during each iteration, privacy is amplified by Theorem 3.1. We present additional discussions about privacy and efficiency.

4.1 Privacy Amplification and Efficiency Improvement

Algorithm 1 SUB-LDA

Input: Document corpus G, Prior parameters τ,β, Subsampling ratio γ, Topic number K, Clipping bound $clip$
Output: Trained document-topic distribution Θ, topic-word distribution Φ, accumulate privacy loss $\varepsilon = T \cdot \varepsilon_{subsample}(\alpha)$
// Initialization
for $d_m \in G$ **do**
  for $w = t \in d_m$ **do**
    Sample topic: $k \sim Mult\left(\frac{1}{K} \cdot \mathbf{I}_K\right)$
    Initialize word count n_k^t and n_m^k
  end for
end for
// Collapsed Gibbs Sampling
Set $Iter = 0$
while not convergent or $Iter <= ITER$ **do**
  for $d_m \in G$ **do**
    Take a batch of word $\mathbb{W}_t$ from d_m according to subsampling ratio γ
    for $w = t \in \mathbb{W}_t$ **do**
      Add noise to each n_k^t independently: $n_k^t \leftarrow n_k^t + \eta$
      Clip: $\left(n_k^t\right)^{\text{temp}} \leftarrow \min\left\{n_k^t, clip\right\}$
      Compute sampling distribution $\mathbf{p}$: $p_k \propto \frac{\left(n_k^t\right)^{temp}+\beta}{\sum_{t=1}^{V}\left(n_k^t+\beta\right)} \cdot \frac{n_m^k+\tau}{\sum_{k=1}^{K}\left(n_m^k+\tau\right)}$
      Sample topic and update n_k^t and n_m^k via $\widetilde{\mathbf{p}}$
    end for
  end for
  $Iter \leftarrow Iter + 1$
end while
Output Trained document-topic distribution Θ, topic-word distribution Φ, accumulate privacy loss $\varepsilon = Iter \cdot \varepsilon_{subsample}(\alpha)$

Given privacy budget $\varepsilon(\alpha)$ of RDP in each iteration, noise η is drawn independently from Gaussian distribution $N\left(0, \sigma^2\right)$, where $\sigma^2 = \frac{\alpha}{2\varepsilon(\alpha)}$. Since SUB-LDA conducts Poisson subsampling before counting and noise injection, privacy budget $\varepsilon_{subsample}(\alpha)$ in each iteration is obtained by equation (3) and $\varepsilon_{subsample}(\alpha)$. Intuitively, the smaller the γ, the better the privacy amplification.

To discuss the efficiency improvement, we discover that the running time of each iteration is proportional to the sum of sampling times for each

Symbol	Meaning
τ, β	Hyper-parameters of Dirichlet distribution
G	Text corpus
n_m^k, n_k^t	Count of topic k in document m and count of word t in topic k
K	Topic amount
V	Amount of unique words in corpus
γ	Subsampling ratio
M	A randomized mechanism
$M \circ$ PoissonSubsample	A randomized mechanism equipped with Poisson Subsampling
$\varepsilon(\alpha)$	RDP privacy loss of order α of a randomized mechanism in one iteration
$\varepsilon_{subsample}(\alpha)$	RDP privacy loss of order α of a randomized mechanism equipped with Poisson Subsampling in one iteration
N_d	The length of document d
t	Running time of one CGS step for a single word
I	Efficiency-privacy function

Table 1: Notations for SUB-LDA

document d in each iteration. Let N_d be the length of document d and t be the time of conducting one CGS for a single word. Then, the total average time T of each iteration of CGS is

$$T = t \cdot \sum_{d=1}^{D} N_d \gamma \tag{4}$$

Obviously, smaller γ induces shorter total time; thus, CGS is more efficient. To evaluate privacy amplification and efficiency improvement synthetically, we propose the efficiency–privacy function I, which is defined as follows:

$$I = T \cdot e^{(\alpha-1)\varepsilon_{\text{subsample}}(\alpha)}. \tag{5}$$

For a given text corpus dataset, a smaller value of I indicates better efficiency and privacy amplification. In the following analysis, we omit the constant $t \cdot \sum_{d=1}^{D} N_d$, and I is simplified as the following kernel:

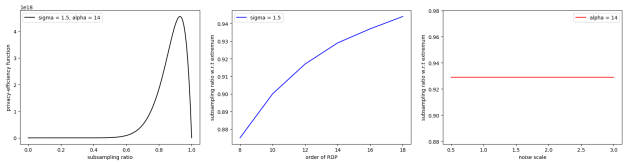
$$I = \gamma \cdot e^{(\alpha-1)\varepsilon_{\text{subsample}}(\alpha)}. \tag{6}$$

We find an important property of I, which is expressed in Lemma 4.1.

Lemma 4.1. There exists a $\gamma_0 \in (0, 1]$, where efficiency–privacy function I is monotonically increasing in $(0, \gamma_0]$ and monotonically decreasing in $(\gamma_0, 1]$.

Lemma 4.1 indicates that we could improve efficiency and amplify privacy simultaneously by decreasing the value of γ in a certain range of subsampling ratio γ. The proof of Lemma 4.1 is presented in the appendix.

We plot properties of efficiency–privacy function I in Figure 1. We observe an extremum of efficiency–privacy function I. Moreover, the value



(a) image of I (b) extremum of I w.r.t α (c) extremum of I w.r.t σ

Figure 1: Properties of Efficiency Function I

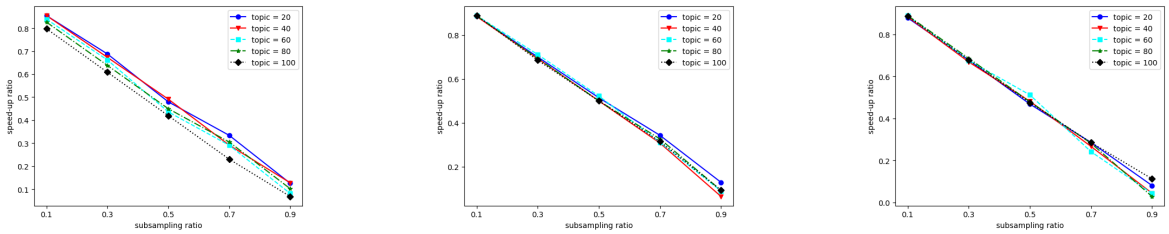
of the extremum increases when order α increases. However, the value of the extremum is unchanged when noise scale σ increases.

4.2 Subsampling actually amplifies privacy?

Does Poisson subsampling actually amplify the privacy of HDP-LDA? The answer could be yes or no. If we concentrate on one single iteration and fix the total iteration number $ITER$, the privacy budget actually shrinks, and we can conclude that Poisson subsampling amplifies privacy. If we concentrate on the whole training process of SUB-LDA, we cannot reach the exact same conclusion. Poisson subsampling actually amplifies privacy of each iteration of HDP-LDA. Nevertheless, the efficiency improvement is at the cost of more iterations to reach convergence (as the latent topics are updated for a subset of words in a document). Thus, the accumulated privacy loss $\varepsilon = Iter \cdot \varepsilon_{subsample}(\alpha)$ of SUB-LDA may increase. We show results in our experiments.

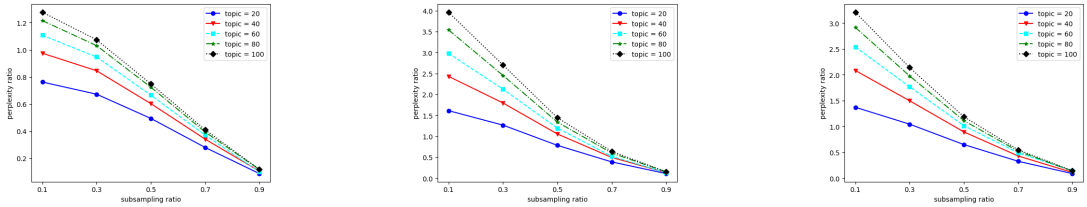
5 Experiment results

This section reports on our evaluation of SUB-LDA. We implement our method on three real-word datasets: 20 Newsgroups dataset (Lang,



(a) 20Newsgroups (b) NIPS (c) ENRON

Figure 2: Speed-up Ratio with Respect to Subsampling Ratio



(a) 20Newsgroups (b) NIPS (c) ENRON

Figure 3: Perplexity Ratio with Respect to Subsampling Ratio

1995), NIPS[1], and ENRON[2]. The statistics of these datasets are shown in Table 2.

Dataset	Amount of words	Amount of unique words	Amount of documents
20Newsgroups	908,262	138,203	9,740
NIPS	1,900,000	12,419	1,500
ENRON	6,400,000	28,102	37,861

Table 2: Statistics of Datasets

$K = 100$	SUB-LDA		
Subsampling ratio	$Iter$	$\varepsilon_{subsample}(\alpha)$	$\varepsilon = Iter \cdot \varepsilon_{subsample}(\alpha)$
0.1	92	0.05	4.6
0.3	83	0.73	59.86
0.5	70	1.18	82.6
0.7	66	1.48	97.86
0.9	61	1.63	99.43
1	42	2	84

Table 3: Privacy Guarantee Difference under Convergence (NIPS)

5.1 Efficiency improvement of SUB-LDA

In our experiments, we set the order α of RDP as 14 and the original privacy budget $\varepsilon(\alpha) = 2$. We vary the subsampling ratio γ from 0.1 to 0.9 with the step being 0.2. Obviously, when $\gamma = 1$, SUB-LDA is simply HDP-LDA. The topic amount varies from 20 to 100 with the step being 20. We omit the convergence condition and use $ITER = 100$ to stop the iterations. We record the average running time t_{sub} of each SUB-LDA iteration and the average running time t_{hdp} of each HDP-LDA iteration. The speed-up ratio is calculated as follows:

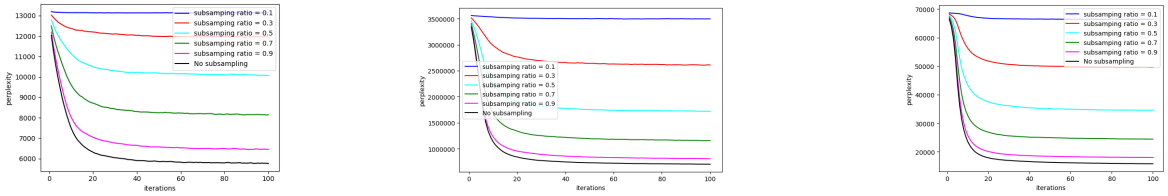
$$\text{Speed-up ratio} = \frac{|t_{sub} - t_{hdp}|}{t_{hdp}}. \quad (7)$$

The results are shown in Figure 2. We conclude that SUB-LDA would have better efficiency if we choose a smaller subsampling ratio. The amount of topic K indicates the complexity of the LDA model. Larger K often results in better efficiency improvement, which indicates that SUB-LDA could be suitable for a complex LDA model.

5.2 Effectiveness difference between SUB-LDA and HDP-LDA

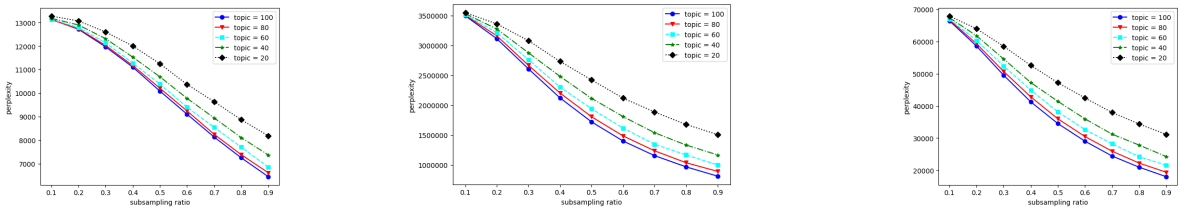
We choose perplexity to evaluate the model's utility. We focus on the impacts of Poisson subsampling of SUB-LDA on perplexity. After $ITRE = 100$ iterations, we record the perplexity per_{sub} of SUB-LDA and the perplexity per_{hdp} of HDP-LDA. We utilize the perplexity ratio in equation (8) to show the difference of effectiveness between SUB-LDA and HDP-LDA. Lower perplexity always indicates better generalization ability of the LDA model. Subsampling has non-negligible impacts on the model's perplexity. Figure 3 shows that perplexity would have a smaller difference when the value of the subsampling ratio is larger. Furthermore,

[1]https://archive.ics.uci.edu/ml/datasets/bag+of+words
[2]https://archive.ics.uci.edu/ml/datasets/bag+of+words



(a) 20Newsgroups (b) NIPS (c) ENRON

Figure 4: Convergence of SUB-LDA



(a) 20Newsgroups (b) NIPS (c) ENRON

Figure 5: Changes of Perplexity

we conclude that the effectiveness difference between SUB-LDA and HDP-LDA may be related to the complexity of the LDA model. More complex models often have bigger differences of perplexity.

$$\text{Perplexity ratio} = \frac{|per_{sub} - per_{hdp}|}{per_{hdp}}. \quad (8)$$

5.3 Convergence difference between SUB-LDA and HDP-LDA

In this subsection, we fix the topic amount as 100 and track the changes of perplexity during iterations. The results are shown in Figure 4. A smaller subsampling ratio results in higher perplexity. Moreover, the convergence of SUB-LDA is influenced by subsampling. Often SUB-LDA with a larger subsampling ratio has a faster convergence rate.

5.4 Privacy guarantee difference between SUB-LDA and HDP-LDA under convergence

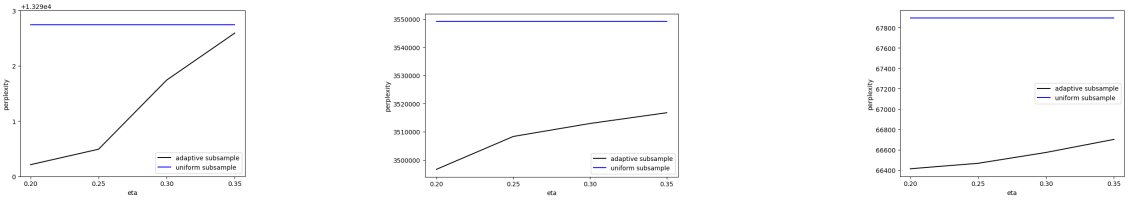
In this subsection, the topic amount is fixed as 100. We utilize the convergence condition to stop the burn-in process. Given i-th iteration, per_i denotes the perplexity of this iteration. The difference value of perplexity of the i-iteration is defined as $D_i = |per_i - per_{i-1}|$. Given a threshold $\hat{D}$, we consider that SUB-LDA reaches convergence if the following condition is satisfied for some value T and s:

$$D_i \leq \hat{D}, i = T, T+1, \ldots, T+s. \quad (9)$$

We use theorem 6 in (Zhu and Wang, 2019) to approximate $\varepsilon_{subsample}(\alpha)$ in each iteration. We then calculate the accumulative privacy budget of SUB-LDA, and the results of NIPS are shown in Table 3. Unsurprisingly, $\varepsilon_{subsample}(\alpha)$ shrinks when the subsampling ratio decreases, but we obtain a larger value of $Iter$. From Table 3, we observe that the accumulative privacy losses of SUB-LDA with a subsampling ratio of 0.9 and 0.7 are greater than those with subsampling ratio of 1. The results of Table 3 is to provide insights on the synthetical impacts of iterations and subsampling ratio towards privacy guarantee. Thus, in practice, we can find a suitable subsampling ratio not the smallest ratio to provide a rigorous privacy guarantee.

5.5 Relationship between efficiency–privacy function and perplexity

Obviously, it is difficult to analyze properties of perplexity. Nevertheless, we discover that efficiency–privacy function I tends to have similarities to perplexity. We show the values of perplexity with respect to each dataset in Figure 5 after SUB-LDA terminates. In Figure 1, we discover that the gradient of I first increases and then decreases. The



(a) 20Newsgroups (b) NIPS (c) ENRON

Figure 6: Perplexity Produced by AS Method

change rate (absolute value) of perplexity displayed in Figure 5 also tends to increase first and then decrease(these curves are concave first and then convex). This is reasonable in practice, since the improvements of efficiency and privacy are usually at the cost of variation of the model's utility. This indicates that analysis of perplexity could be substituted for analysis of the efficiency–privacy function.

5.6 An effective method to improve model's utility in practice

In our subsampling experiments, we discover that the model's perplexity produced by SUB-LDA usually tends to have few changes when the subsampling ratio is small (e.g., $\gamma = 0.1$). To improve utility in this case, we propose an AS method. The AS method is based on the fact that a small subset of frequent words has much higher probabilities than the other words given a certain topic k. Thus, we can increase the subsampling ratio of frequent words of topic k while decreasing the subsampling ratio of infrequent words. Give corpus dataset G, we denote $G = \{N_1, \dots, N_k, \dots N_K\}_{k=1}^{K}$ as the partition of G in terms of topics in the i-th iteration. For each $N_k = \sum_{t=1}^{V} n_k^t$, the AS method first constructs a frequent word subset, namely, $\left\{t \in \{1, 2, ..., V\} : \sum_t n_k^t \geq qN_k\right\}$, where q is fixed beforehand. For n-th word $w_{m,n} = t$ with topic $z_{m,n} = k$ in document d_m, AS sets the subsampling ratio as follows:

$$\gamma_t^i\big|_{w_{m,n}=t, z_{m,n}=k} = \begin{cases} v\gamma\left(\frac{1}{\gamma} > v > 1\right), t \in \left\{t : \sum_t n_k^t \geq qN_k\right\} \\ 0, t \notin \left\{t : \sum_t n_k^t \geq qN_k\right\} \end{cases} \tag{10}$$

Denote $\left|G^{sub}\right|$ and $\left|\bar{G}^{sub}\right|$ as the size of the word subset produced by the conventional subsampling method (we call this a uniform subsample) and the AS method. Then, we have

$$\frac{E\left(\left|G^{sub}\right|\right)}{E\left(\left|\bar{G}^{sub}\right|\right)} = \frac{1}{vq} = \eta. \tag{11}$$

$$\frac{\mathrm{Var}\left(\left|G^{sub}\right|\right)}{\mathrm{Var}\left(\left|\bar{G}^{sub}\right|\right)} = \frac{1-\gamma}{qv(1-v\gamma)}. \tag{12}$$

The AS method takes η as input. We apply the AS method to each dataset under topic $k = 20$ and $\gamma = 0.1$. The results are shown in Figure 6. We observe that prominent improvements of utility are achieved for NIPS and ENRON. For 20Newsgroups, we achieve similar utility, since the scale of 20Newsgroups is small compared to NIPS and ENRON.

Privacy guarantees for the uniform subsample and the adaptive subsample are provided in Lemma 5.1.

Lemma 5.1. Suppose a randomized mechanism M satisfies (ε, δ)-DP. M equipped with uniform subsampling satisfies (ε', δ')-DP. M equipped with adaptive subsampling satisfies $(\bar{\varepsilon}, \bar{\delta})$-DP. Then,

$$\bar{\varepsilon} \leq \varepsilon' - log(\eta q), \bar{\delta} \leq \frac{\gamma}{\eta q}\delta. \tag{13}$$

6 Conclusion and Future Works

In this study, we combine Poisson subsampling with HDP-LDA to improve efficiency and amplify privacy in LDA model training. We find that subsampling naturally improves efficiency. Moreover, we propose a metric to evaluate the efficiency–privacy improvement via efficiency–privacy function I. We discuss the properties of I. We then conduct comprehensive experiments to evaluate the efficiency improvements and privacy amplification effects. In future works, we plan to combine SUB-LDA with distributed CGS algorithms that satisfy local DP to boost the efficiency and privacy guarantee.

Acknowledgements

Hong Chen is the corresponding author. This work is supported by National Natural Science Foundation of China (62072460, 61772537, 61772536, 62076245, 62172424), Beijing Natural Science Foundation (4212022).

References

Borja Balle, Gilles Barthe, and Marco Gaboardi. 2020. Privacy profiles and amplification by subsampling. *Journal of Privacy and Confidentiality*, 10(1).

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Chain Monte Carlo. 2004. Markov chain monte carlo and gibbs sampling. *Lecture notes for EEB*, 581.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Chuan Hu, Huiping Cao, and Qixu Gong. 2017. Subgibbs sampling: a new strategy for inferring lda. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 907–912. IEEE.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier.

Aaron Q Li, Amr Ahmed, Sujith Ravi, and Alexander J Smola. 2014. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 891–900.

Jun S Liu. 2008. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media.

David JC MacKay and David JC Mac Kay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275. IEEE.

Ilya Mironov, Kunal Talwar, and Li Zhang. 2019. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*.

Mijung Park, James Foulds, Kamalika Chaudhuri, and Max Welling. 2016. Variational bayes in private settings (vips). *arXiv preprint arXiv:1611.00340*.

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577.

Alastair J Walker. 1977. An efficient method for generating discrete random variables with general distributions. *ACM Transactions on Mathematical Software (TOMS)*, 3(3):253–256.

Yansheng Wang, Yongxin Tong, and Dingyuan Shi. 2020. Federated latent dirichlet allocation: A local differential privacy based framework. In *AAAI*, pages 6283–6290.

Yi Wang, Xuemin Zhao, Zhenlong Sun, Hao Yan, Lifeng Wang, Zhihui Jin, Liubin Wang, Yang Gao, Jia Zeng, Qiang Yang, et al. 2014. Towards topic modeling for big data. *arXiv preprint arXiv:1405.4402*.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. 2019. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235. PMLR.

Han Xiao and Thomas Stibor. 2010. Efficient collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of 2nd Asian Conference on Machine Learning*, pages 63–78.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 937–946.

Jinhui Yuan, Fei Gao, Qirong Ho, Wei Dai, Jinliang Wei, Xun Zheng, Eric Po Xing, Tie-Yan Liu, and Wei-Ying Ma. 2015. Lightlda: Big topic models on modest computer clusters. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1351–1361.

Lele Yut, Ce Zhang, Yingxia Shao, and Bin Cui. 2017. Lda* a robust and large-scale topic modeling system. *Proceedings of the VLDB Endowment*, 10(11):1406–1417.

Fangyuan Zhao, Xuebin Ren, Shusen Yang, Qing Han, Peng Zhao, and Xinyu Yang. 2020. Latent dirichlet allocation model training with differential privacy. *IEEE Transactions on Information Forensics and Security*.

Fangyuan Zhao, Xuebin Ren, Shusen Yang, and Xinyu Yang. 2019. On privacy protection of latent dirichlet allocation model training. *arXiv preprint arXiv:1906.01178*.

Tianqing Zhu, Gang Li, Wanlei Zhou, Ping Xiong, and Cao Yuan. 2016. Privacy-preserving topic model for tagging recommender systems. *Knowledge and information systems*, 46(1):33–58.

Yuqing Zhu and Yu-Xiang Wang. 2019. Poission subsampled rényi differential privacy. In *International Conference on Machine Learning*, pages 7634–7642.

A Appendix

Proof of Lemma 4.1. Let D and D' be two neighboring datasets where $D' = D \cup \{d_{n+1}\}$ is satisfied. Denote $\mu_1 = M(D)$ and $\mu_2 = M(D')$. Given subsampling ratio γ and supposing that $p = M\circ$ PoissonSubsample (D) and $q = M\circ$ PoissonSubsample (D'), we have $e^{(\alpha-1)\cdot\varepsilon_{subsample}(\alpha)} = \mathbb{E}_q\left[(p/q)^\alpha\right]$. Meanwhile,

$$\mathbb{E}_q\left[(p/q)^\alpha\right] = \mathbb{E}_{\mu_0}\left[((1-\gamma)+\gamma\mu_1/\mu_0)^\alpha\right]. \tag{14}$$

Then,

$$I = \gamma\cdot\mathbb{E}_{\mu_0}\left[((1-\gamma)+\gamma\mu_1/\mu_0)^\alpha\right]. \tag{15}$$

We obtain the first derivative of I.

$$\begin{aligned}
&\frac{\partial I}{\partial\gamma} = \mathbb{E}_{\mu_0}\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^\alpha \\
&+\alpha\gamma\mathbb{E}_{\mu_0}\left[\left(\frac{\mu_1}{\mu_0}-1\right)\left(1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right)^{\alpha-1}\right] \\
&= \mathbb{E}_{\mu_0}\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^\alpha + \alpha\mathbb{E}_{\mu_0}\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^\alpha \\
&-\alpha\mathbb{E}_{\mu_0}\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^{\alpha-1}
\end{aligned} \tag{16}$$

Suppose that γ_0 satisfy $\frac{\partial I}{\partial\gamma_0} = 0$. To analyze $\frac{\partial^2 I}{\partial\gamma_0^2}$, let

$$g_\alpha(\gamma) = \mathbb{E}_{\mu_0}\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^\alpha. \tag{17}$$

Then,

$$(1+\alpha)g_\alpha(\gamma_0) = \alpha g_{\alpha-1}(\gamma_0). \tag{18}$$

We need the following lemma to decide the sign of $\frac{\partial^2 I}{\partial\gamma_0^2}$.

Lemma. For all integers $\alpha > 1$, $g_\alpha(\gamma) \geq [g_{\alpha-1}(\gamma)]^{\frac{\alpha}{\alpha-1}}$.

Proof: Due to the convexity of $f(x) = x^{\frac{\alpha}{\alpha-1}} (\alpha > 1)$, we have

$$\begin{aligned}
&g_\alpha(\gamma) = \mathbb{E}_{\mu_0}\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^\alpha \\
&= \mathbb{E}_{\mu_0}\left(\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^{\alpha-1}\right)^{\frac{\alpha}{\alpha-1}} \\
&\geq \left(\mathbb{E}_{\mu_0}\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^{\alpha-1}\right)^{\frac{\alpha}{\alpha-1}} = [g_{\alpha-1}(\gamma)]^{\frac{\alpha}{\alpha-1}}
\end{aligned}$$

In particular, $g_\alpha(\gamma) \geq [g_1(\gamma)]^\alpha$. Moreover, $g_1(\gamma) = \mathbb{E}_{\mu_0}\left[1-\gamma+\gamma\frac{\mu_1}{\mu_0}\right]^1 = 1$. Thus, $g_\alpha(\gamma) \geq [g_1(\gamma)]^\alpha = 1$.

We now decide the sign of $\frac{\partial^2 I}{\partial\gamma_0^2}$.

$$\begin{aligned}
&\frac{\partial^2 I}{\partial\gamma_0^2} \\
&= \frac{\alpha(\alpha-1)}{\gamma_0}\left[g_\alpha(\gamma_0) - 2g_{\alpha-1}(\gamma_0) + g_{\alpha-2}(\gamma_0)\right] - \frac{2}{\gamma}g_\alpha(\gamma_0) \\
&\leq \frac{\alpha(\alpha-1)}{\gamma_0}\left[g_\alpha(\gamma_0) - 2g_{\alpha-1}(\gamma_0) + [g_{\alpha-1}(\gamma_0)]^{\frac{\alpha-2}{\alpha-1}}\right] - \frac{2}{\gamma}g_\alpha \\
&= \frac{\alpha(\alpha-1)}{\gamma_0}\left[g_\alpha(\gamma_0) - (2-c)g_{\alpha-1}(\gamma_0)\right] - \frac{2}{\gamma}g_\alpha(\gamma_0)
\end{aligned}$$

where $c = [g_{\alpha-1}(\gamma)]^{\frac{-1}{\alpha-1}}$. We have $c \geq 1$. Together with equation (14), we have

$\frac{\partial^2 I}{\partial\gamma_0^2} = \frac{g_\alpha(\gamma_0)}{\gamma_0}\left[(1-c)\alpha^2 - \alpha + c - 2\right]$.

Let $h(\alpha) = (1-c)\alpha^2 - \alpha + c - 2$. When $c \geq 1$, $h(\alpha) \leq h(2) = -3c < 0$. Thus, $\frac{\partial^2 I}{\partial\gamma_0^2} < 0$. This proves Lemma 4.1.

Proof of Lemma 5.1. According to Theorem 13 in (Balle et al., 2020), we have

$$\varepsilon' = \log\left(1+\gamma\left(e^\varepsilon-1\right)\right), \delta' \leq \gamma\delta. \tag{19}$$

$$\bar{\varepsilon} = \log\left(1+v\gamma\left(e^\varepsilon-1\right)\right), \bar{\delta} \leq v\gamma\delta. \tag{20}$$

For ε' and $\bar{\varepsilon}$, we have
$\varepsilon' - \bar{\varepsilon} = \log\frac{1+\gamma(e^\varepsilon-1)}{1+v\gamma(e^\varepsilon-1)} \geq \log\frac{1}{v} = \log q\eta$.
Thus,

$$\bar{\varepsilon} \leq \varepsilon' - log(\eta q), \bar{\delta} \leq \frac{\gamma}{\eta q}\delta. \tag{21}$$

.