An end-to-end Differentially Private Latent Dirichlet Allocation Using a Spectral Algorithm

Christopher DeCarolis [1] **Mukul Ram** [1] **Seyed Esmaeili** [1] **Yu-Xiang Wang** [2] **Furong Huang** [1]

Abstract

We provide an end-to-end differentially private spectral algorithm for learning LDA, based on matrix/tensor decompositions, and establish theoretical guarantees on utility/consistency of the estimated model parameters. We represent the spectral algorithm as a computational graph. Noise can be injected along the edges of this graph to obtain differential privacy. We identify *subsets of edges*, named "configurations", such that adding noise to all edges in such a subset guarantees differential privacy of the end-to-end spectral algorithm. We characterize the sensitivity of the edges with respect to the input and thus estimate the amount of noise to be added to each edge for any required privacy level. We then characterize the utility loss for each configuration as a function of injected noise. Overall, by combining the sensitivity and utility characterization, we obtain an end-to-end differentially private spectral algorithm for LDA and identify which configurations outperform others under specific regimes. We are the first to achieve utility guarantees under a required level of differential privacy for learning in LDA. We additionally show that our method systematically outperforms differentially private variational inference.

1. Introduction

Topic modeling has been used extensively in document categorization, social sciences, machine translation and so forth. Learning topic modeling involves projecting high dimensional observations (documents) to a lower dimensional latent structure (topics), and outputting a model parameter estimation that describes the generative process of observed documents. This paper focuses on the popular topic model — *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003). There exist multiple learning algorithms for LDA, but the output of these algorithms may leak sensitive information in domains where privacy is a concern. This can limit the applicability of LDA in legal, financial, and medical domains. For instance, consider a situation in which the corpus D contains medical records, an adversary could potentially trace a learned topic t of an LDA learning algorithm back to an individual document d. This is a realistic threat model because topic t is high-dimensional and may contain a unique combination of words that only appear in d. We refer readers to (Carlini et al., 2019) for a concrete example of a learned high-dimensional machine learning model leaking credit card and social security numbers. Differential privacy (DP) (Dwork et al., 2006) is a formal definition of privacy that provides *provable and quantifiable* protection against such re-identification attacks. A generic method to convert an algorithm A to be differentially private is to add *sufficient noise* to A's output.

The existing state-of-the-art differentially private algorithm for learning LDA is differentially private variational inference (DP VI) (Park et al., 2016; 2020), in which noise is added at each iteration of variational inference to guarantee privacy. However, VI (Blei et al., 2003)-based LDA — even without privacy considerations — is not guaranteed to consistently learn LDA in *polynomial time*[1]. After all, it aims at solving a non-convex optimization problem that maximizes the likelihood function with (a variational approximation) of expectation-maximization.

The spectral learning method for LDA (Anandkumar et al., 2014a), on the other hand, circumvents the nonconvex optimization problem by solving a moment-matching equation using tensor decomposition, thereby enjoying provable computational efficiency and statistical consistency.

The **goals** of our work are twofolds. **(1)** to introduce a fam-

[*]Equal contribution [1]Department of Computer Science, University of Maryland [2]Department of Computer Science, UC Santa Barbara. Correspondence to: Furong Huang <furongh@cs.umd.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]Note that VI is shown to be statistically consistent (Wang & Blei, 2018) if the *optimal variational posterior* can be found, but it requires a potentially unbounded number of iterations. The DP extension has a total privacy loss that composes over the many iterations, therefore cannot afford to run many iterations.

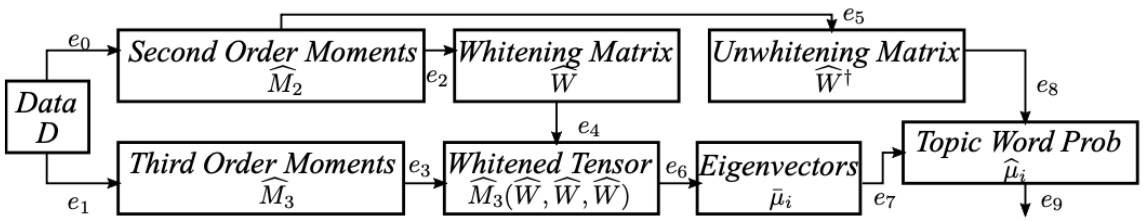


Figure 1. Algorithmic flow of end-to-end spectral learning algorithm to learning LDA topic model.

ily differentially private extensions to spectral-LDA that are guaranteed to be achieve a prescribed budget of differential privacy for all possible input datasets; **(2)** to show that they are able to provably recover high quality estimates of the LDA model parameters and to compare the privacy-utility tradeoff of these methods using theory and experiments.

Figure 1 illustrates a computation graph of the spectral LDA algorithm. Each edge represents a potential place where noise could be added. We define *configurations* as subsets of edges E of edges $\{e_i\}_{i=0}^{9}$. When E is a *cut* that separates the input and the output, differentially privately releasing (e.g., adding noise to) all nodes preceding the edges in E guarantees the overall differential privacy according to the composition theorem and the closure to post processing. For instance, privately releasing nodes preceding $E = (e_0, e_2)$ provides no privacy as the non-private information could flow to the output through the path below. However, when $E = \{e_5, e_6\}$, then such information-flow is cut off which guarantees overall differential privacy.

Summary of results. Our main contributions are:

- **(1)** We provide bounds for the sensitivities of intermediate quantities on the computation graph and identify four *configurations* of interest. For each configuration, we propose methods that achieve either pure-ϵ-DP or approximate (ϵ, δ)-DP for all choices of $\epsilon, \delta > 0$. Whenever applicable, we design data-dependent DP mechanisms that exploit a small local sensitivity and provide differential privacy even when the global sensitivity is large or unbounded.
- **(2)** We analyze the impact of the noise-injected by our algorithms and establish high-probability error bounds for estimating true model parameters. In some configurations, we show that the impact of differential privacy is in a low-order term, which says that for a large dataset, the utility cost of ensuring differential privacy is *almost for free*.
- **(3)** We conduct empirical studies with synthetic and real-life datasets, which confirm that the DP spectral algorithm systematically outperforms DP variational inference.

Compared to differentially private VI, the proposed approach is advantageous in that it (1) retains consistency guarantees, (2) is computationally efficient, (3) achieves higher accuracy in synthetic and real data experiments, moreover, (4) does not require performing composition across multiple iterations. We note that empirically VI is known to be more data-efficient than spectral learning methods for topic modeling *when privacy is not a concern*. Interestingly, we observe that for almost all experiments, our proposed *differentially private* spectral learning algorithm outperforms its VI counterpart in all commonly accepted ranges of privacy budgets ($\epsilon \leq 1$, $\delta < 1/n$). This difference should be attributed to the simpler mathematical structures of spectral learning methods, which allows for more efficient use of a given privacy budget.

2. Related Work

There are a few works that are private extensions of variational inference (Schein et al., 2019; Park et al., 2020; 2017). Among these, Schein et al. (2019); Park et al. (2020) use topic models as examples, even though the model of (Schein et al., 2019) is a Poisson factorization model, rather than LDA. (Park et al., 2020) contains an updated set of experiments to (Park et al., 2016) on LDA which shows competitive perplexity scores.

Our work focuses on LDA parameter estimation based on spectral algorithms which, unlike EM-based algorithms (Park et al., 2017; 2016), guarantee parameter recovery if a mild set of assumptions are met (Anandkumar et al., 2012; 2014b). The spectral estimation method relies on matrix decomposition and tensor decomposition methods. Thus, differentially private PCA and tensor decomposition are related to our objective.

Differentially private PCA is an established topic, and $(\epsilon, 0)$ differentially private PCA was achieved using the exponential mechanism in (Chaudhuri et al., 2012; Kapralov & Talwar, 2013). The algorithm in (Kapralov & Talwar, 2013) provides guarantees but with complexity $O(d^6)$; in

contrast, (Chaudhuri et al., 2012) introduces an algorithm that is near optimal but without an analysis of convergence time. Although (ϵ, δ) differential privacy is a more loose definition of differential privacy, it leads to better utility. Comparative experimental results show that the (ϵ, δ) PCA algorithm of (Imtiaz & Sarwate, 2016) outperform $(\epsilon, 0)$ significantly, and (Dwork et al., 2014b) introduce a simple input perturbation algorithm which achieves near optimal utility. In our work, we follow the (ϵ, δ) definition and use (Dwork et al., 2014b) to obtain a differentially private matrix decomposition when needed.

Differentially private tensor decomposition is studied in (Wang & Anandkumar, 2016) with an incoherence basis assumption. It is not clear the extent to which such an assumption holds in topic modeling. The authors exclude the possibility of input perturbation as that causes the privacy parameter to be lower bounded by the dimension ($\epsilon = \Omega(d)$) which is prohibitive. However, the same analysis on the tensor of a reduced dimension would conclude that $\epsilon = \Omega(k)$, which is acceptable for a reduced dimension whitened tensor as $k \ll d$.

3. Preliminaries and Notations

Latent Dirichlet Allocation is characterized by two model parameters: $\boldsymbol{\alpha}$, the dirichlet parameter of the topic prior, and $\boldsymbol{\mu}$, the topic word matrix. $\boldsymbol{\alpha}$ parameterizes a dirichlet distribution, which determines the topic mixture in each document, $\boldsymbol{\mu}$ controls the word distribution per topic. We provide a detailed explanation of LDA in Appendix B. We use d to denote the number of distinct words in a vocabulary, N to denote the total number of documents, k to denote the number of topics. The topic prior Dirichlet distribution is parameterized by $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ and $\alpha_0 = \sum_{i=1}^{k} \alpha_i$. For each document n, topic proportion is θ_n, document length is l_n, and word frequency vector is denoted as c_n. Word tokens are denoted by x. Let D, D' be two datasets. We say datasets D and D' are adjacent (denoted by $D \sim D'$) if we can form D' by *replacing* exactly one document from D.

Definition 1 (**(ϵ, δ)-Differential Privacy**). *Let $\mathcal{A} : D \to Y$ be a randomized algorithm. If $\forall D \sim D', \forall S \subseteq Y$ $\mathbb{P}[\mathcal{A}(D) \in S] \leq e^{\epsilon}\mathbb{P}[\mathcal{A}(D') \in S] + \delta$, then $\mathcal{A}$ is (ϵ, δ)-DP (differentially private).*

Differential privacy provides any individual data point a degree of *plausible deniability* in the sense that attackers, even with arbitrary side-information, could not infer whether the individual is in the dataset or not.

Definition 2 (**Local / Global Sensitivity**). *The local sensitivity $\Delta_f(D) := \max_{D'|D' \sim D} \|f(D) - f(D')\|$ and the global sensitivity $\Delta_f := \max_D \Delta_f(D)$.*

The norm $\|\cdot\|$ could be any vector ℓ_p norm, and when the distinction matters, we say (local or global) ℓ_p sensitivity. Many differentially private algorithms, including those that we will build upon, are based on perturbing $f(D)$ with a noise. The level of the noise is calibrated using the sensitivity to ensure DP for some prescribed budgets ϵ, δ (see more details in Appendix A).

4. Differentially Private LDA Topic Model

The *method of moments* principle — dating back at least to (Pearson, 1894) — provides another class of algorithms for learning LDA by computation upon *data moments*. Notably, the method of moments algorithm based on spectral tensor decomposition (Anandkumar et al., 2012; 2014a) guarantees consistent recovery of the topic-word distribution (i.e. LDA model parameters) under the constraint that the *third order data moment tensor* can be uniquely decomposed (the *third order data moment* denotes the expected co-occurrence of triplets of words in a document).

To briefly describe the spectral algorithm of learning LDA, we define the first, second, and third order LDA moments in Lemma 3. Then, using the properties of LDA, we derive unbiased estimators of the LDA parameters by decomposing the LDA moments into factors that correspond to each μ_i, formalized in Lemma 3. We show that as long as we empirically estimate the moments M_1, M_2, and M_3 without bias, we obtain the model parameters α and μ via tensor decomposition on the empirically estimated moments.

Lemma 3 (**LDA moments and Moment Decompositions Recover Model Parameters**). *Let random variables x_1, x_2 and x_3 denote the first, second and third tokens in a document. Tokens are represented as one-hot encodings, i.e., $x_1 = e_v$ if the first token is the v-th word in the dictionary. We define the first, second, and third order moments of LDA M_1, M_2 and M_3 as $M_1 \overset{\text{def}}{=} \mathbb{E}[x_1]$, $M_2 \overset{\text{def}}{=} \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0+1}\mathbb{E}[x_1] \otimes \mathbb{E}[x_1]$ and $M_3 \overset{\text{def}}{=} \mathbb{E}[x_1 \otimes x_2 \otimes x_3] + \frac{2\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)}\mathbb{E}[x_1] \otimes \mathbb{E}[x_1] \otimes \mathbb{E}[x_1] - \frac{1}{\alpha_0+2}\Big(\mathbb{E}[x_1 \otimes x_2 \otimes \mathbb{E}[x_3]] + \mathbb{E}[x_1 \otimes \mathbb{E}[x_2] \otimes x_3] + \mathbb{E}[\mathbb{E}[x_1] \otimes x_2 \otimes x_3]\Big)$. The LDA moments relate to the model parameters α and μ through matrix/tensor decomposition as follows*

$$M_1 = \sum_i^k \frac{\alpha_i}{\alpha_0}\mu_i, \; M_2 = \sum_i^k \frac{\alpha_i}{\alpha_0(\alpha_0+1)}\mu_i \otimes \mu_i,$$
$$M_3 = \sum_i^k \frac{2\alpha_i}{\alpha_0(\alpha_0+1)(\alpha_0+2)}\mu_i \otimes \mu_i \otimes \mu_i. \quad (1)$$

The proof is given in Appendix E. Note that α_0 is prespecified and thus data-independent. Using the properties of LDA, the moments are decomposed as factors shown in Lemma 3, and the factors μ_i correspond to the LDA model

parameters we aim to estimate. According to Lemma 3, decomposing on matrix M_2 only will not result in correct recovery of μ_i as there are no unique μ_i's unless $\mu_i \perp\!\!\!\perp \mu_{i'}$ and $\alpha_i \neq \alpha_{i'}$. The word distributions under different topics are only linearly independent instead of orthogonal. However, tensor decomposition on M_3 will yield a unique decomposition (Anandkumar et al., 2014a).

Method of Moments & Tensor Decomposition Inspired by Lemma 3, we conclude that tensor decomposition on M_3 will result in consistent estimation of the LDA parameters α and μ_i. We have no access to population moments M_1, M_2 and M_3, but do have access to word frequency vectors c_n. To solve this problem, we empirically estimate the moments M_1, M_2, M_3 as in Equations (17)(18)(19) given the observations of word frequency vectors c_n, and obtain the model parameters α and μ by implementing tensor decomposition on those empirically estimated moments. In Lemma 26 in Appendix C, we prove that the empirical moment estimators are unbiased.

The method of moments uses the property of data moments of the LDA model (in Lemma 3) to estimate the parameters of topic model α and $\mu_i, \forall i \in k$. The algorithm flow is depicted in Figure 1 and consists of the following steps: **(1)** Using c_n for document $\forall n \in [N]$, estimate $\hat{M}_2$ and $\hat{M}_3$ using equation (18) (e_0 in Figure 1) and equation (19) (e_1 in Figure 1). **(2)** Apply SVD on $\hat{M}_2$ to obtain an estimation of the whitening matrix $\widehat{W} \overset{\text{def}}{=} \widehat{U}\widehat{\Sigma}^{-\frac{1}{2}}$, where $\widehat{U}$ and $\widehat{\Sigma}$ are the top k singular vectors and singular values of $\hat{M}_2$ (e_2 in Figure 1). **(3)** Whiten the tensor $\widehat{\mathcal{T}} = \hat{M}_3(\widehat{W}, \widehat{W}, \widehat{W})$ using multilinear operations [2] on $\hat{M}_3$ with $\widehat{W}$ (e_3 and e_4 in Figure 1). **(4)** Implement tensor decomposition on the whitened tensor $\widehat{\mathcal{T}}$ and denote the resulting eigenvectors as $\bar{\mu}_i, \forall i \in [k]$ (e_6 in Figure 1). **(5)** Obtain the un-whitening matrix $\widehat{W}^\dagger = \widehat{\Sigma}^{\frac{1}{2}}\widehat{U}^\top$ (e_5 in Figure 1). **(6)** Un-whiten the singular vectors to obtain LDA parameters: $\widehat{\mu}_i \propto (\widehat{W}^\dagger)^\top \bar{\mu}_i$ and $\widehat{\alpha}_i$, $\forall i \in k$ (e_7 and e_8 in Figure 1). **(7)** Project $\widehat{\mu}_i$ onto a simplex to get the final estimate. The spectral algorithm guarantees the correct learning of topic models (see Lemma 29).

Differentially Private LDA Problem Statement We assume that the corpus of data is held by a trusted curator and that an analyst will query for the parameters of the topic model. The curator has to output the model parameters α_i, μ_i in a differentially private manner with respect to the documents. While it is easy to achieve differential privacy, the challenge is in guaranteeing high utility. We will use the Gaussian mechanism described in Proposition 22 in this paper to achieve (ϵ, δ)-differentially private topic modeling for each of the configurations. We will compute sensitivities of edges in each configuration in Section 5 to obtain the noise level that must be added to each edge. Our derived utility loss results are demonstrated in Section 6.

[2] The (i, j, k)-th entry of the multilinear operation $\hat{M}_3(\hat{W}, \hat{W}, \hat{W})$ is $\sum_{m,n,l}[\hat{M}_3]_{m,n,l}\hat{W}_{m,i}\hat{W}_{n,j}\hat{W}_{l,k}$. $\hat{W}$ is $d \times k$ and $\hat{M}_3$ is $d \times d \times d$, thus $\hat{M}_3(\hat{W}, \hat{W}, \hat{W})$ is $k \times k \times k$.

5. Sensitivity of Nodes in Algorithmic Flow

The most straightforward method of making an algorithm differentially private is to add noise to the output. However, it is also possible to achieve differential privacy by adding noise earlier in the computation. As long as we privately release intermediate components (nodes) along a *cut* of the algorithm's computation graph (with bounded global sensitivity), differential privacy can be achieved via the composition theorem. For the spectral LDA algorithm, we list possible cuts on the computational graph as a *configuration*. Adding noise along different configurations can be helpful when trying to minimize utility loss for a fixed level of differential privacy, because the amount of noise required to reach a given privacy level differs based on where it is added. In fact, the amount of noise that needs to be injected is dependent upon the *sensitivity* of the nodes. Therefore, in order to determine the ideal regimes for each configuration, it is necessary to calculate the sensitivities of the various nodes defined on the computation graph. In this section, we calculate the sensitivities for the nodes used in each configuration, and in Section 6 we provide a utility analysis for each configuration.

Δ_2	global sensitivity of $\hat{M}_2$
Δ_3	global sensitivity of $\hat{M}_3$
$\Delta_{\widehat{\mathcal{T}}}(D)$	local sensitivity of $\widehat{\mathcal{T}}$
$\Delta_{\bar{\mu}}(D)$, $\Delta_{\bar{\alpha}}(D)$	local sensitivity of $\bar{\mu}_i, \bar{\alpha}_i$
$\Delta_{\mu}(D)$, $\Delta_{\alpha}(D)$	local sensitivity of μ_i, α_i
$\sigma_k(\hat{M}_2)$, $\sigma_k(\widehat{\mathcal{T}})$	k-th singular value of $\hat{M}_2$,$\widehat{\mathcal{T}}$
γ_s	$\frac{1}{4}\min_{i\in[k]}\sigma_i(\widehat{\mathcal{T}}) - \sigma_{i-1}(\widehat{\mathcal{T}})$
$\tau_{\epsilon,\delta}$	$\frac{2\ln 1.25/\delta}{\epsilon^2}$

Theorem 4 (Global sensitivity of second and third order LDA moments)**.** *Let Δ_2 and Δ_3 be the ℓ_1 sensitivities for $\hat{M}_2$ and $\hat{M}_3$ respectively. Both Δ_2 and Δ_3 are upper bounded by $O(\frac{1}{N})$.*

Theorem 5 (Local sensitivity of the whitened tensor $\widehat{\mathcal{T}}$)**.** *The ℓ_1 sensitivity of the whitened tensor $\widehat{\mathcal{T}}$, denoted as $\Delta_{\widehat{\mathcal{T}}}(D)$, is upper bounded by $\Delta_{\widehat{\mathcal{T}}}(D) = O(\frac{k^{1.5}}{N(\sigma_k(\hat{M}_2))^{1.5}})$.*

Theorem 6 (Local sensitivity of the output of tensor decomposition $\bar{\mu}_i, \bar{\alpha}_i$)**.** *Let $\bar{\mu}_1, \ldots, \bar{\mu}_k$ and $\bar{\alpha}_1, \ldots, \bar{\alpha}_k$ be the results of tensor decomposition before unwhitening. The sensitivity of $\bar{\mu}_i$, denoted as $\Delta_{\bar{\mu}}(D)$, and the sensitivity of $\bar{\alpha}_i$, denoted as $\Delta_{\bar{\alpha}}(D)$, are both upper bounded by $O(\frac{k^2}{\gamma_s N(\sigma_k(\hat{M}_2))^{1.5}})$, where $\gamma_s = \min_{i\in[k]}\frac{\sigma_i(\widehat{\mathcal{T}})-\sigma_{i+1}(\widehat{\mathcal{T}})}{4}$.*

Theorem 7 (Local sensitivity of the final output μ_i, α_i)**.** *The sensitivities $\Delta_\mu(D)$ and $\Delta_\alpha(D)$ of the final output are*

upper bounded by $O(\frac{k^2\sqrt{\sigma_1(\hat{M}_2)}}{\gamma_s N \sigma_k^{1.5}(\hat{M}_2)})$.

Remark. The sensitivities before the whitening are $O(\frac{1}{N})$. The whitening step increases the sensitivity by $\frac{k^{1.5}}{\sigma_k(\hat{M}_2)^{1.5}}$, leading to $O(\frac{k^{1.5}}{N(\sigma_k(\hat{M}_2))^{1.5}})$. Further, the simultaneous power method for tensor decomposition increases the sensitivity by $\frac{k^{0.5}}{\gamma_s}$, leading to $O(\frac{k^2}{\gamma_s N(\sigma_k(\hat{M}_2))^{1.5}})$. The unwhitening increases the sensitivity by $\sqrt{\sigma_1(\hat{M}_2)}$, leading to $O(\frac{k^2\sqrt{\sigma_1(\hat{M}_2)}}{\gamma_s N(\sigma_k(\hat{M}_2))^{1.5}})$. While we used big-O notation to present interpretable bounds, *explicit bounds* are required to implement our algorithm. A summary of these sensitivity is presented in the appendix.

5.1. Data-dependent Privacy Calibration

Theorem 5, 6 and 7 are local sensitivities, which are functions of the input data set. Adding noise proportional to the local sensitivity does not guarantee differential privacy as the local sensitivity may be sensitive to adding/removing of individuals and lead to the identification of individuals.

Two seminal solutions to this problem include the smooth sensitivity framework (Nissim et al., 2007) and the propose-test-release (PTR) framework (Dwork & Lei, 2009). The idea of the smooth sensitivity framework is to construct a smooth upper bound of the local sensitivity that is insensitive and to calibrate noise with a heavier tail that satisfies certain "dilation" and "shift" properties to achieve pure-DP. The PTR framework involves proposing bounds of the local sensitivity and testing its validity. If the test is passed, we calibrate the noise according to the proposed test. PTR is often easier to use but can only provide an (ϵ, δ)-DP with $\delta > 0$.

In our problem, the smooth sensitivity itself is unbounded, thus we cannot apply the smooth sensitivity framework naively. Instead, we use a variant of propose-test-release framework that releases a confidence bound of the local sensitivity in a differentially private manner, and calibrates noise accordingly, similar to the idea in (Blocki et al., 2012) and a more recent example in the context of data-adaptive differentially private linear regression (Wang, 2018). We formalize the idea using the following lemma.

Lemma 8. *Let $\Delta_f(D)$ be the local sensitivity of a function f on a fixed data set D. Let $\tilde{\Delta}_f(D)$ obeys $(\epsilon_1, 0)$-DP and that $\mathbb{P}[\Delta_f(D) \geq \tilde{\Delta}_f(D)] \leq \delta_1$ (where the probability is only over the randomness in releasing $\tilde{\Delta}_f(D)$). Then the algorithm releases $f(D) + Z(\epsilon, \delta, \tilde{\Delta}_f(D))$ that is $(\epsilon_1 + \epsilon, \delta_1 + \delta)$-DP, where $Z(\epsilon, \delta, \tilde{\Delta}_f(D))$ is any way of calibrating the noise for privacy (for Gaussian mechanism, one can take $Z(\epsilon, \delta, \tilde{\Delta}_f(D)) = \mathcal{N}(0, \frac{\tilde{\Delta}_f(D)^2}{2\epsilon^2}\big(\sqrt{\epsilon + \log(1/\delta)} + \sqrt{\log(1/\delta)}\big)^2)$).*

The proof is in Appendix G.6. In our problem, the local sensitivities depend on the data only through $\sigma_k(\hat{M}_2)$ and γ_s. A natural idea would be to privately release $\sigma_k(\hat{M}_2)$ and γ_s and construct a high-confidence upper bound of the local sensitivity through a high-confidence lower bound of $\sigma_k(\hat{M}_2)$ and γ_s. We will show the global sensitivities of $\sigma_k(\hat{M}_2)$ and $\sigma_i(\widehat{\mathcal{T}})$ are small, and release $\sigma_k(\hat{M}_2)$ and $\sigma_i(\widehat{\mathcal{T}})$ differentially privately.

Lemma 9 (Global Sensitivity of $\sigma_k(\hat{M}_2)$ and γ_s)**.** *The sensitivities of $\sigma_k(\hat{M}_2)$ and γ_s are each $2/N$.*

The proof is in Appendix G.7.

Calibrating Noise Using Lemma 8 and Lemma 9, we describe an algorithm that guarantees $(\epsilon_1+\epsilon_1'+\epsilon, \delta_1+\delta_1'+\delta_2)$-DP under local sensitivity $\tilde{\Delta}_f(D)$ in Procedure 1.

6. Differentially Private Spectral Algorithm

In Figure 1, each node corresponds to an intermediate objective required for a final output estimation and each edge denotes certain operation required as a step of the spectral learning algorithm. We consider injecting noise to a subset E of edges $\{e_i\}_{i=0}^9$ that separates the input and the output (a cut). When E is a cut, differentially privately releasing all nodes preceding the edges in E under bounded global sensitivity guarantees the overall differential privacy according to the composition theorem and the closure to post processing. We call such a subset of edges as a "configuration" if adding noise to all edges in this configuration guarantees differential privacy of the overall algorithm.

In this section, We achieve $(\epsilon_1 + \epsilon_1' + \epsilon, \delta_1 + \delta_1' + \delta_2)$-DP under local sensitivity $\tilde{\Delta}_f(D)$ in Procedure 1 Four configurations are identified as in Table 1. $\tilde{\sigma}_k$ and $\tilde{\gamma_s}$ are determined by a choice of (ϵ_1, δ_1) and (ϵ_1', δ_1'). In what follows, if noise is added to edge e_i, then ϵ_i refers to the associated differential privacy parameter.

Config. 1 has a global ℓ_1 sensitivity $O(1/N)$ and we could obtain pure-DP if we add Laplace noise instead.

In Config. 2, the whitening matrix results from a noiseless $\hat{M}_2$, but the pseudo-inverse results from a noisy $\hat{M}_2$. We add noise to a tensor of a smaller dimension, at the expense of an increased sensitivity by a factor of $\frac{k^{3/2}}{\sigma_k^{3/2}(\hat{M}_2)}$.

Config. 3 adds noise to the output of the simultaneous tensor power method and thus the sensitivity after the output of the simultaneous power iteration increases by a factor of $\frac{1}{\gamma_s}$ compared to Config. 2.

Config. 4 is arguably the simplest, as the previous configurations involve the composition of multiple differentially private outputs whereas this method only adds noise to one branch. Adding noise to μ_i instead of $\bar{\mu}_i$ means that the noise vector increases in dimension from k to d.

Procedure 1 $(\epsilon_1+\epsilon_1'+\epsilon, \delta_1+\delta_1'+\delta)$-Differential Privacy (DP) Noise Calibration

Input: local sensitivity of the configuration: $\Delta_f(D)$, non-DP output of the configuration: $f(\mathrm{D})$
Output: $(\epsilon_1+\epsilon_1'+\epsilon, \delta_1+\delta_1'+\delta)$-DP output

$\widehat{\sigma}_k = \sigma_k(\hat{M}_2) + \mathsf{Lap}(\Delta_2/\epsilon_1)$ ▷ $(\epsilon_1, 0)$-DP release of $\sigma_k(\hat{M}_2)$ via Laplacian mechanism
$\tilde{\sigma}_k = \max\{0, \widehat{\sigma}_k - \frac{\Delta_2}{\epsilon_1}\log(\frac{1}{2\delta_1})\}$ ▷ high probability lower bound of $\widehat{\sigma}_k$: $\mathbb{P}(\tilde{\sigma}_k < \widehat{\sigma}_k) \geq 1-\delta_1$
if config # > 2 **then**
  $\widehat{\gamma}_s = \gamma_s + \mathsf{Lap}(\Delta_3/\epsilon_1')$ ▷ $(\epsilon_1', 0)$-DP release of γ_s via Laplacian mechanism
  $\tilde{\gamma}_s = \max\{0, \widehat{\gamma}_s - \frac{\Delta_3}{\epsilon_1'}\log(\frac{1}{2\delta_1'})\}$ ▷ high probability lower bound of $\widehat{\gamma}_s$: $\mathbb{P}(\tilde{\gamma}_s < \widehat{\gamma}_s) \geq 1-\delta_1'$
  Obtain $\tilde{\Delta}_f(D)$ — a high prob. upper bound of $\Delta_f(D)$ — by replacing $\sigma_k(\hat{M}_2)$ with $\tilde{\sigma}_k$ and γ_s with $\tilde{\gamma}_s$ in $\Delta_f(D)$
else
  Obtain $\tilde{\Delta}_f(D)$ by replacing $\sigma_k(\hat{M}_2)$ with $\tilde{\sigma}_k$ in $\Delta_f(D)$
  $\epsilon_1' = 0, \delta_1' = 0$
end if
Return $f(\mathrm{D}) + \mathcal{N}(0, \tilde{\Delta}_f(D)^2\tau_{\epsilon,\delta})$

Table 1. The four configurations identified for DP spectral method for LDA.

Configs	Edges	DP Mechanism
Config. 1	(e_2, e_3, e_5)	perturb $\hat{M}_2$ with $\mathcal{N}(0, \Delta_2^2\tau_{\epsilon_2,\delta_2})$ for (ϵ_2,δ_2)-DP W perturb $\hat{M}_3$ with $\mathcal{N}(0, \Delta_3^2\tau_{\epsilon_3,\delta_3})$ for (ϵ_3,δ_3)-DP $\widehat{M}_3$ perturb $\hat{M}_2$ with $\mathcal{N}(0, \Delta_2^2\tau_{\epsilon_5,\delta_5})$ for (ϵ_5,δ_5)-DP $\widehat{W}^\dagger$
Config. 2	(e_5, e_6)	perturb $\hat{M}_2$ with $\mathcal{N}(0, \Delta_2^2\tau_{\epsilon_5,\delta_5})$ for (ϵ_5,δ_5)-DP $\widehat{W}^\dagger$ perturbation $\widehat{\mathcal{T}}$ with $\mathcal{N}(0, \tilde{\Delta}_{\widehat{\mathcal{T}}}(D)^2\tau_{\epsilon_6,\delta_6})$ for $(\epsilon_1+\epsilon_6, \delta_1+\delta_6)$-DP $\widehat{\mathcal{T}}$
Config. 3	(e_5, e_7)	perturb $\hat{M}_2$ with $\mathcal{N}(0, \Delta_2^2\tau_{\epsilon_5,\delta_5})$ for (ϵ_5,δ_5)-DP $\widehat{W}^\dagger$ perturb $\bar{\mu}_i$ with $\mathcal{N}(0, \tilde{\Delta}_{\bar{\mu}_i}(D)^2\tau_{\epsilon_7,\delta_7})$ for $(\epsilon_1+\epsilon_1'+\epsilon_7, \delta_1+\delta_1'+\delta_7)$-DP $\bar{\mu}$
Config. 4	(e_9)	perturb $\widehat{\mu}_i$ with $\mathcal{N}(0, \tilde{\Delta}_{\mu_i}(D)^2\tau_{\epsilon_9,\delta_9})$ for $(\epsilon_1+\epsilon_1'+\epsilon_9, \delta_1+\delta_1'+\delta_9)$-DP $\widehat{\mu}$

Though it is possible to perform input perturbation, we exclude this option because this ℓ_2 sensitivity does not decay with the number of records. Therefore the utility of input perturbation is poor even with many records.

6.1. Utility Guarantees

For each configuration, we compute the noise needed to obtain (ϵ, δ) differential privacy based on sensitivity, thereby characterizing the utility with necessary noise. The utility of each configuration is listed in Theorems 10, 12, 14 and 16. Proofs of all utility derivations are in Appendix H.

From Lemma 29, we know the utility loss of the non-DP is upper bounded by $\|\mu_i - \hat{\mu}_i\|_2 \leq O(\frac{(\alpha_0+1)^2k^3}{p_{\min}^2\sigma_k(\mu)\sqrt{N}}) = \tilde{O}(\frac{k^3}{\sqrt{N}})$, where $p_{\min} = \min_i \frac{\alpha_i}{\alpha_0}$ and $\tilde{O}$ hides dependencies on quantities other than k, d, N and γ_s.

The utility losses consist of two $\tilde{O}$ terms – the first $\tilde{O}$ term is a bound of non-private learning and the second $\tilde{O}$ term bounds the different between the private estimator and the non-private estimator. Notice that the second $\tilde{O}$ term is negligible for large N when ϵ is a constant. Therefore, the impact of differential privacy is in a low-order term, which says that for a large dataset, the utility cost of ensuring differential privacy is *almost for free*.

Theorem 10 (Config. 1 Utility Loss)**.** *The utility loss* $\|\mu_i - \mu_i^{\mathsf{DP}}\|$ *using Config. 1 to guarantee* $(\epsilon_2+\epsilon_3+\epsilon_5, \delta_2+\delta_3+\delta_5)$*-DP is* $O(\frac{(\alpha_0+1)^2k^3}{p_{\min}^2\sigma_k(\mu)\sqrt{N}}) + O(\frac{\sqrt{\sigma_1(\hat{M}_2)k}}{\gamma_s}((\frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_2,\delta_2})^3 + \frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_3,\delta_3}) + \frac{\sqrt{\sigma_1(\hat{M}_2)d}}{\sigma_k(\hat{M}_2)N}\tau_{\epsilon_5,\delta_5} + \sqrt{\sigma_1(\hat{M}_2) + \frac{\sqrt{d}}{N}\tau_{\epsilon_5,\delta_5}}\frac{\sqrt{k}}{\gamma_s}\Big[(\frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)}\tau_{\epsilon_2,\delta_2})^3 + \frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_3,\delta_3}\Big])$.

Remark 11. *The order of the Config. 1 utility loss to guarantee* (ϵ, δ)*-DP is*

$$\tilde{O}\Big(\frac{k^3}{\sqrt{N}}\Big) + \tilde{O}\Big(\Big(\frac{k^{0.5}}{\gamma_s}\Big(\frac{\sqrt{d}}{N} + (\frac{\sqrt{d}}{N})^{1.5}\Big) + \frac{\sqrt{d}}{N}\Big)\frac{\log\frac{1}{\delta}}{\epsilon^2}\Big) \tag{2}$$

Theorem 12 (Config. 2 Utility Loss)**.** *The utility loss* $\|\mu_i - \mu_i^{DP}\|$ *using Config. 2 to guarantee* $(\epsilon_1+\epsilon_5+\epsilon_6,\ \delta_1+\delta_5+\delta_6)$*-DP is* $O(\frac{(\alpha_0+1)^2k^3}{p_{\min}^2\sigma_k(\mu)\sqrt{N}}) + O(\frac{\sqrt{\sigma_1(\hat{M}_2)k^{2.5}}}{\gamma_s N\tilde{\sigma_k}^{3/2}}\tau_{\epsilon_6,\delta_6} + \frac{\sqrt{\sigma_1(\hat{M}_2)d}}{\sigma_k(\hat{M}_2)N}\tau_{\epsilon_5,\delta_5} + \sqrt{\sigma_1(\hat{M}_2) + \frac{\sqrt{d}}{N}\tau_{\epsilon_5,\delta_5}}\frac{k^{2.5}\tau_{\epsilon_6,\delta_6}}{\gamma_s N\tilde{\sigma_k}^{3/2}})$.

Remark 13. *The order of the Config. 2 utility loss to guarantee* (ϵ,δ)*-DP is*

$$\tilde{O}\Big(\frac{k^3}{\sqrt{N}}\Big)+\tilde{O}\Big(\Big(\frac{k^{0.5}}{\gamma_s}\Big(\frac{k^{0.75}}{N}+\frac{k^2}{N}(\frac{\sqrt{d}}{N})^{0.5}\Big)+\frac{\sqrt{d}}{N}\Big)\frac{\log\frac{1}{\delta}}{\epsilon^2}\Big) \tag{3}$$

Theorem 14 (Config. 3 Utility Loss)**.** *The utility loss* $\|\mu_i - \mu_i^{\mathsf{DP}}\|$ *using Config. 3 to guarantee* $(\epsilon_1+\epsilon_1'+\epsilon_5+\epsilon_7,\ \delta_1+\delta_1'+\delta_5+\delta_7)$*-DP is* $O(\frac{(\alpha_0+1)^2k^3}{p_{\min}^2\sigma_k(\mu)\sqrt{N}}) + O(\frac{\sqrt{\sigma_1(\hat{M}_2)k^{2.5}}}{\tilde{\gamma_s}N\tilde{\sigma}_k^{3/2}}\tau_{\epsilon_7,\delta_7} + \frac{\sqrt{\sigma_1(\hat{M}_2)d}}{\sigma_k(\hat{M}_2)N}\tau_{\epsilon_5,\delta_5} + \sqrt{\sigma_1(\hat{M}_2) + \frac{\sqrt{d}}{N}\tau_{\epsilon_5,\delta_5}}\frac{k^2\tau_{\epsilon_7,\delta_7}}{\tilde{\gamma_s}N\tilde{\sigma}_k^{3/2}})$.

Remark 15. *The order of the Config. 3 utility loss to guarantee* (ϵ,δ)*-DP is*

$$\tilde{O}\Big(\frac{k^3}{\sqrt{N}}\Big)+\tilde{O}\Big(\Big(\frac{k^{0.5}}{\gamma_s}\Big(\frac{k^{0.75}}{N}+\frac{k^{1.5}}{N}(\frac{\sqrt{d}}{N})^{0.5}\Big)+\frac{\sqrt{d}}{N}\Big)\frac{\log\frac{1}{\delta}}{\epsilon^2}\Big) \tag{4}$$

Theorem 16 (Config. 4 Utility Loss)**.** *The utility loss* $\|\mu_i - \mu_i^{\mathsf{DP}}\|$ *using Config. 4 to guarantee* $(\epsilon_1+\epsilon_1'+\epsilon_9,\ \delta_1+\delta_1'+\delta_9)$ *is* $O(\frac{(\alpha_0+1)^2k^3}{p_{\min}^2\sigma_k(\mu)\sqrt{N}}) + O(\frac{\sqrt{\sigma_1(\hat{M}_2)dk^2}}{\tilde{\gamma_s}N\tilde{\sigma}_k^{3/2}}\tau_{\epsilon_9,\delta_9})$.

Remark 17. *The order of the Config. 4 utility loss to guarantee* (ϵ,δ)*-DP is*

$$\tilde{O}\Big(\frac{k^3}{\sqrt{N}}\Big) + \tilde{O}\Big(\frac{k^{0.5}}{\gamma_s}\frac{\sqrt{d}}{N}\frac{\log\frac{1}{\delta}}{\epsilon^2}\Big) \tag{5}$$

6.2. Comparison of Configurations

We present a pairwise comparison between the utilities of different configurations using $\tilde{O}$-order utility losses. The O-order utility losses are too complex for comparison in theory, but we will implement experiments for comparisons. As we illustrate in the remarks in the previous subsection 6.1, the utility loss difference are marked as blue.

Remark 18. ***Configuration 1 vs. 2:*** *When square root of the dimension (vocabulary size)* $\sqrt{d}$ *is smaller than total number of documents* N*, the dominating term in the blue is* $\tilde{O}(\frac{\sqrt{d}}{N})$ *for Config. 1 utility loss, and it is larger than the* $\tilde{O}(\frac{k^2}{N}(\frac{\sqrt{d}}{N})^{0.5})$ *term in Config. 2. Therefore, for smaller* d *Config. 2 is preferred over Config. 1.*

More importantly when d *is large, Config. 1 requires adding noise to the third order data moment* $\hat{M}_3$*, and thus explicitly forms the large third order data moment object* $\hat{M}_3$ *of size* $d\times d\times d$*. As a result, Config. 1 does not scale to large scale real-world experiments such as LDA on Wikipedia documents. In the experiments, for other configurations, we never explicitly form* $\hat{M}_3$*; the whitened third order moment* $\widehat{\mathcal{T}}$ *of size* $k\times k\times k$ *is formed instead.*

Remark 19. ***Configuration 2 vs. 3:*** *If we do not consider Procedure 1 of calibrating local sensitivity, the utility loss for Config. 3 seem to be lower than that of Config. 2 by a factor of* $\tilde{O}(k^{0.5})$ *in the last term of the utility loss differences colored blue. However, during the local sensitive calibration, Config. 3 requires extra differential private release of* γ_s*, which could cause the utility loss of Config. 3 to be larger than Config. 2. To understand how the two compares,* γ_s *is crucial and should be analyzed case by case.*

Remark 20. ***Configuration 3 vs. 4:*** *When* $\tilde{O}(k^{0.75}) \geq \tilde{O}(d^{0.5})$ *and* N *is sufficiently large, Config. 4 is preferred over Config. 3, and vice versa. Therefore, smaller* k *(relative to* d*) prefers Config. 3 and larger* k *(relative to* d*) prefers Config. 4.*

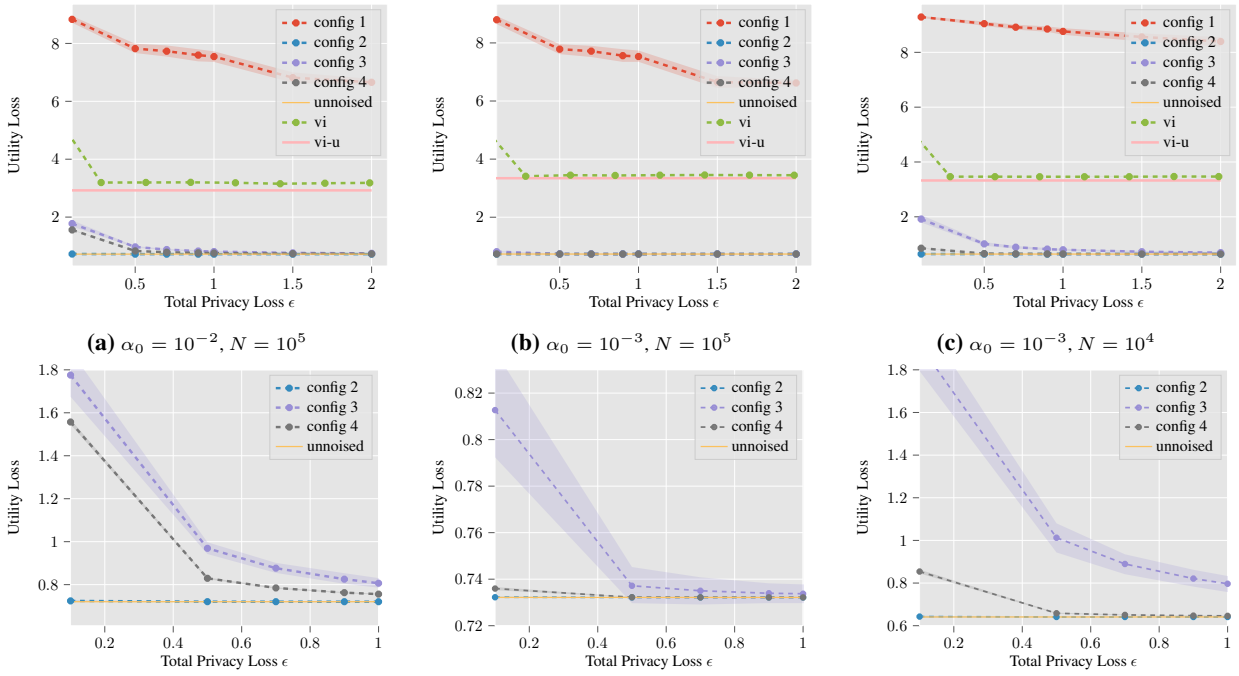
6.3. Comparison with DP VI

Without privacy constraints, variational inference estimates, although could be trapped in local optima, could sometimes achieve lower error than spectral methods in practice,. However, this can differ significantly in the differential privacy setting. Due to the fact that the DP VI algorithm requires adding noise across multiple iterations, compounded with the non-convexity of the likelihood function, empirical performance is often compromised. The guaranteed consistency of the spectral algorithm makes it a more attractive option in the differential privacy case.

7. Experiments

In a suite of synthetic experiments, we simulate documents from an LDA model parameterized by varying choice of α and μ. Each are randomly sampled to ensure that bursty use of a single word under a certain topic is possible in our experiment.Therefore, our setting covers a wide range of hyper-parameters and captures some common irregularities in distributional properties. Under this synthetic setting, we have access to the underlying parameters of the latent dirichlet allocation, and can thus directly calculate error with respect to the true parameters. This is not feasible with real data. We compare the empirical loss of each configuration under different hyperparameter settings. In addition, we compare all configurations of our spectral algorithm against differentially private variational inference (Park et al., 2016) run under the same settings. Our algorithm universally outperforms state-of-the-art VI quantitatively.

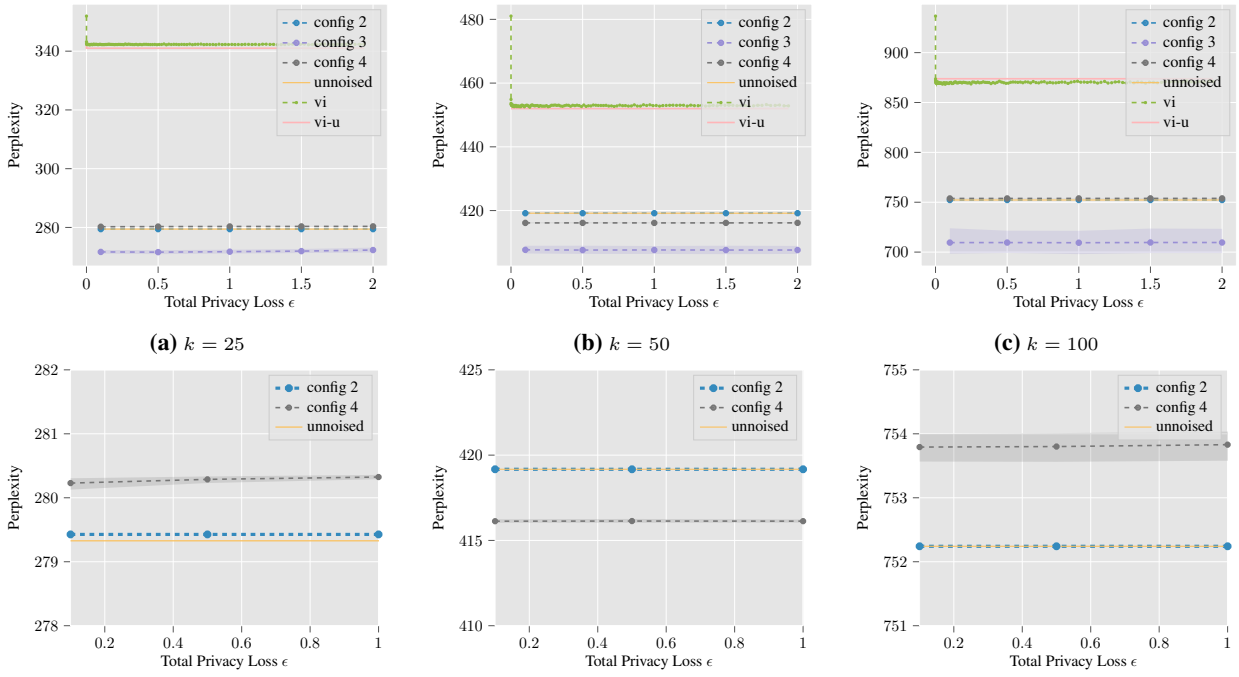
To evaluate Configuration 1, we set the vocabulary size and



(a) $\alpha_0 = 10^{-2}$, $N = 10^5$ **(b)** $\alpha_0 = 10^{-3}$, $N = 10^5$ **(c)** $\alpha_0 = 10^{-3}$, $N = 10^4$

(d) $\alpha_0 = 10^{-2}$, $N = 10^5$ (zoom in) **(e)** $\alpha_0 = 10^{-3}$, $N = 10^5$ (zoom in) **(f)** $\alpha_0 = 10^{-3}$, $N = 10^4$ (zoom in)

Figure 2. Error of **our method under all configurations** vs **the differentially private VI** over varying total privacy loss ϵ_{total} (in the range of 0.1 to 2) while fixing the $\delta = 10^{-5}$. vi-u and unnoised denote the non-DP version of VI and spectral algorithm respectively. $d = 50$, $k = 5$.



(a) $k = 25$ **(b)** $k = 50$ **(c)** $k = 100$

(d) $k = 25$ (zoom in) **(e)** $k = 50$ (zoom in) **(f)** $k = 100$ (zoom in)

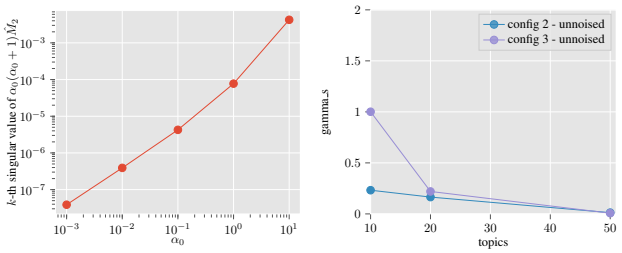
Figure 3. Perplexity scores of **our method under all configurations** vs **the differentially private VI** on Wikipedia data over varying total privacy loss ϵ_{total} while fixing the $\delta = 10^{-4}$. Number of words $d = 8000$, number of documents $N = 50000$, $\alpha_0 = 0.01$.

the number of topics to be small ($d = 50$, $k = 5$) in our synthetic settings. Configuration 1 requires calculating the unwhitened third order moment, which is computationally infeasible for large d or k.

Evaluation Metric: Our experiments evaluate the loss between the ground-truth μ and the estimated $\widehat{\mu}^{\mathsf{DP}}$ via a (ϵ, δ) differentially private algorithm across varying total privacy loss ϵ. The distribution of privacy budget across edges in each configuration is set to be uniform for simplicity. We release only differentially private likelihoods by additionally perturbing the sufficient statistics, as described in (Park et al., 2016).

VI vs Spectral: Figure 2 exhibits the error for varying total privacy loss ϵ on different datasets. Under all configurations except for configuration 1, our differentially private spectral algorithm outperforms differentially private variational inference, and has higher utility under the same level of privacy.

Config. 2 vs Config. 3: As described in Remark 19, the comparison between Config. 2 and Config 3 is unclear and should be analyzed case by case. In synthetic experiments with $d = 50$ and $k = 5$, Config. 2 outperforms Config. 3 as well as Config. 4. This is due to the noised $\tilde{\gamma}_s$ (in Procedure 1). We show the difference between noised $\tilde{\gamma}_s$ and unnoised γ_s in Figure 4b. Config. 2's gap between noised $\tilde{\gamma}_s$ and unnoised γ_s is always smaller than Config. 3's when $k < 50$, suggesting Config. 2 is preferred for smaller k. However, the difference between the gaps decreases as the number of topics increase, suggesting that Config. 3's performance would improve as k increases.



(a) $\alpha_0(\alpha_0 + 1)\sigma_k(\hat{M}_2)$ **(b)** $|\tilde{\gamma}_s - \gamma_s|$

Figure 4. Visualization of **(a)** the k^{th} singular values of $\hat{M}_2$ and **(b)** the smallest singular value gap of $\widehat{T}$ using $100k$ documents.

Small α_0 vs Large α_0: The concentration parameter of the topic distribution α_0 plays an important role in the utility loss. An interesting observation is that the spectral method's performance is more advantageous at smaller values of α_0. This leads to less mixing between topics in each document. Config. 4's performance is affected by α_0 more than other configurations. As α_0 gets smaller, the utility loss for Config. 4 converges to that of Config. 3.

Small Corpus vs Large Corpus: Figure 2c considers the limited data setting, $N = 10^4$. Config. 4's advantage decreases as the number of documents decreases. Config. 2 exhibits robustness with a decreased number of documents.

Wikipedia Dataset: We implement our methods on the wikipedia dataset and verify the performance by comparing with differentially private variational inference. The vocabulary size is truncated to be $d = 8000$. Config. 1 is not scalable, since adding noise to the third order moment (dimensionality $d \times d \times d$ $\hat{M}_3$) is infeasible due to memory constraints. Storing $\hat{M}_3$ when $d = 8000$ requires 2 terabytes of memory. We therefore only run Config. 2 - 4, in which dimensionality reduction is used, subverting the need to explicitly form $\hat{M}_3$.

As shown in Figure 3 where the held-out perplexity scores on Wikipedia are compared with variational inference, our method achieves better perplexities under the same privacy levels. As we observe in the Wiki results in Figure 3, performance of Config.3 is improved under larger number of topics k, confirming our theory.

An interesting observation from Figure 3 is that sometimes DP-algorithms which introduces noises could help the algorithm to train better, in analogy to the well-known result of noisy gradient descent escapes from saddle point (while gradient descent gets trapped) in nonconvex optimization (Ge et al., 2015). Config. 3 achieves better results than the unnoised spectral method.

8. Conclusion

We provide an end-to-end analysis of differentially private Latent Dirichlet Allocation model using a spectral algorithm. The algorithm involves a dataflow that permits different locations for injecting noise and features a delicate data-dependent method that calibrates the noise to a differentially privately released high-probability upper bound of the local-sensitivities. We present a detailed utility analysis which shows that the proposed methods can provably recover the model parameters. To the best of out knowledge, these are *the first* differentially private topic methods that come with a provable consistency guarantee. Moreover, private spectral-LDA methods dominates the current state-of-the-art —differentially private variational inference — in all our experiments, which provides a compelling empirical example of spectral learning methods becoming a more preferable choice when differential privacy is required.

While we focused on LDA, the same technique can be used in other models that can be learned using a tensor-spectral approach. We expect similar improvements in private unsupervised learning to hold for stochastic block models, Gaussian mixture models and hidden Markov models.

Acknowledgment

YW was supported by the start-up grant at UCSB Computer Science and generous gifts from Amazon Web Services, Adobe and NEC Labs. FH is supported by startup fund from Department of Computer Science of University of Maryland, National Science Foundation IIS-1850220 CRII Award 030742- 00001, DOD-DARPA-Defense Advanced Research Projects Agency Guaranteeing AI Robustness against Deception (GARD), Laboratory for Physical Sciences at University of Maryland, and Adobe, Capital One and JP Morgan faculty fellowships.

References

Anandkumar, A., Foster, D. P., Hsu, D. J., Kakade, S. M., and Liu, Y.-K. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pp. 917–925, 2012.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014a.

Anandkumar, A., Ge, R., and Janzamin, M. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *arXiv preprint arXiv:1402.5180*, 2014b.

Balle, B. and Wang, Y.-X. Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising. In *International Conference on Machine Learning (ICML-18)*, pp. 403–412, 2018.

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

Blocki, J., Blum, A., Datta, A., and Sheffet, O. The johnson-lindenstrauss transform itself preserves differential privacy. In *IEEE Symposium on Foundations of Computer Science (FOCS-12)*, pp. 410–419. IEEE, 2012.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 267–284, 2019.

Chaudhuri, K., Sarwate, A., and Sinha, K. Near-optimal differentially private principal components. In *Advances in Neural Information Processing Systems*, pp. 989–997, 2012.

Dwork, C. and Lei, J. Differential privacy and robust statistics. In *ACM symposium on Theory of computing (STOC-09)*, volume 9, pp. 371–380, 2009.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

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

Dwork, C., Talwar, K., Thakurta, A., and Zhang, L. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 11–20. ACM, 2014b.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points — online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pp. 797–842, 2015.

Hsu, D., Kakade, S., Zhang, T., et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.

Imtiaz, H. and Sarwate, A. D. Symmetric matrix perturbation for differentially-private principal component analysis. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 2339–2343. IEEE, 2016.

Kapralov, M. and Talwar, K. On differentially private low rank approximation. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1395–1414. SIAM, 2013.

Mironov, I. Rényi differential privacy. In *Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In *ACM symposium on Theory of computing (STOC-07)*, pp. 75–84. ACM, 2007.

Park, M., Foulds, J., Chaudhuri, K., and Welling, M. Private topic modeling. *arXiv preprint arXiv:1609.04120*, 2016.

Park, M., Foulds, J. R., Choudhary, K., and Welling, M. DP-EM: differentially private expectation maximization. In *International Conference on Artificial Intelligence*

and Statistics (AISTATS-17), volume 54, pp. 896–904. PMLR, 2017.

Park, M., Foulds, J., Chaudhuri, K., and Welling, M. Variational bayes in private settings (vips). *Journal of Artificial Intelligence Research*, 68:109–157, 2020.

Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Schein, A., Wu, Z., Schofield, A., Zhou, M., and Wallach, H. Locally private bayesian inference for count modeling. In *International Conference on Machine Learning*, 2019.

Stewart, G. W. Matrix perturbation theory, 1990.

Stewart, G. W. Perturbation theory for the singular value decomposition. Technical report, 1998.

Tao, T. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Tomioka, R. and Suzuki, T. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

Wang, P.-A. and Lu, C.-J. Tensor decomposition via simultaneous power iteration. In *International Conference on Machine Learning*, pp. 3665–3673, 2017.

Wang, Y. and Anandkumar, A. Online and differentially-private tensor decomposition. In *Advances in Neural Information Processing Systems*, pp. 3531–3539, 2016.

Wang, Y. and Blei, D. M. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, pp. 1–15, 2018.

Wang, Y.-X. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

Zou, J. Y., Hsu, D. J., Parkes, D. C., and Adams, R. P. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, pp. 2238–2246, 2013.