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, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

[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.

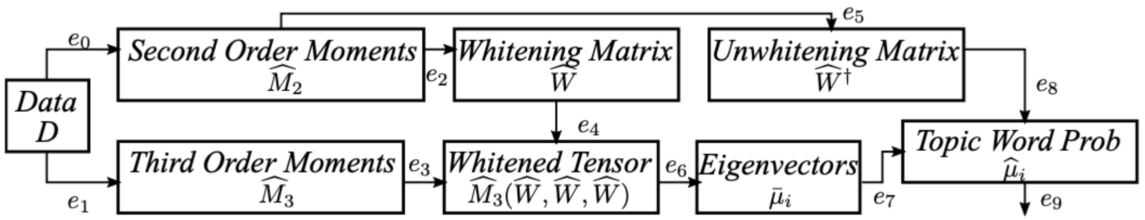


Figure 1. Algorithmic flow of end-to-end spectral learning algorithm to learning LDA topic model.

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*

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

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.

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*

[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 $k \times k \times k$.

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 $Z(\epsilon, \delta, \tilde{\Delta}_f(D)) = \mathcal{N}(0, \frac{\tilde{\Delta}_f(D)^2}{2\epsilon^2}(\sqrt{\epsilon + \log(1/\delta)} + \sqrt{\log(1/\delta)})^2)$).*

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 $\gamma_s$ 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 $\gamma_s$ 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 $\|\mu_i - \hat{\mu_i}\|_2 \leq O(\frac{(\alpha_0+1)^2k^3}{p_{\mathsf{min}}^2\sigma_k(\mu)\sqrt{N}}) = \tilde{O}(\frac{k^3}{\sqrt{N}})$, where $p_{\mathsf{min}} = \min_i \frac{\alpha_i}{\alpha_0}$ 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 $\left\|\mu_i - \mu_i^{\mathsf{DP}}\right\|$ using Config. 1 to guarantee $(\epsilon_2+\epsilon_3+\epsilon_5, \delta_2+\delta_3+\delta_5)$-DP is $O(\frac{(\alpha_0+1)^2k^3}{p_{\mathsf{min}}^2\sigma_k(\mu)\sqrt{N}}) + O(\frac{\sqrt{\sigma_1(\hat{M}_2)k}}{\gamma_s}((\frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_2,\delta_2})^3 + \frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_3,\delta_3}) + \frac{\sqrt{\sigma_1(\hat{M}_2)d}}{\sigma_k(\hat{M}_2)N}\tau_{\epsilon_5,\delta_5} + \sqrt{\sigma_1(\hat{M}_2) + \frac{\sqrt{d}}{N}\tau_{\epsilon_5,\delta_5}}\frac{\sqrt{k}}{\gamma_s}\Big[(\frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)}\tau_{\epsilon_2,\delta_2})^3 + \frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_3,\delta_3}\Big])$.*

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

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

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

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.*

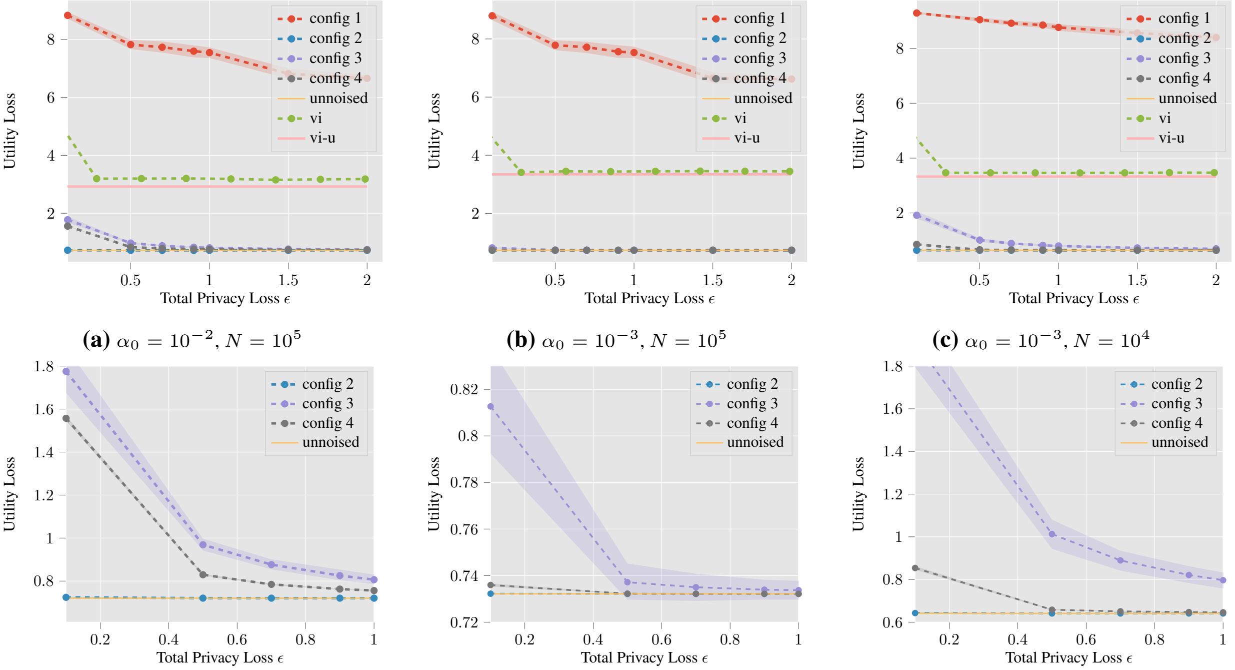
6.3. Comparison with DP VI

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.

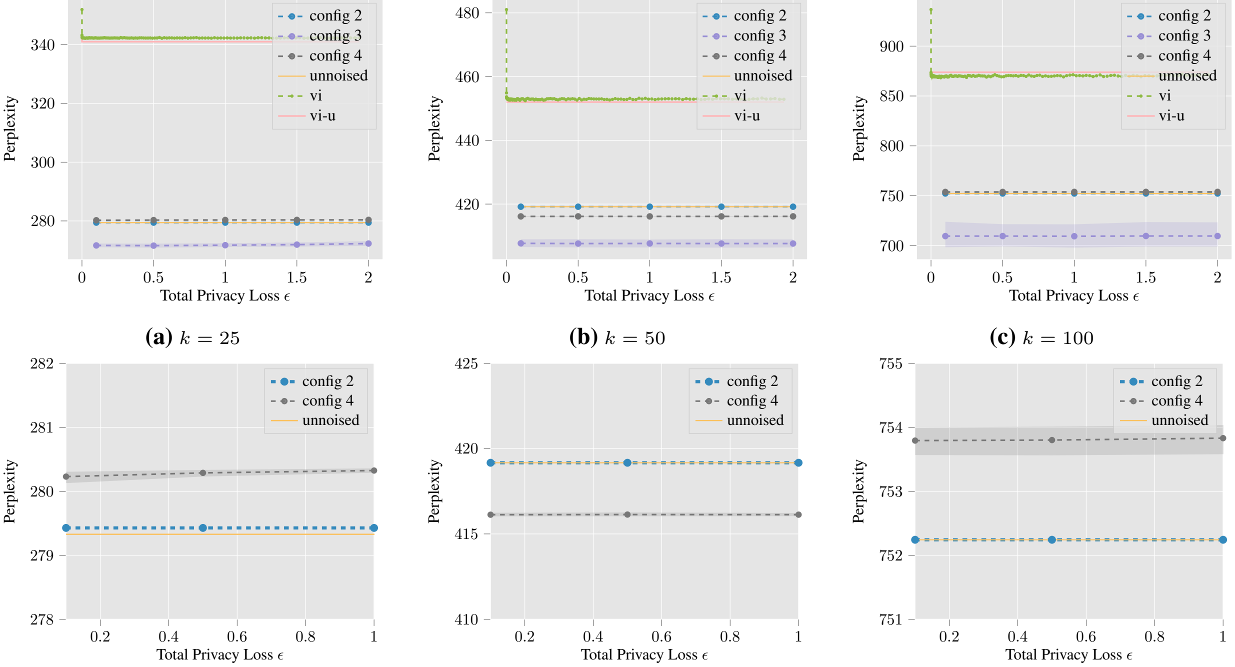
To evaluate Configuration 1, we set the vocabulary size and



(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)

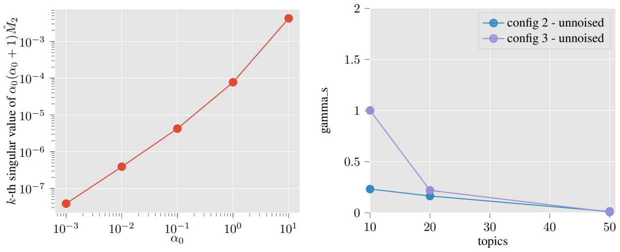
Figure 3. Perplexity scores of **our method under all configurations** vs **the differentially private VI** on Wikipedia data over varying total privacy loss ϵ_{total} while fixing the $\delta = 10^{-4}$. Number of words $d = 8000$, number of documents $N = 50000$, $\alpha_0 = 0.01$.

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 $\widehat{\mathcal{T}}$ 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.

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

A. Differential Privacy Review

Differential privacy was developed in (Dwork et al., 2006) and has been increasingly adopted as the *de facto* mathematical definition for privacy in statistical, machine learning and data science applications. We include additional information in this section that is relevant to this paper, but will defer more exposition to recent book (Dwork et al., 2014a) and the references therein.

Definition 21 (**Utility Loss & Error**). *Let $f : D \to Y$ be a random algorithm and $f^{\mathsf{DP}}(X)$ be the differentially private version of f. For some value $x \in D$, let $y \in Y$ be the ground truth value. Then define $\left\|f(x) - f^{\mathsf{DP}}(X)\right\|_F$ as the **utility loss** for this input. Additionally, define $\left\|y - f^{\mathsf{DP}}(X)\right\|_F$ as the **error** for this input.*

The gaussian mechanism makes a random algorithm differentially private by adding specifically designed Gaussian noise to the output.

Proposition 22. *[**Gaussian mechanism**] Let $f : D \to Y$ ($Y \subset \mathbb{R}^k$) be a random algorithm with ℓ_2 sensitivity Δ_f. Let $g \in \mathbb{R}^k$ and each coordinate g_i be sampled i.i.d. from $\mathcal{N}(0, \Delta^2_{f,\epsilon,\delta})$, where $\Delta_{f,\epsilon,\delta} = \Delta_f \; \tau_{\epsilon,\delta} = \frac{\Delta_f\sqrt{2\ln(1.25/\delta)}}{\epsilon}$. Then the output $f_{\mathsf{DP}} = f + g$ is (ϵ, δ) differentially private if $0 < \epsilon \leq 1$.*

The above bound is used for theoretical purposes only, a tighter and more general calibration of the Gaussian mechanism that does not require $\epsilon \leq 1$ is to set

$$\sigma = \frac{\Delta_f}{2\epsilon}\left(\sqrt{\epsilon + \log(1/\delta)} + \sqrt{\log(1/\delta)}\right).$$

Moreover, the optimal calibration (no closed-form formula available) was proposed in (Balle & Wang, 2018) and is available through, e.g., *autodp.calibrator*: `https://github.com/yuxiangw/autodp`.

Differential privacy composes over multiple DP releases.

Proposition 23. *[**Composition theorem**] Let $f_1^{\mathsf{DP}}(X), \ldots, f_n^{\mathsf{DP}}(X)$ be n differentially private algorithms with privacy parameters $(\epsilon_1, \delta_1), \ldots, (\epsilon_n, \delta_n)$. Then $g^{\mathsf{DP}}(X) = f(f_1^{\mathsf{DP}}(X), \ldots, f_n^{\mathsf{DP}}(X))$ is $(\epsilon_1 + \ldots + \epsilon_n, \delta_1 + \ldots + \delta_n)$ differentially private.*

This is what we called a simple composition where epsilon increases linearly. There is also an advanced composition where privacy loss for accessing for k times obey that $\sqrt{k}$, see, e.g., Section 3.5 of (Dwork et al., 2014a). Increasingly, the advanced composition and other privacy loss computation has been conducted numerically using modern tools such as Concentrated Differential privacy (Bun & Steinke, 2016) and Renyi Differential Privacy (or equivalently the moments accountant) (Mironov, 2017). We used simple composition in our theoretical analysis and for calibrating noise to privacy so as to be comparable to older literature that does not take advantage of the modern tool (Park et al., 2016; Wang & Anandkumar, 2016).

B. Latent Dirichlet Allocation

LDA, despite being a bag of words model, allows modeling of the mixed topics in a document to account for the more general case in which a document belongs to several different latent classes (topics) simultaneously. Latent Dirichlet Allocations has two major model parameters: topic prior $\boldsymbol{\alpha}$ and topic-word matrix $\boldsymbol{\mu}$. Topic prior $\boldsymbol{\alpha}$ determines the topic proportions and the topic word matrix controls the word distribution per topic.

Topic Proportions The proportion of words in topics, known as *topic proportion* (denoted as θ_n for document n), is drawn from a Dirichlet distribution (topic prior) parameterized by $\alpha = (\alpha_1, \ldots, \alpha_k)$, with density $P_\alpha(\theta = \theta_n) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k}\Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_{n,i}^{\alpha_k - 1}$, where $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.

Topic-Word Matrix Under a topic i, tokens in the documents are assumed to be generated in a conditionally independent manner through μ_i, i.e., token $x_1 \sim \mathrm{Cat}(d, \mu_i)$ where $\mathrm{Cat}(d, \mu_i)$ denotes the categorical distribution. Under different topics, these conditional distributions μ_i are linearly independent, $\forall i \in [k]$.

With the definition of the two major parameters, we now describe the generative model of LDA topic model. The process involves generating topics first, followed by tokens.

Topic Generation LDA remains simple as each token in the corpus belongs to one of the k topics only, although tokens in the same document could belong to different topics. We denote the topic of token j in document n as $z_{n,j}$. Therefore, topics generated are categorical $z_{n,j} \in [k]$ and distributed according to θ_n, i.e., $z_{n,j} \sim \text{Cat}(k, \theta_n)$ where $\text{Cat}(k, \theta_n)$ denotes the categorical distribution.

Word Generation Let x denote the tokens. After determining the topic of the token j, $z_{n,j}$, token j is generated conditionally independently through $\mu_{z_{n,j}}$, i.e., token $\sim \text{Cat}(d, \mu_{z_{n,j}})$. In a document n, if the j'^{th} token $x_{n,j'}$ is the v-th word in the dictionary, then $x_{n,j'} = e_v$ where e_v is a one-hot encoding, i.e., $x_{n,j'}(j) = 0 \; \forall j \neq v$ and $x_{n,j'}(j) = 1$ if $j = v$. Let l_n be the length of document n, random realizations of token x, i.e., $\{x_{n,j'}\}_{j'=1}^{l_n}$, are i.i.d.

Term-Document Matrix The term-document matrix $D \in \mathbb{N}_0^{d \times N}$. The n^{th} column in D is denoted by c_n, where its j^{th} component $c_n(j) =$ number of times word j in the vocabulary appeared in document n. This means that $c_n = \sum_{j'=1}^{l_n} x_{n,j'}$ where l_n is the number of words in document n. Clearly, $l_n = \sum_j^d c_n(j) = \|c_n\|_1$.

C. Method of Moments for Latent Dirichlet Allocation

Empirical Moment Estimators The moments that we obtain are not the population moments but rather empirically estimated moments from the given data set. We list the forms of first, second, and third order empirical moment estimators for the single topic case as shown in (Zou et al., 2013). Given a document n, the following quantities are calculated.

$$\tilde{\tilde{M}}_1^n = \frac{c_n}{l_n} \tag{6}$$

$$\tilde{\tilde{M}}_2^n = \frac{1}{2\binom{l_n}{2}}(c_n \otimes c_n - \text{diag}(c_n)) \tag{7}$$

$$\begin{aligned}\tilde{\tilde{M}}_3^n = \frac{1}{6\binom{l_n}{3}}\Big(& c_n \otimes c_n \otimes c_n + 2\sum_{i=1}^{d} c_n(i)(e_i \otimes e_i \otimes e_i) \\ & - \sum_{i=1}^{d}\sum_{j=1}^{d} c_n(i)c_n(j)(e_i \otimes e_i \otimes e_j + e_i \otimes e_j \otimes e_j + e_j \otimes e_i \otimes e_j)\Big)\end{aligned} \tag{8}$$

The empirically estimated moments are the averages of these quantities over the entire data set. Specifically,

Lemma 24. *Single Topic Empirical Moment Estimators(Propositions 3 and 4 in (Zou et al., 2013))*

$$\hat{\mathbb{E}}[x_1] = \frac{1}{N}\sum_{n=1}^{N} \tilde{\tilde{M}}_1^n \tag{9}$$

$$\hat{\mathbb{E}}[x_1 \otimes x_2] = \frac{1}{N}\sum_{n=1}^{N} \tilde{\tilde{M}}_2^n \tag{10}$$

$$\hat{\mathbb{E}}[x_1 \otimes x_2 \otimes x_3] = \frac{1}{N}\sum_{n=1}^{N} \tilde{\tilde{M}}_3^n \tag{11}$$

$$\tag{12}$$

Further these moments are unbiased, i.e.:

$$\mathbb{E}[\hat{\mathbb{E}}[x_1]] = \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_1^n] = \mathbb{E}[x_1] \tag{13}$$

$$\mathbb{E}[\hat{\mathbb{E}}[x_1 \otimes x_2]] = \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_2^n] = \mathbb{E}[x_1 \otimes x_2] \tag{14}$$

$$\mathbb{E}[\hat{\mathbb{E}}[x_1 \otimes x_2 \otimes x_3]] = \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_3^n] = \mathbb{E}[x_1 \otimes x_2 \otimes x_3] \tag{15}$$

$$\tag{16}$$

Note that this lemma implies that: $\mathbb{E}[\tilde{\tilde{M}}_1^n] = \mathbb{E}[x_1], \mathbb{E}[\tilde{\tilde{M}}_2^n] = \mathbb{E}[x_1 \otimes x_2]$, and that $\mathbb{E}[\tilde{\tilde{M}}_3^n] = \mathbb{E}[x_1 \otimes x_2 \otimes x_3]$ for any sampled document n.

We extend the single topic moment estimators of (Zou et al., 2013) to the LDA case.

Lemma 25. *Empirical Moment estimators for LDA*

$$\hat{M}_1 = \frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_1^n \tag{17}$$

$$\hat{M}_2 = \frac{1}{N}\sum_{n=1}^{N}\left[\tilde{\tilde{M}}_2^n\right] - \frac{a}{2\binom{N}{2}}\left[\sum_{m,n=1}^{N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m - \sum_{n=1}^{N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^n\right] \tag{18}$$

$$\hat{M}_3 = \left[\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_3^n + \mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3 + \mathbf{b}\right] \tag{19}$$

where

$$\mathbf{B}_1 \overset{\mathsf{def}}{=} \frac{b}{2\binom{N}{2}}\left[\Big(\sum_{n=1}^{N}\tilde{\tilde{M}}_2^n\Big) \otimes \Big(\sum_{n=1}^{N}\tilde{\tilde{M}}_1^n\Big)\right], \tag{20}$$

$$\mathbf{b} \overset{\mathsf{def}}{=} c\left[\Big(\sum_{n=1}^{N}\tilde{\tilde{M}}_1^n\Big) \otimes \Big(\sum_{n=1}^{N}\tilde{\tilde{M}}_1^n\Big) \otimes \Big(\sum_{n=1}^{N}\tilde{\tilde{M}}_1^n\Big)\right], \tag{21}$$

$\mathbf{B}_2$ *and* $\mathbf{B}_3$ *are formed from* $\mathbf{B}_1$ *by permuting, i.e.,* $[\mathbf{B}_2]_{ijk} = [\mathbf{B}_1]_{ikj}$ *and* $[\mathbf{B}_3]_{ijk} = [\mathbf{B}_1]_{kij}$. *Further,* $a = \frac{\alpha_0}{\alpha_0+1}, b = \frac{-\alpha_0}{\alpha_0+2}, c = \frac{2\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)}$.

Now we prove that these estimators are unbiased.

Lemma 26 (The LDA Moment Estimators are Unbiased)**.** *The estimators defined in definition 25 are unbiased, i.e.,*

$$\mathbb{E}[\hat{M}_1] = M_1 \tag{22}$$

$$\mathbb{E}[\hat{M}_2] = M_2 \tag{23}$$

$$\mathbb{E}[\hat{M}_3] = M_3 \tag{24}$$

Proof. **First order moment**:

$$\mathbb{E}[\hat{M}_1] = \mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_1^n] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\tilde{\tilde{M}}_1^n] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\frac{c_n}{l_n}] \tag{25}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{l_n}\mathbb{E}[c_n] = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{l_n}\mathbb{E}[\sum_{i=1}^{l_n}x_{n,i}] = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{l_n}\sum_{i=1}^{l_n}\mathbb{E}[x_{n,i}] \tag{26}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{l_n}\sum_{i=1}^{l_n}\mathbb{E}[x_1] = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{l_n}l_n\mathbb{E}[x_1] = \frac{1}{N}N\mathbb{E}[x_1] = \mathbb{E}[x_1] = M_1 \tag{27}$$

Second order moment: The first term of $\hat{M}_2$ is actually the estimator the single-topic second order moment and $\mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_2^n]] = \mathbb{E}[x_1 \otimes x_2]$ see proposition 3 in (Zou et al., 2013) and its appendix for the proof. Now we have:

$$
\begin{aligned}
&\mathbb{E}\left[\frac{a}{2\binom{N}{2}}\left[\sum_{m,n=1}^{N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m - \sum_{n=1}^{N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^n\right]\right] && (28)\\
=&\mathbb{E}\left[\frac{a}{2\binom{N}{2}}\left[\sum_{\substack{m=1\\n=1\\m\neq n}}^{N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m + \sum_{n=1}^{N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^n - \sum_{n=1}^{N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^n\right]\right] && (29)\\
=&\mathbb{E}\left[\frac{a}{2\binom{N}{2}}\sum_{\substack{m=1\\n=1\\m\neq n}}^{N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m\right] && (30)\\
=&\frac{a}{2\binom{N}{2}}\sum_{\substack{m=1\\n=1\\m\neq n}}^{N}\mathbb{E}[\tilde{\tilde{M}}_1^n] \otimes \mathbb{E}[\tilde{\tilde{M}}_1^n] = \frac{a}{2\binom{N}{2}}\sum_{\substack{m=1\\n=1\\m\neq n}}^{N}\mathbb{E}[x_1] \otimes \mathbb{E}[x_1] && (31)\\
=&\frac{a}{N(N-1)}N(N-1)\mathbb{E}[x_1] \otimes \mathbb{E}[x_1] = a\mathbb{E}[x_1] \otimes \mathbb{E}[x_1] && (32)
\end{aligned}
$$

Thus, we have that: $\mathbb{E}[\hat{M}_2] = \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0+2}\mathbb{E}[x_1] \otimes \mathbb{E}[x_1] = M_2$.

Third order moment: Similar to the second order moment, the first term of $\hat{M}_3$ is the estimator the single-topic second order moment and $\mathbb{E}[\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_3^n]] = \mathbb{E}[x_1 \otimes x_2 \otimes x_3]$ as shown in proposition 4 in (Zou et al., 2013) and proved in its appendix. We need to prove that (1): $\mathbb{E}[\mathbf{B}_1] = b\mathbb{E}[x_1 \otimes x_2 \otimes \mathbb{E}[x_3]]$, note that $\mathbb{E}[x_3] = M_1$ and (2): $\mathbb{E}[\mathbf{b}] = cE[x_1] \otimes E[x_1] \otimes E[x_1] = cM_1 \otimes M_1 \otimes M_1 \otimes M_1$. Since $\mathbf{B}_2$ and $\mathbf{B}_3$ are permuted version of $\mathbf{B}_1$ their proofs follow from the proof of $\mathbf{B}_1$.

For $\mathbf{B}_1$ we simplify the expression and then show that the expectation of the resultant is equal to the desired moment:

$$
\begin{aligned}
\mathbb{E}[\mathbf{B}_1] =&\frac{b}{2\binom{N}{2}}\mathbb{E}\left[\Big(\sum_{n=1}^{N}\tilde{\tilde{M}}_2^n\Big) \otimes \Big(\sum_{n=1}^{N}\tilde{\tilde{M}}_1^n\Big) - \sum_{n=1}^{N}\Big(\tilde{\tilde{M}}_2^n \otimes \tilde{\tilde{M}}_1^n\Big)\right] && (33)\\
=&\frac{b}{2\binom{N}{2}}\mathbb{E}\left[\sum_{\substack{m=1,n=1\\m\neq n}}^{N}\Big(\tilde{\tilde{M}}_2^n \otimes \tilde{\tilde{M}}_1^m\Big) + \sum_{n=1}^{N}\Big(\tilde{\tilde{M}}_2^n \otimes \tilde{\tilde{M}}_1^n\Big) - \sum_{n=1}^{N}\Big(\tilde{\tilde{M}}_2^n \otimes \tilde{\tilde{M}}_1^n\Big)\right] && (34)\\
=&\frac{b}{2\binom{N}{2}}\mathbb{E}\left[\sum_{\substack{m=1,n=1\\m\neq n}}^{N}\Big(\tilde{\tilde{M}}_2^n \otimes \tilde{\tilde{M}}_1^m\Big)\right] && (35)\\
=&\frac{b}{2\binom{N}{2}}\sum_{\substack{m=1,n=1\\m\neq n}}^{N}\mathbb{E}\Big[\tilde{\tilde{M}}_2^n\Big] \otimes \mathbb{E}\Big[\tilde{\tilde{M}}_1^m\Big] && (36)\\
=&\frac{b}{2\binom{N}{2}}\sum_{\substack{m=1,n=1\\m\neq n}}^{N}\mathbb{E}[x_1 \otimes x_2] \otimes \mathbb{E}[x_3] && (37)\\
=&\frac{b}{N(N-1)}N(N-1)\mathbb{E}[x_1 \otimes x_2] \otimes \mathbb{E}[x_3] && (38)\\
=&b\mathbb{E}[x_1 \otimes x_2] \otimes \mathbb{E}[x_3] && (39)\\
=&b\mathbb{E}\big[x_1 \otimes x_2 \otimes \mathbb{E}[x_3]\big] && (40)
\end{aligned}
$$

For $\mathbf{b}$ identity 38 is applied, this leads to the following

$$\mathbb{E}[\mathbf{b}] = \frac{c}{6\binom{N}{3}}\mathbb{E}\Bigg[\Bigg(\sum_{i=1}^{N}(\tilde{\tilde{M}}_1^n)^{\otimes^3} + 3\sum_{\substack{n=1,m=1\\n\neq m}}^{N,N}(\tilde{\tilde{M}}_1^n)^{\otimes^2}\tilde{\tilde{M}}_1^m + \sum_{\substack{n=1,m=1,p=1\\n\neq m,m\neq p,p\neq n}}^{N,N,N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m \otimes \tilde{\tilde{M}}_1^p \tag{41}$$

$$-3\sum_{m=1}^{N}\Big(\sum_{n=1}^{N}\Big(\tilde{\tilde{M}}_1^n\Big)^{\otimes^2} \otimes \Big(\tilde{\tilde{M}}_1^m\Big)\Big) + 2\sum_{n=1}^{N}\Big(\tilde{\tilde{M}}_1^n\Big)^{\otimes^3}\Bigg)\Bigg] \tag{42}$$

$$= \frac{c}{6\binom{N}{3}}\mathbb{E}\Bigg[\sum_{\substack{n=1,m=1,p=1\\n\neq m,m\neq p,p\neq n}}^{N,N,N}\tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m \otimes \tilde{\tilde{M}}_1^p\Bigg] \tag{43}$$

$$= \frac{c}{N(N-1)(N-2)}(N)(N-1)(N-2)\mathbb{E}[\tilde{\tilde{M}}_1^n] \otimes \mathbb{E}[\tilde{\tilde{M}}_1^m] \otimes \mathbb{E}[\tilde{\tilde{M}}_1^p] \tag{44}$$

$$= c\mathbb{E}[x_1] \otimes \mathbb{E}[x_1] \otimes \mathbb{E}[x_1] \tag{45}$$

Combing these results and plugging the values for a, b,and c we get:

$$\mathbb{E}[\hat{M}_3] = \mathbb{E}[x_1 \otimes x_2 \otimes x_3] - \frac{\alpha_0}{\alpha_0 + 2}\Bigg(\mathbb{E}[x_1 \otimes x_2 \otimes \mathbb{E}[x_3]] + \mathbb{E}[x_1 \otimes \mathbb{E}[x_2] \otimes x_3] + \mathbb{E}[\mathbb{E}[x_1] \otimes x_2 \otimes x_3]\Bigg) \tag{46}$$

$$+ \frac{2\alpha_0^2}{(\alpha_0 + 1)(\alpha_0 + 2)}\mathbb{E}[x_1] \otimes \mathbb{E}[x_1] \otimes \mathbb{E}[x_1] = M_3 \tag{47}$$

□

D. Lemmas regarding Dirichlet Moments

This section introduces two lemmas regarding the moments of the dirichlet distribution that will be useful for the proof of Lemma 3.

D.1. Dirichlet Moments

Lemma 27. *The first, second and third moments of dirichlet distribution are*

$$\mathbb{E}[\theta] = \frac{1}{\alpha_0}\alpha \tag{48}$$

$$\mathbb{E}[\theta \otimes \theta] = \frac{1}{\alpha_0(\alpha_0 + 1)}[\alpha \otimes \alpha + \sum_{t=1}^{T}\alpha_t e_t \otimes e_t] \tag{49}$$

$$\mathbb{E}[\theta \otimes \theta \otimes \theta] = \frac{1}{\alpha_0(\alpha_0 + 1)(\alpha_0 + 2)}[\alpha \otimes \alpha \otimes \alpha + \sum_{t=1}^{T}\alpha_t e_t \otimes e_t \otimes \alpha$$

$$+ \sum_{t=1}^{T}\alpha_t \alpha \otimes e_t \otimes e_t + \sum_{t=1}^{T}\alpha_t e_t \otimes \alpha \otimes e_t + 2\sum_{t=1}^{T}\alpha_t e_t \otimes e_t \otimes e_t] \tag{50}$$

D.2. Raw Moments

Lemma 28.

$$\mathbb{E}[x_1] = \mu\mathbb{E}[\theta] \tag{51}$$

$$\mathbb{E}[x_1 \otimes x_2] = \mu\mathbb{E}[\theta \otimes \theta]\mu^\top \tag{52}$$

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \mathbb{E}[\theta \otimes \theta \otimes \theta](\mu, \mu, \mu) \tag{53}$$

Proof. **First Order Moments** Let us omit n and use x_1 to denote a token in any document, and we will use x_2 and x_3 to denote other two tokens in the same document. The the expectation of a token is

$$\mathbb{E}[x_1] = \mathbb{E}[x_2] = \mathbb{E}[x_3] = \mathbb{E}[\mathbb{E}[x_1|\theta]] = \mu\mathbb{E}[\theta] \tag{54}$$

This is called the first order moment.

Second Order Moments The second order moment is defined as

$$\begin{aligned}
\mathbb{E}[x_1 \otimes x_2] &= \mathbb{E}[\mathbb{E}[x_1 \otimes x_2 | \theta]] && (55)\\
&= \sum_{i,i'} \mathbb{E}[x_1 \otimes x_2 | z_{n,j} = e_i, z_{n,k} = e_{i'}] P(z_{n,j} = e_i, z_{n,k} = e_{i'}) && (56)\\
&= \sum_{i,i'} \mathbb{E}[x_1 | z_{n,j} = e_i] \otimes \mathbb{E}[x_2 | z_{n,k} = e_{i'}] P(z_{n,j} = e_i, z_{n,k} = e_{i'}) && (57)\\
&= \sum_{i,i'} \mu e_i \otimes (\mu e_{i'}) P(z_{n,j} = e_i, z_{n,k} = e_{i'}) && (58)\\
&= \mu \sum_{i,i'} e_i \otimes e_{i'} P(z_{n,j} = e_i, z_{n,k} = e_{i'}) \mu^\top && (59)\\
&= \mu \mathbb{E}[\theta \otimes \theta] \mu^\top && (60)
\end{aligned}$$

Third Order Moments The third order moment is defined as

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \mathbb{E}[\mathbb{E}[x_1 \otimes x_2 \otimes x_3 | \theta]] = \mathbb{E}[\theta \otimes \theta \otimes \theta](\mu, \mu, \mu) \quad (61)$$

To clarify the notations, $x \otimes y$ is a $length(x)$-by-$length(y)$ matrix which has entries $[x \otimes y]_{i,j} = x_i y_j$. And $\mathbb{E}[\theta \otimes \theta \otimes \theta](\mu, \mu, \mu)$ is a tucker with core tensor $\mathbb{E}[\theta \otimes \theta \otimes \theta]$ and projection μ in all three modes. □

E. Proof of Lemma 3

The lemma relates the LDA moments to the model parameters α and μ.

Proof. In order to prove this relation, we combine Lemmas 27 and Lemma 28 to prove the forms of M_1, M_2 and M_3 in Lemma 3 as follows.

$$\begin{aligned}
M_1 &= \mathbb{E}[x_1] = \mu \mathbb{E}[\theta] = \sum_{i=1}^{k} \frac{\alpha_i}{\alpha_0} \mu_i && (62)\\
M_2 &= \mathbb{E}[x_1 \otimes x_2] - \frac{\alpha_0}{\alpha_0 + 1} \mathbb{E}[x_1] \otimes \mathbb{E}[x_1] && (63)\\
&= \mathbb{E}[\theta \otimes \theta](\mu, \mu) - \frac{1}{\alpha_0(\alpha_0 + 1)} M_1 \otimes M_1 && (64)\\
&= \sum_{i=1}^{k} \frac{\alpha_i}{\alpha_0(\alpha_0 + 1)} \mu_i \otimes \mu_i && (65)\\
M_3 &= \mathbb{E}[x_1 \otimes x_2 \otimes x_3] - \frac{1}{\alpha_0 + 2} (\mathbb{E}[x_1 \otimes x_2 \otimes \mathbb{E}[x_3]] + \mathbb{E}[x_1 \otimes \mathbb{E}[x_2] \otimes x_3] \\
&\quad + \mathbb{E}[\mathbb{E}[x_1] \otimes x_2 \otimes x_3]) + \frac{2}{\alpha_0(\alpha_0 + 1)(\alpha_0 + 2)} \mathbb{E}[x_1] \otimes \mathbb{E}[x_1] \otimes \mathbb{E}[x_1] && (66)\\
&= \mathbb{E}[\theta \otimes \theta \otimes \theta](\mu, \mu, \mu) && (67)\\
&\quad - \frac{1}{\alpha_0 + 2} \{\mathbb{E}[\theta \otimes \theta \otimes \mathbb{E}[\theta]] - \mathbb{E}[\theta \otimes \mathbb{E}[\theta] \otimes \theta] - \mathbb{E}[\mathbb{E}[\theta] \otimes \theta \otimes \theta]\}(\mu, \mu, \mu) && (68)\\
&\quad + \frac{2}{\alpha_0(\alpha_0 + 1)(\alpha_0 + 2)} M_1 \otimes M_1 \otimes M_1 && (69)\\
&= \sum_{i}^{k} \frac{2\alpha_i}{\alpha_0(\alpha_0 + 1)(\alpha_0 + 2)} \mu_i \otimes \mu_i \otimes \mu_i && (70)
\end{aligned}$$

□

F. Correctness of Method of Moments for Latent Dirichlet Allocation

Lemma 29 (Correctness of Method of Moments in Learning LDA (Anandkumar et al., 2012))**.** *Applying the method of moments over a corpus of N documents sampled iid. There exist universal constants $C_1, C_2 \geq 0$ such that if $N > C_1((\alpha_0+1)/p_{\mathsf{min}}^2\sigma_k(\mu)^2)$, then $\|\mu_i - \hat{\mu_i}\|_2 \leq C_2\frac{(\alpha_0+1)^2k^3}{p_{\mathsf{min}}^2\sigma_k(\mu)\sqrt{N}}$, where $p_{\mathsf{min}} = \min_i \frac{\alpha_i}{\alpha_0}$, μ is a matrix of stacked word-topic vectors, i.e. $\mu = [\mu_1| \ldots |\mu_k]$.*

G. Sensitivity Proofs

In proving the sensitivities for $\hat{M}_2$ and $\hat{M}_3$ we rely on the fact that frequently in the calculations, we encounter probability vectors, matrices, and tensors where the elements sum to 1. This is identical to the stating that the l_1 norm equals 1. Further, we note the following Lemma which essentially states that taking the outer product of a vector with a probability vector or probability matrix does not increase the l_q norm of the vector and in fact keeps it the same if $q = 1$.

Lemma 30 (Multiplying by probabilities does not change the norm)**.** *Let v_p, M_p be a probability vector, matrix, respectively and let v, u be ordinary vectors, matrices, respectively. Then the following holds:*

$$\left\|uv_p^T\right\|_q \leq \|u\|_q \text{, which is equal if } q = 1. \tag{71}$$

$$\text{If } T = M_p \otimes u \text{, then } \|T\|_q = \|M_p \otimes u\|_q \leq \|u\|_q \text{, which is equal if } q = 1. \tag{72}$$

Proof.

$$\left\|uv_p^T\right\|_q = (\sum_{i,j} |u_i v_{pj}|^q)^{1/q} = (\sum_i |u_i|^q \sum_j |v_{pj}|^q)^{1/q} = \|v\|_q \|u\|_q \leq \|u\|_q \,. \tag{73}$$

Where we used the fact that $\|x\|_1 \geq \|x\|_q$ for any $q \geq 1$ and that $\|v_p\|_1 = 1$. Thus the above inequality is tight if $q = 1$.

$$\|T\|_q = \|M_p \otimes u\|_q = \Big(\sum_{i,j,k} |M_{p_{i,j}} u_k|^q\Big)^{1/q} = \Big(\sum_k |u_k|^q \sum_{i,j} |M_{p_{i,j}}|^q\Big)^{1/q} = \|u\|_q \|M_p\|_q \leq \|u\|_q \,. \tag{74}$$

Where we used the fact that for any matrix M, $\|M\|_1 \geq \|M\|_q$ for any $q \geq 1$ [*] and that $\|M_p\|_1 = 1$. Thus the above inequality is tight if $q = 1$. □

Proposition 31. $\tilde{\tilde{M}}_1^n$ is a probability vector*,* $\tilde{\tilde{M}}_2^n$ is a probability matrix*, and* $\tilde{\tilde{M}}_3^n$ is a probability tensor.

Proof. The proof is immediate as these moments correspond to join probability estimates (Zou et al., 2013), specifically:

$$\tilde{\tilde{M}}_1^n(i) = \mathbb{P}[x_1 = i] \tag{75}$$
$$\tilde{\tilde{M}}_1^n(i,j) = \mathbb{P}[x_1 = i, x_2 = j] \tag{76}$$
$$\tilde{\tilde{M}}_1^n(i,j,k) = \mathbb{P}[x_1 = i, x_2 = j, x_3 = k] \tag{77}$$

□

G.1. Proof for Theorem 4 (sensitivity for $\hat{M}_2$)

Let Δ_2 be the l_1 sensitivity for $\hat{M}_2$, then Δ_2 is $\frac{2}{N} + \frac{\alpha_0}{\alpha_0+1}\frac{4}{N}$=$O(\frac{1}{N})$.

[*] These norms are obtained by extending the vector definition to matrices or simply vectorizing the matrix and then calculating the norm.

Proof. Let $\hat{M}_2$ and $\hat{M}'_2$ be two second order LDA moments generated from two neighboring corpora, WLOG assume the difference is in the n^{th} record, i.e. $D = [c_1|\dots|c_{N-1}|c_N]$ and $D' = [c_1|\dots|c_{N-1}|c'_N]$ then:

$$\hat{M}_2 - \hat{M}'_2 = \frac{1}{N}(\tilde{\tilde{M}}_2^{N} - \tilde{\tilde{M}}_2^{N'}) - \frac{a}{2\binom{N}{2}}\left(\left[\tilde{\tilde{M}}_1^{N} \otimes \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) + \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) \otimes \tilde{\tilde{M}}_1^{N}\right]\right. \tag{78}$$

$$\left. - \left[\tilde{\tilde{M}}_1^{N'} \otimes \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) + \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) \otimes \tilde{\tilde{M}}_1^{N'}\right]\right) \tag{79}$$

$$= \frac{1}{N}(\tilde{\tilde{M}}_2^{N} - \tilde{\tilde{M}}_2^{N'}) - \frac{a}{2\binom{N}{2}}\left((\tilde{\tilde{M}}_1^{N} - \tilde{\tilde{M}}_1^{N'}) \otimes \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) + \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) \otimes (\tilde{\tilde{M}}_1^{N} - \tilde{\tilde{M}}_1^{N'})\right) \tag{80}$$

$$= \frac{1}{N}(\tilde{\tilde{M}}_2^{N} - \tilde{\tilde{M}}_2^{N'}) - \frac{a}{N}\left((\tilde{\tilde{M}}_1^{N} - \tilde{\tilde{M}}_1^{N'}) \otimes \left(\frac{1}{N-1}\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) + \left(\frac{1}{N-1}\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) \otimes (\tilde{\tilde{M}}_1^{N} - \tilde{\tilde{M}}_1^{N'})\right) \tag{81}$$

Note that according to proposition (31) $\tilde{\tilde{M}}_1^{N}$ and $\tilde{\tilde{M}}_1^{N'}$ are probability vectors and $\tilde{\tilde{M}}_2^{N}$ and $\tilde{\tilde{M}}_2^{N'}$ are probability matrices. Further, $\left(\frac{1}{N-1}\sum_{n=1}^{N-1}\tilde{\tilde{M}}_2^{n}\right)$ is also a probability matrix since it's the normalized sum of probability matrices. We upper bound the l_1 norm of the expression by applying the triangular inequality and using lemma (30) for the terms involving a tensor product. This leads to the following:

$$\left\|\hat{M}_2 - \hat{M}'_2\right\|_1 \leq \frac{2}{N} + \frac{4a}{N} = \frac{2}{N} + \frac{\alpha_0}{\alpha_0+1}\frac{4}{N} = O(\frac{1}{N}) \tag{82}$$

$$\tag{83}$$

a was replaced by its expression as in the above $a = \frac{\alpha_0}{\alpha_0+1}$ in the above. □

G.2. Proof for Theorem 4 (sensitivity for $\hat{M}_3$)

Let Δ_3 be the l_1 sensitivity for $\hat{M}_3$, then Δ_3 is $\frac{2}{N} + \frac{4\alpha_0}{\alpha_0+2}\frac{1}{N} + \frac{12\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)}\frac{(N-1)}{N(N-2)} = O(\frac{1}{N})$.

Proof. Following a similar setting as in G.1 we have the two moments $\hat{M}_3$ and $\hat{M}'_3$ generated from two neighboring corpora. First we note that the expression of $\hat{M}_3$ and $\hat{M}'_3$ have the following form: $\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}_3^{n} + \mathbf{B}_1 + \mathbf{B}_2 + \mathbf{B}_3 + \mathbf{b}$. Effectively there are three kinds of terms: **(a)** $\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}$, **(b)** $\mathbf{B}_1$, and **(c)**$\mathbf{b}$. Since $\mathbf{B}_2$ and $\mathbf{B}_3$are permuted versions of $\mathbf{B}_1$ they have a similar behavior

(a) $\frac{1}{N}\sum_{n=1}^{N}\tilde{\tilde{M}}$**:** The first term difference between $\hat{M}_3$ and $\hat{M}'_3$ would result in $\frac{1}{N}(\tilde{\tilde{M}}_3^{N} - \tilde{\tilde{M}}_3^{N'})$.

$$\frac{1}{N}\left\|\tilde{\tilde{M}}_3^{N} - \tilde{\tilde{M}}_3^{N'}\right\|_1 \leq \frac{1}{N}\left(\left\|\tilde{\tilde{M}}_3^{N}\right\|_1 + \left\|\tilde{\tilde{M}}_3^{N'}\right\|_1\right) \leq \frac{2}{N} \tag{84}$$

Note that both $\tilde{\tilde{M}}_3^{N}$ and $\tilde{\tilde{M}}_3^{N'}$ are probability tensors.

(b) $\mathbf{B}_1$**:** Based on the minimized expression, the $\mathbf{B}_1$ term difference between $\hat{M}_3$ and $\hat{M}'_3$ is equal to:

$$\mathbf{B}_1 - \mathbf{B}'_1 = \frac{b}{2\binom{N}{2}}\left[\tilde{\tilde{M}}_2^{N} \otimes \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) + \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_2^{n}\right) \otimes \tilde{\tilde{M}}_1^{N}\right. \tag{85}$$

$$\left. - \tilde{\tilde{M}}_2^{N'} \otimes \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) - \left(\sum_{n=1}^{N-1}\tilde{\tilde{M}}_2^{n}\right) \otimes \tilde{\tilde{M}}_1^{N'}\right] \tag{86}$$

$$= \frac{b}{N}\left[\left(\tilde{\tilde{M}}_2^{N} - \tilde{\tilde{M}}_2^{N'}\right) \otimes \left(\frac{1}{N-1}\sum_{n=1}^{N-1}\tilde{\tilde{M}}_1^{n}\right) + \left(\frac{1}{N-1}\sum_{n=1}^{N-1}\tilde{\tilde{M}}_2^{n}\right) \otimes \left(\tilde{\tilde{M}}_1^{N} - \tilde{\tilde{M}}_2^{N}\right)'\right] \tag{87}$$

Note that $\frac{1}{N-1}\sum\limits_{n=1}^{N-1}\tilde{\tilde{M}}_1^n$ and $\frac{1}{N-1}\sum\limits_{n=1}^{N-1}\tilde{\tilde{M}}_2^n$ are probability vectors and matrices, respectively. Thus lemma 30 can be used to upper bound the l_1 norm, leading to the following:

$$\|\mathbf{B}_1 - \mathbf{B'}_1\|_1 \leq \frac{|b|}{N}(2+2) = \frac{4|b|}{N} = \frac{4\alpha_0}{\alpha_0+2}\frac{1}{N} \tag{88}$$

(c) b: Based on the minimized expression, the $\mathbf{b}$ term difference between $\hat{M}_3$ and $\hat{M}_3'$ is equal to:

$$\begin{aligned}
\mathbf{b} - \mathbf{b}' =& \frac{c}{6\binom{N}{3}}\Bigg[\Bigg(\tilde{\tilde{M}}_1^N \otimes \big(\sum_{\substack{m=1,p=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^m \otimes \tilde{\tilde{M}}_1^p\big) + \big(\sum_{\substack{n=1,p=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^N \otimes \tilde{\tilde{M}}_1^p\big) &(89)\\
&+ \big(\sum_{\substack{n=1,m=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m\big) \otimes \tilde{\tilde{M}}_1^N\Bigg) - \Bigg(\tilde{\tilde{M}}_1^{N'} \otimes \big(\sum_{\substack{m=1,p=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^m \otimes \tilde{\tilde{M}}_1^p\big) &(90)\\
&- \big(\sum_{\substack{n=1,p=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^{N'} \otimes \tilde{\tilde{M}}_1^p\big) - \big(\sum_{\substack{n=1,m=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m\big) \otimes \tilde{\tilde{M}}_1^N\Bigg)\Bigg] &(91)\\
=& \frac{c(N-1)}{N(N-2)}\Bigg[\big(\tilde{\tilde{M}}_1^N - \tilde{\tilde{M}}_1^{N'}\big) \otimes \big(\frac{1}{(N-1)^2}\sum_{\substack{m=1,p=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^m \otimes \tilde{\tilde{M}}_1^p\big) &(92)\\
&+ \big(\frac{1}{(N-1)^2}\sum_{\substack{n=1,p=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^n \otimes \big(\tilde{\tilde{M}}_1^N - \tilde{\tilde{M}}_1^{N'}\big) \otimes \tilde{\tilde{M}}_1^p\big) &(93)\\
&+ \big(\frac{1}{(N-1)^2}\sum_{\substack{n=1,m=1\\\text{distinct}}}^{N-1} \tilde{\tilde{M}}_1^n \otimes \tilde{\tilde{M}}_1^m\big) \otimes \big(\tilde{\tilde{M}}_1^N - \tilde{\tilde{M}}_1^{N'}\big)\Bigg] &(94)
\end{aligned}$$

Similarly, we have probability tensors so we use lemma 30 to bound the l_1 norm. This results in:

$$\|\mathbf{b} - \mathbf{b}'\|_1 \leq \frac{c(N-1)}{N(N-2)}(2+2+2) = \frac{6c(N-1)}{N(N-2)} = \frac{12\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)}\frac{(N-1)}{N(N-2)} \tag{95}$$

Combing the results from **(a), (b)** and **(c)**, we have the following bound:

$$\Delta_3 \leq \frac{2}{N} + \frac{4\alpha_0}{\alpha_0+2}\frac{1}{N} + \frac{12\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)}\frac{(N-1)}{N(N-2)} = O(\frac{1}{N}) \tag{96}$$

□

G.3. Proof for Theorem 5 (sensitivity for $\hat{M}_3(\hat{W}, \hat{W}, \hat{W})$)

As explained before, the whitened tensor is denoted as $\widehat{\mathcal{T}}$ for simplicity. Therefore we denote the sensitivity of $\hat{M}_3(\hat{W}, \hat{W}, \hat{W})$ as $\Delta_{\widehat{\mathcal{T}}}(D)$. Theorem 5 states that $\Delta_{\widehat{\mathcal{T}}}(D) = O(\frac{k^{3/2}}{N\sigma_k(\hat{M}_2)^{3/2}})$.

We need the following Lemma to prove Theorem 5.

Lemma 32. $\left\|\hat{W}' - \hat{W}\right\|_F \leq \frac{\sqrt{2k}\Delta_2}{\sigma_k(\hat{M}_2)\sqrt{\sigma_k(\hat{M}_2) - \Delta_2}}$

Proof. We follow an analysis similar to (Anandkumar et al., 2012). Note that the whitening matrix $\hat{W}$ is defined such that:

$$\hat{W}^T \hat{M}_{2,k} \hat{W} = I. \tag{97}$$

Analogously for the neighboring corpus,

$$\hat{W'}^T \hat{M'}_{2,k} \hat{W'} = I. \tag{98}$$

Let E_{M_2} denote the perturbation introduced to $\hat{M}_2$ by changing a single record. Because the spectral gap of the perturbation introduced by modifying a single record is small according to the condition, applying the original whitening matrix to the neighboring data base moment $\hat{M'}_2$ would lead to a rank k matrix of size $k \times k$. Therefore, $\hat{W}^T \hat{M'}_{2,k} \hat{W}$ is a rank k matrix of size $k \times k$, which can be factorized as:

$$\hat{W}^T \hat{M'}_{2,k} \hat{W} = ADA^T \tag{99}$$

where A are the singular vectors of $\hat{W}^T \hat{M'}_{2,k} \hat{W}$, and D is a diagonal matrix of the corresponding singular values of $\hat{W}^T \hat{M'}_{2,k} \hat{W}$. This also leads to $\hat{W'} = \hat{W} A D^{\frac{-1}{2}} A^T$. Using this, we observe:

$$\left\| \hat{W'} - \hat{W} \right\| = \left\| \hat{W'} - \hat{W'} A D^{\frac{1}{2}} A^T \right\| = \left\| \hat{W'} (I - A D^{\frac{1}{2}} A^T) \right\| \leq \left\| \hat{W'} \right\| \left\| I - A D^{\frac{1}{2}} A^T \right\| \tag{100}$$

Now we bound $\left\| I - A D^{\frac{1}{2}} A^T \right\|$:

$$\begin{aligned} \left\| I - A D^{\frac{1}{2}} A^T \right\| &= \left\| A^T A - \hat{W'} A D^{\frac{1}{2}} A^T \right\| = \left\| I - D^{\frac{1}{2}} \right\| && (101) \\ &\leq \left\| (I - D^{\frac{1}{2}})(I + D^{\frac{1}{2}}) \right\| \leq \| (I - D) \| && (102) \\ &= \| I - ADA^T \| = \left\| \hat{W}^T \hat{M}_{2,k} \hat{W} - \hat{W'}^T \hat{M'}_{2,k} \hat{W'} \right\| && (103) \\ &\leq \left\| \hat{W} \right\|^2 \left\| \hat{M}_{2,k} - \hat{M'}_{2,k} \right\| \leq \left\| \hat{W} \right\|^2 \| E_{M_2} \| && (104) \end{aligned}$$

We know that

$$\begin{aligned} \left\| \hat{W} \right\|^2 &\leq \frac{1}{\sigma_k(\hat{M}_2)} && (105) \\ \left\| \hat{W'} \right\| &\leq \frac{1}{\sqrt{\sigma_k(\hat{M'}_2)}} \leq \frac{1}{\sqrt{\sigma_k(\hat{M}_2) - \| E_{M_2} \|_2}} \leq \frac{1}{\sqrt{\sigma_k(\hat{M}_2) - \Delta_2}} && (106) \end{aligned}$$

Weyl's theorem was used in the last bound in Equation (106). Bounding the Frobenius norm, would result in the following:

$$\left\| \hat{W'} - \hat{W} \right\|_F \leq \sqrt{2k} \left\| \hat{W'} - \hat{W} \right\| \leq \frac{\sqrt{2k} \, \| E_{M_2} \|}{\sigma_k(\hat{M}_2) \sqrt{\sigma_k(\hat{M}_2) - \| E_{M_2} \|}} \leq \frac{\sqrt{2k} \Delta_2}{\sigma_k(\hat{M}_2) \sqrt{\sigma_k(\hat{M}_2) - \Delta_2}}, \tag{107}$$

where we have used the fact that the l_1 norm upper bounds the spectral norm of a matrix, since it upper bounds the Frobenius. □

Now we are ready to prove Theorem 5.

Proof. $\hat{M'}_3 = \hat{M}_3 + E_3$.

$$\begin{aligned} \left\| \hat{M}_3(\hat{W}, \hat{W}, \hat{W}) - \hat{M'}_3(\hat{W'}, \hat{W'}, \hat{W'}) \right\|_F &= \| \hat{M}_3(\hat{W}, \hat{W}, \hat{W}) - \hat{M}_3^{LDA}(\hat{W'}, \hat{W'}, \hat{W'}) - E_3(\hat{W'}, \hat{W'}, \hat{W'}) \|_F && (108) \\ &\leq \left\| \hat{M}_3^{LDA}(W \hat{-} W', W \hat{-} W', W \hat{-} W') \right\|_F + \left\| E_3(\hat{W'}, \hat{W'}, \hat{W'}) \right\|_F && (109) \\ &\leq \left\| \hat{M}_3 \right\|_F \left\| \hat{W} - \hat{W'} \right\|_F^3 + \| \Delta_3 \|_F \left\| \hat{W'} \right\|_F^3 && (110) \end{aligned}$$

We have used the fact that the Frobenius norm of the difference between the tensors is bounded above by the l_1 norm of the difference Δ_3. To bound the l_1 norm of $\hat{M}_3$ we use an analysis similar to calculating Δ_3. Again we note that the l_1 norm upper bounds the Frobenius norm:

$$\left\| \hat{M}_3 \right\|_F \leq \left\| \hat{M}_2 \right\|_1 = 1 + \frac{6\alpha_0}{\alpha_0 + 2} \frac{N}{N-1} + \frac{6\alpha_0^2}{(\alpha_0 + 1)(\alpha_0 + 2)} \frac{N^3}{N(N-1)(N-2)} \tag{111}$$

Combining all the expressions we get:

$$\begin{aligned}\Delta_{\widehat{\mathcal{T}}}(D) &= \left\|\hat{M}_3(\hat{W},\hat{W},\hat{W}) - \hat{M'}_3(\hat{W'},\hat{W'},\hat{W'})\right\|_F && (112)\\ &\leq (1+\frac{6\alpha_0}{\alpha_0+2}\frac{N}{N-1}+\frac{6\alpha_0^2}{(\alpha_0+1)(\alpha_0+2)}\frac{N^3}{N(N-1)(N-2)})\\ &\times\frac{(2k)^{3/2}(\Delta_2)^3}{\sigma_k(\hat{M}_2)^3(\sigma_k(\hat{M}_2)-\Delta_2)^{3/2}}+\frac{\Delta_3 k^{3/2}}{(\sigma_k(\hat{M}_2)-\Delta_2)^{3/2}} && (113)\\ &= O(\frac{k^{3/2}}{N\sigma_k(\hat{M}_2)^{3/2}}) && (114)\end{aligned}$$

We see that if N is larger than $d^{3/2}$, then $N\sigma_k(\hat{M}_2)^{3/2} \geq 1$ as $\sigma_i(\hat{M}_2)$ is in the order of $1/d$. □

G.4. Proof for Theorem 6 (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 $\Delta_{\bar{\mu}}(D) \leq O(\frac{k^2}{\gamma_s N(\sigma_k(\hat{M}_2))^{3/2}})$, where $\gamma_s = \min_{i\in[k]}\frac{\sigma_i-\sigma_{i+1}}{4}$, σ_i is the i^{th} eigenvalue of $\hat{M}_3(\hat{W},\hat{W},\hat{W})$.

Proof. The proof follows from the result of the simultaneous tensor power method (Theorem 1 in (Wang & Lu, 2017)). Replacing the original eigenvectors with those resulting from database D leads to tensor $\hat{M}_3(\hat{W},\hat{W},\hat{W})$, then the tensor resulting from corpus D' with one record changed yields $\hat{M'}_3(\hat{W'},\hat{W'},\hat{W'})$ where the spectral norm of the error is upper bounded by ϵ, if $\Delta_{\widehat{\mathcal{T}}}(D)$ is sufficiently small $\Delta_{\widehat{\mathcal{T}}}(D) \leq \frac{\gamma_s\epsilon}{2\sqrt{k}}$. Therefore we get $\left\|\bar{\mu}_i - \bar{\mu'_i}\right\|_2 \leq \frac{2\sqrt{k}\Delta_{\widehat{\mathcal{T}}}(D)}{\gamma_s}$ and $|\bar{\alpha}_i - \bar{\alpha'}_i| \leq \frac{2\sqrt{k}\Delta_{\widehat{\mathcal{T}}}(D)}{\gamma_s}$. □

G.5. Proof for Theorem 7 (sensitivity of the final output μ_i, α_i)

We now prove the sensitivity of the final output μ_i, α_i: $\Delta_\mu(D) = O(\frac{k^2\sqrt{\sigma_1(\hat{M}_2)}}{\gamma_s N\sigma_k^{3/2}(\hat{M}_2)})$.

Proof. We point out a number of things. Tensor decomposition outputs are: $\bar{\mu}_i, \bar{\alpha}_i, i \in [k]$, where, $\bar{\alpha}_i = \frac{2\sqrt{(\alpha_0+1)\alpha_0}}{(\alpha_0+2)\sqrt{\alpha_i}}$. In order to recover the desired word topic vector μ, we have to "unwhiten" to get the μ_i and α_i before whitening, i.e. $\mu_i = \frac{1}{\sqrt{\alpha_i^r}}(W^T)^\dagger\bar{\mu}_i$, where $\frac{1}{\sqrt{\alpha_i^r}} = \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\bar{\alpha}_i$. The sensitivity would be:

$$\begin{aligned}\max_{D,D'}\|\mu_i - \mu\prime_i\| &\leq \max_{D,D'}\left\{\left\|\frac{1}{\sqrt{\alpha_i^r}}(W^T)^\dagger\bar{\mu}_i - \frac{1}{\sqrt{\alpha_i^{r,'}}}(W^{T,'})^\dagger\bar{\mu}'_i\right\|_2\right\} && (115)\\ &\leq \max_{D,D'}\left\{\frac{1}{\sqrt{\alpha_i^r}}\left\|(W^T)^\dagger\right\|\|\bar{\mu}_i - \bar{\mu}'_i\| + \frac{1}{\sqrt{\alpha_i^r}}\left\|W^\dagger - (W')^\dagger\right\| + \left\|(W')^\dagger\right\||\frac{1}{\sqrt{\alpha_i^r}} - \frac{1}{\sqrt{\alpha_{i,\prime}^r}}|\right\} && (116)\end{aligned}$$

We note the following:

(i) $\max_{D,D'}|\frac{1}{\sqrt{\alpha_i^r}} - \frac{1}{\sqrt{\alpha_{i,\prime}^r}}| = \max_{D,D'}|\frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\bar{\alpha}_i - \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\bar{\alpha}'_i| \leq \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\max_{D,D'}|\bar{\alpha}_i - \bar{\alpha}'_i| \leq \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\frac{2\sqrt{k}\Delta_{\widehat{\mathcal{T}}}(D)}{\gamma_s}$, where the above follows from the simultaneous power iteration method.

(ii) $\max_{i\in[k]}\frac{1}{\sqrt{\alpha_i^r}} \leq \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\max_{i\in[k]}\bar{\alpha}_i = \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\sigma_1(\widehat{\mathcal{T}})$

(iii) $\max\left\|((W\prime)^T)\dagger\right\| \leq \sqrt{\sigma_1(\hat{M\prime}_2)} \leq \sqrt{\sigma_1(\hat{M}_2)+\Delta_2}$

(iv) Following an analysis similar to that in 32, we obtain $\left\|W^\dagger - (W')^\dagger\right\| \leq \frac{\sqrt{\sigma_1(\hat{M}_2)}}{\sigma_k(\hat{M}_2)}\Delta_2$.

Combining all of this together leads to the following

$$
\begin{aligned}
\max_{D,D'} \|\mu_i - \mu'_i\| \leq & \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}} \sigma_1(\widehat{\mathcal{T}}) \sqrt{\sigma_1(\hat{M}_2)} \frac{2\sqrt{k}\Delta_{\widehat{\mathcal{T}}}(D)}{\gamma_s} + \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}} \sigma_1(\widehat{\mathcal{T}}) \frac{\sqrt{\sigma_1(\hat{M}_2)}}{\sigma_k(\hat{M}_2)} \Delta_2 \\
& + \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}} \sqrt{\sigma_1(\hat{M}_2) + \Delta_2} \frac{2\sqrt{k}\Delta_{\widehat{\mathcal{T}}}(D)}{\gamma_s} && (117) \\
& = O(\frac{k^2\sqrt{\sigma_1(\hat{M}_2)}}{\gamma_s N \sigma_k^{3/2}(\hat{M}_2)}) && (118)
\end{aligned}
$$

□

G.6. Proof for Lemma 8

Let $\tilde{LS}$ denote the local sensitivity. We prove a slightly more general version where the construction of $\tilde{LS}$ is (ϵ_1, δ_1)-DP, and it is a valid upper bound with probability $\geq 1 - \delta_3$.

Lemma 33. *Let* LS *be the* ℓ_p *local sensitivity of a function* f *on a fixed data set. Let* $\tilde{\mathsf{LS}}$ *obeys* (ϵ_1, δ_1)*-DP and that* $\mathbb{P}[\mathsf{LS} \geq \tilde{\mathsf{LS}}] \leq \delta_3$ *(where the probability is only over the randomness in releasing* $\tilde{\mathsf{LS}}$*). Then the algorithm releases* $f(DATA) + Z(\epsilon, \delta, \tilde{\mathsf{LS}})$ *that is* $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2 + \delta_3)$*-DP, where* $Z(\epsilon_2, \delta_2, \tilde{\mathsf{LS}})$ *is any way of calibrating the noise for privacy which takes the local sensitivity as if it is a global sensitivity.*

Proof. Let x, x' be two adjacent data sets and the overall output be $O := f(\text{DATA}) + Z(\epsilon_2, \delta_2, \tilde{LS})$. Let $S_1 \subset \text{Range}(f), S_2 \subset \mathbb{R}_+$ be any measurable sets.

Let E be the measurable set of $\tilde{\mathrm{LS}}$ that represents the event that $\tilde{\mathrm{LS}} \geq \mathrm{LS}$.

$$
\begin{aligned}
& \mathbb{P}[(O, \tilde{\mathrm{LS}}) \in S_1 \times S_2 | x] && (119) \\
= & \mathbb{P}[(O, \tilde{\mathrm{LS}}) \in S_1 \times (S_2 \cap E) | x] + \mathbb{P}[(O, \tilde{\mathrm{LS}}) \in S_1 \times S_2 \cap E^c | x] && (120) \\
\leq & \mathbb{P}[(O, \tilde{\mathrm{LS}}) \in S_1 \times (S_2 \cap E) | x] + \delta_3 && (121) \\
\leq & e^{\epsilon_1 + \epsilon_2} \mathbb{P}[(O, \tilde{\mathrm{LS}}) \in S_1 \times (S_2 \cap E) | x'] + \delta_1 + \delta_2 + \delta_3 && (122) \\
\leq & e^{\epsilon_1 + \epsilon_2} \mathbb{P}[(O, \tilde{\mathrm{LS}}) \in S_1 \times S_2 | x'] + \delta_1 + \delta_2 + \delta_3 && (123)
\end{aligned}
$$

The fourth line holds due to the fact that under event the E, $\tilde{\mathrm{LS}}$ is always a valid upper bound of the local sensitivity, therefore, conditioning on the σ-field induced by $E \cap S_2$ for any S_2, O is an (ϵ_2, δ_2)-DP release. By the simple composition Theorem of (ϵ, δ)-DP (Dwork et al., 2014a)[Theorem B.1,], by taking the measurable set of interest to be $S_1 \times (S_2 \cap E)$, we have that

$$
\mathbb{P}[(O, \tilde{\mathrm{LS}}) \in S_1 \times (S_2 \cap E) | x] \leq e^{\epsilon_1 + \epsilon_2} \mathbb{P}[(O, \tilde{\mathrm{LS}}) \in S_1 \times (S_2 \cap E) | x'] + \delta_1 + \delta_2
$$

which wraps up the proof. □

The proof of Lemma 8 is a corollary which takes $\delta_1 = 0$.

G.7. Proof for Sensitivity of singular values $\sigma_k(\hat{M}_2)$ (Lemma 9)

Proof. We first prove that the global sensitivity of $\sigma_k(\hat{M}_2)$ is $1/n$. By Weyl's lemma (Stewart, 1998)[Theorem 1], for any matrix X, any i, the singular value $|\sigma_i(X) - \sigma_i(X + E)| \leq \|E\|_2$. In our case, E is coming from adding or removing one data point and we know that $\|E\|_2 \leq \|E\|_F \leq \|E\|_{1,1} \leq 2/n$, hence the bound.

Now we prove that the global sensitivity of $\gamma_s = \min_{i \in [k]} \frac{\sigma_i(\widehat{\mathcal{T}}) - \sigma_{i+1}(\widehat{\mathcal{T}})}{4}$. For any tensor $\widehat{\mathcal{T}}$, we consider a polyadic form or the so called tensor decomposition form, and denote the singular values as the amplitude of the components in the polyadic form. As shown in Section G.2, $|\sigma_i(\widehat{\mathcal{T}}) - \sigma_i(\widehat{\mathcal{T}} + \mathcal{E})| \leq \|\mathcal{E}\| \leq 1$, where $\mathcal{E}$ comes from adding or removing one data point.

□

H. Utility Proofs

Before starting the utility proofs, we point out a number of things. Tensor decomposition outputs:$\bar{\mu}_i, \bar{\alpha}_i, i \in [k]$. Where, $\bar{\alpha}_i = \frac{2\sqrt{(\alpha_0+1)\alpha_0}}{(\alpha_0+2)\sqrt{\alpha_i}}$. In order to recover the desired word topic vector μ, we have to 'reverse whiten', i.e. $\mu_i = \frac{1}{\sqrt{\alpha_i^r}}(W^T)^\dagger \bar{\mu}_i$, where $\frac{1}{\sqrt{\alpha_i^r}} = \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\bar{\alpha}_i$. We need to establish the distance between the non-differentially private output and the differentially private output, i.e. $\left\|\mu_i - \mu_i^{DP}\right\|$. This can be upper bounded similar to G.5 by the following:

$$\left\|\mu_i - \mu_i^{DP}\right\| \leq \frac{1}{\sqrt{\alpha_i^r}} \left\|(W^T)^\dagger\right\| \left\|\bar{\mu}_i - \bar{\mu}_i^{DP}\right\| + \frac{1}{\sqrt{\alpha_i^r}} \left\|W^\dagger - (W^{DP})^\dagger\right\| + \left\|(W^{DP})^\dagger\right\| \left|\frac{1}{\sqrt{\alpha_i^r}} - \frac{1}{\sqrt{\alpha_{i,DP}^r}}\right| \tag{124}$$

For this we frequently need to bound the following: $\left\|\bar{\mu}_i - \bar{\mu}_i^{DP}\right\|$, $\left\|W^\dagger - (W^{DP})^\dagger\right\|$, $\left\|(W^{DP})^\dagger\right\|$, $|\frac{1}{\sqrt{\alpha_i^r}} - \frac{1}{\sqrt{\alpha_{i,DP}^r}}|$, and $|\bar{\alpha}_i - \bar{\alpha}_i^{DP}|$.

We point out the following facts before preceding.

Fact 34. $|\frac{1}{\sqrt{\alpha_i^r}} - \frac{1}{\sqrt{\alpha_{i,DP}^r}}| \leq |\frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\bar{\alpha}_i - \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\bar{\alpha}_i^{DP}| \leq \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}|\bar{\alpha}_i - \bar{\alpha}_i^{DP}|$.

Fact 35. $\left\|(W^T)^\dagger\right\| \leq \sqrt{\sigma_1(\hat{M}_2)}$.

Fact 36. $\frac{1}{\sqrt{\alpha_i^r}} = \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\bar{\alpha}_i \leq \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\sigma_1(\widehat{\mathcal{T}})$.

H.1. Perturbation on $\hat{M}_2$, $\hat{M}_3$ Config. 1 (e_3, e_4, e_8): Proof for Theorem 10

Similar to the perturbation on (e_6, e_8). We have that

$$\left\|W^\dagger - (W^{DP})^\dagger\right\| \leq \frac{\sqrt{\sigma_1(\hat{M}_2)}\,\|E_{8,G}\|}{\sigma_k(\hat{M}_2)} \tag{125}$$

$$\left\|(W^{DP})^\dagger\right\| \leq \sqrt{\sigma_1(\hat{M}_2) + \|E_{8,G}\|} \tag{126}$$

Now the perturbed tensor can be represented as $\hat{M}_3^{DP} = \hat{M}_3 + E_{3,G}$, where $E_{3,G}$ is symmetric Gaussian noise that has been added to the original tensor. Similar to the sensitivity analysis for the whitened tensor, we have that the error Φ can be bounded as follows:

$$\|\Phi\|_2 = \left\|\hat{M}_3(\hat{W}, \hat{W}, \hat{W}) - \hat{M}_3^{DP}(W^{DP}, W^{DP}, W^{DP})\right\|_2 \tag{127}$$

$$\leq \left\|\hat{M}_3\right\| \left\|W - W^{DP}\right\|^3 + \|E_{3,G}\| \left\|W^{DP}\right\| \tag{128}$$

Following an analysis similar to bounding $\left\|W^\dagger - (W^{DP})^\dagger\right\|$, we get that $\left\|W^\dagger - (W^{DP})^\dagger\right\| \leq \frac{\|E_{8,G}\|}{\sigma_k(\hat{M}_2)\sqrt{\frac{\sigma_k(\hat{M}_2)}{2}}}$. According to 43 we have that with high probability $\|E_{3,G}\| = O(\sqrt{d}\Delta_3\tau_{\epsilon_3,\delta_3})$. We note the following $\left\|\bar{\mu}_i - \bar{\mu}_i^{DP}\right\|_2 \leq \frac{2\sqrt{k}\|\Phi\|}{\gamma_s}$ using the simultaneous power iteration of (Wang & Lu, 2017). Similarly we have $|\bar{\alpha}_i - \bar{\alpha}_i^{DP}| \leq \frac{2\sqrt{k}\|\Phi\|}{\gamma_s}$ and that $|\frac{1}{\sqrt{\alpha_i^r}} - \frac{1}{\sqrt{\alpha_{i,DP}^r}}| \leq \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\frac{2\sqrt{k}\|\Phi\|}{\gamma_s}$. This leads to $\left\|\mu_i - \mu_i^{DP}\right\|_2 \leq \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\sigma_1(\widehat{\mathcal{T}})\sqrt{\sigma_1(\hat{M}_2)}\frac{2\sqrt{k}\|\Phi\|}{\gamma_s} + \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\sigma_1(\widehat{\mathcal{T}})\frac{\sqrt{\sigma_1(\hat{M}_2)}}{\sigma_k(\hat{M}_2)}\|E_{8,G}\| + \sqrt{\sigma_1(\hat{M}_2) + \|E_{8,G}\|}\frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\frac{2\sqrt{k}\|\Phi\|}{\gamma_s}$.

Based on the bound on $\|\Phi\|$ we have with high probability $\left\|\mu_i - \mu_i^{DP}\right\|_2 = O(\frac{\sqrt{\sigma_1(\hat{M}_2)k}}{\gamma_s}((\frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_4,\delta_4})^3 + \frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_3,\delta_3}) + \frac{\sqrt{\sigma_1(\hat{M}_2)d}}{\sigma_k(\hat{M}_2)N}\tau_{\epsilon_8,\delta_8} + \sqrt{\sigma_1(\hat{M}_2) + \frac{\sqrt{d}}{N}\tau_{\epsilon_8,\delta_8}}\frac{\sqrt{k}}{\gamma_s}\left[(\frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)}\tau_{\epsilon_4,\delta_4})^3 + \frac{\sqrt{d}}{N\sigma_k(\hat{M}_2)^{3/2}}\tau_{\epsilon_3,\delta_3}\right])$.

H.2. Perturbation on $\widehat{\mathcal{T}}$ and $\hat{M}_2$ Config. 2(e_6, e_8): Proof for Theorem 12

This configuration has two properties: the noise level introduced is low because the whitening step reduces the tensor dimension from $\hat{M}_3 \in \mathbb{R}^{d\times d\times d}$ to $\widehat{\mathcal{T}} = \hat{M}_3(\hat{W}, \hat{W}, \hat{W}) \in \mathbb{R}^{k\times k\times k}$. However, even though the dimension of the tensor is reduced, unless the whitening tensor (resulting from eigendecomposition over $\hat{M}_2$) is stable, the sensitivity of the whitened tensor is not necessarily low.

Note that the sensitivity of $\hat{M}_2$ falls with $\frac{1}{N}$ (Theorem 4). Therefore, we expect the sensitivity of $\hat{M}_3(\hat{W}, \hat{W}, \hat{W})$ to drop with an increasing number of records. As Theorem 5 states, $\Delta_{\widehat{\mathcal{T}}}(D) = O(\frac{k^{3/2}}{N\sigma_k^{3/2}(\hat{M}_2)})$, if $\Delta_2 \leq \sigma_k(\hat{M}_2) - \sigma_{k+1}(\hat{M}_2)$. Thus, given the spectral gap requirement, the sensitivity of the whitened tensor is $\Delta_{\widehat{\mathcal{T}}}(D)$.

$\hat{M}_2$ is used to generate both the whitening and unwhitening matrix, and unlike input perturbation, the sensitivity over $\hat{M}_2$ and $\hat{M}_3$ falls as the dataset size increases (Theorem 4). However, an issue with this configuration is that adding noise to $\hat{M}_3$ leads to higher noise build up prior to the tensor decomposition. Note that by (43) w.h.p the norm of the error is $O(\sqrt{d}\sigma)$, with σ being the variance of the noise (this bound would be $\sqrt{k}\sigma$ if the noise is added to a symmetric tensor of size k). Tensor decomposition methods, in particular (Wang & Lu, 2017) require the spectral norm of the perturbation to the tensor to be lower than a certain threshold. Following arguments similar to (Wang & Anandkumar, 2016), the spectral norm of the error is $O(\frac{\sqrt{d}}{N\epsilon_3})$ and should be below $\frac{\sqrt{k}}{\gamma_s \sigma_k(\widehat{\mathcal{T}})}$. Thus ϵ_3 should satisfy $\epsilon_3 = \Omega(\frac{\sqrt{kd}}{\gamma_s \sigma_k(\widehat{\mathcal{T}})N})$ to establish utility guarantees for tensor decomposition. Following similar arguments, this time using the bound on the spectral norm of the noisy matrices, to guarantee utility, the differentially private whitening W and pseudo-inverse $W^\dagger$ should be close to their non-differentially private values, which requires both ϵ_4 and ϵ_8 to be $\Omega(\frac{\sqrt{d}}{(\sigma_k(\hat{M}_2) - \sigma_{k+1}(\hat{M}_2)N)})$. Although, the privacy parameters have a lower bound of $\sqrt{d}$, the bound also falls with $\frac{1}{N}$.

The spectral norm of the noise added to $\hat{M}_2$ can be bounded by 42 to be $O(\frac{\sqrt{d}}{N}\tau_{\epsilon_8,\delta_8})$ with high probability. Now, if we have $N = \Omega(\frac{\sqrt{d}\tau_{\epsilon_8,\delta_8}}{\sigma_k(\hat{M}_2) - \sigma_{k+1}(\hat{M}_2)})$, then with w.h.p we have that $\|E_{8,G}\| \leq \frac{\sigma_k(\hat{M}_2) - \sigma_{k+1}(\hat{M}_2)}{2}$, where $\|E_{8,G}\|$ is the spectral norm of the Gaussian matrix. This condition enables us to bound $\left\|W^\dagger - (W^{DP})^\dagger\right\|$, in a manner similar to establishing the bounds between $\|W - W'\|$ in 32. Following a similar analysis, given that

$$W^T(\hat{M}_2)_k W = I, \tag{129}$$

$$W^{T,DP}(\hat{M}_2 + E_{8,G})_k W^{DP} = I, \tag{130}$$

$$W^T(\hat{M}_2 + E_{8,G})_k W = ADA^T, \tag{131}$$

we have that $\left\|W^\dagger - (W^{DP})^\dagger\right\| \leq \left\|W^\dagger\right\| \|I - D\|$. We know that $\left\|W^\dagger\right\| \leq \frac{1}{\sqrt{\sigma_k(\hat{M}_2)}}$ and $\|I - D\|$ can be bounded as follows:

$$\|I - D\| \leq \left\|I - ADA^T\right\| \leq \left\|W^T(\hat{M}_2)_k W - W^T(\hat{M}_2 + E_{8,G})_k W\right\| \tag{132}$$

$$\leq \|W\|^2 \left\|(\hat{M}_2)_k - (\hat{M}_2 + E_{8,G})_k\right\| \leq \|W\|^2 \|E_{8,G}\| \leq \frac{\|E_{8,G}\|}{\sigma_k(\hat{M}_2)} \tag{133}$$

This leads to $\left\|W^\dagger - (W^{DP})^\dagger\right\| \leq \frac{\sqrt{\sigma_1(\hat{M}_2)}\|E_{8,G}\|}{\sigma_k(\hat{M}_2)}$.

Moreover, it is immediate by Weyl's theorem that $\left\|(W^{DP})^\dagger\right\| \leq \sqrt{\sigma_1(\hat{M}_2 + E_{8,G})} \leq \sqrt{\sigma_1(\hat{M}_2) + \|E_{8,G}\|}$.

Finally, by the results of simultaneous power iteration (with an argument similar to Theorem 6), if N is sufficiently large, we have that $\left\|\bar{\mu}_i - \bar{\mu}_i^{DP}\right\| \leq \frac{2\sqrt{k}\|E_{6,G}\|}{\gamma_s}$ where $E_{6,G}$ is the Gaussian tensor added to the whitened tensor $\Delta_{\widehat{\mathcal{T}}}(D)$. An identical bound is established for the eigenvalues, i.e. $|\bar{\alpha}_i - \bar{\alpha}_i^{DP}| \leq \frac{2\sqrt{k}\|E_{6,G}\|}{\gamma_s}$.

Now we can state the utility:

$$\begin{aligned} \left\|\mu_i - \mu_i^{DP}\right\| \leq{}& \frac{(\alpha_0 + 2)}{2\sqrt{(\alpha_0 + 1)\alpha_0}}\sigma_1(\widehat{\mathcal{T}})\sqrt{\sigma_1(\hat{M}_2)}\frac{2\sqrt{k}\left\|E_{6,G}\right\|}{\gamma_s} + \frac{(\alpha_0 + 2)}{2\sqrt{(\alpha_0 + 1)\alpha_0}}\sigma_1(\widehat{\mathcal{T}})\frac{\sqrt{\sigma_1(\hat{M}_2)}}{\sigma_k(\hat{M}_2)}\left\|E_{8,G}\right\| \\ &+ \frac{(\alpha_0 + 2)}{2\sqrt{(\alpha_0 + 1)\alpha_0}}\sqrt{\sigma_1(\hat{M}_2) + \|E_{8,G}\|}\frac{2\sqrt{k}\left\|E_{6,G}\right\|}{\gamma_s} \end{aligned} \tag{134}$$

We note that w.h.p we have the following bounds on spectral norms of noisy Gaussian matrix and noisy Gaussian tensor. In particular, $\|E_{6,G}\| = O(\frac{k^2}{N\tilde{\sigma}_k^{3/2}}\tau_{\epsilon_6,\delta_6})$ and $\|E_{8,G}\| = O(\frac{\sqrt{d}}{N}\tau_{\epsilon_8,\delta_8})$. This leads to the following utility

$$\|\mu_i - \mu_i^{DP}\| = O(\frac{\sqrt{\sigma_1(\hat{M}_2)k^{2.5}}}{\gamma_s N\tilde{\sigma}_k^{3/2}}\tau_{\epsilon_6,\delta_6} + \frac{\sqrt{\sigma_1(\hat{M}_2)d}}{\sigma_k(\hat{M}_2)N}\tau_{\epsilon_8,\delta_8} + \sqrt{\sigma_1(\hat{M}_2) + \frac{\sqrt{d}}{N}\tau_{\epsilon_8,\delta_8}}\frac{k^{2.5}\tau_{\epsilon_6,\delta_6}}{\gamma_s N\tilde{\sigma}_k^{3/2}}). \tag{135}$$

H.3. Perturbation on the output of tensor decomposition $\bar{\mu}_i$,$\bar{\alpha}_i$ and $\hat{M}_2$ Config. 3 (e_7, e_8): Proof for Theorem 14

This configuration shares edge 8 with the previous. This enables us to borrow the same bounds for the pseudo-inverse $W^\dagger$. Specifically, we have:

$$\|W^\dagger - (W^{DP})^\dagger\| \leq \frac{\sqrt{\sigma_1(\hat{M}_2)}\,\|E_{8,G}\|}{\sigma_k(\hat{M}_2)} \tag{136}$$

$$\|(W^{DP})^\dagger\| \leq \sqrt{\sigma_1(\hat{M}_2) + \|E_{8,G}\|} \tag{137}$$

In this method, noise is added directly to the eigenvectors and eigenvalues resulting from the tensor decomposition. Therefore, we have:

$$\bar{\mu}_i^{DP} = \bar{\mu}_i + Y, \qquad Y \sim \mathcal{N}(0, \Delta_{\epsilon,\delta}^2 I_k) \tag{138}$$

$$\bar{\alpha}_i^{DP} = \bar{\alpha}_i + n_i, \qquad n_i \sim \mathcal{N}(0, \Delta_{\epsilon,\delta}^2) \tag{139}$$

where $\Delta_{\epsilon,\delta} = \frac{\sqrt{2k}\Delta_{\hat{\mathcal{T}}}(D)}{\gamma_s}\tau_{\epsilon_7,\delta_7}$ with $\tau_{\epsilon_7,\delta_7} = \frac{\sqrt{2ln(1.25/\delta_7)}}{\epsilon_7}$. This leads to the following bound:

$$\begin{aligned}\|\mu_i - \mu_i^{DP}\| \leq & \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\sigma_1(\widehat{\mathcal{T}})\sqrt{\sigma_1(\hat{M}_2)}\,\|Y\| + \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\sigma_1(\widehat{\mathcal{T}})\frac{\sqrt{\sigma_1(\hat{M}_2)}}{\sigma_k(\hat{M}_2)}\,\|E_{8,G}\| \\ & + \frac{(\alpha_0+2)}{2\sqrt{(\alpha_0+1)\alpha_0}}\sqrt{\sigma_1(\hat{M}_2) + \|E_{8,G}\|}|n_i|\end{aligned} \tag{140}$$

As before w.h.p $\|E_{6,G}\| = O(\frac{\sqrt{d}}{N}\tau_{\epsilon_6,\delta_6})$.The following bounds hold on $\|Y\|$ and $|n_i|$, because they are a Gaussian vector and variable. In particular, w.h.p. $\|Y\| = O(\frac{k^{5/2}}{N\tilde{\sigma}_k^{3/2}\tilde{\gamma_s}}\tau_{\epsilon_7,\delta_7})$ and $|n_i| = O(\frac{k^2}{N\tilde{\sigma}_k^{3/2}\tilde{\gamma_s}}\tau_{\epsilon_7,\delta_7})$. This leads to the following utility: $O(\frac{\sqrt{\sigma_1(\hat{M}_2)k^{2.5}}}{\tilde{\gamma_s}N\tilde{\sigma}_k^{3/2}}\tau_{\epsilon_7,\delta_7} + \frac{\sqrt{\sigma_1(\hat{M}_2)d}}{\sigma_k(\hat{M}_2)N}\tau_{\epsilon_8,\delta_8} + \sqrt{\sigma_1(\hat{M}_2) + \frac{\sqrt{d}}{N}\tau_{\epsilon_8,\delta_8}}\frac{k^2\tau_{\epsilon_7,\delta_7}}{\tilde{\gamma_s}N\tilde{\sigma}_k^{3/2}})$.

H.4. Perturbation on the final output μ_i, α_i Config. 4 ($e9$): Proof for Theorem 16

In this configuration, we add noise proportional to the output's sensitive

$$\mu_i^{DP} = \mu_i + Z, \text{ where } Z \sim \mathcal{N}(0, \Delta_{\epsilon,\delta}^2 I_k) \tag{141}$$

where $\Delta_{\epsilon,\delta} = \Delta_\mu(D)\tau_{\epsilon_9,\delta_9}$, with $\tau_{\epsilon_9,\delta_9} = \frac{\sqrt{2ln(1.25/\delta_9)}}{\epsilon_9}$. Similar to the previous analysis, since Z is Gaussian, then w.h.p. $\|Z\| = O(\frac{\sqrt{d\sigma_1(\hat{M}_2)}k^2}{N\tilde{\gamma_s}\tilde{\sigma}_k^{3/2}})$. We have the utility $O(\frac{\sqrt{\sigma_1(\hat{M}_2)d}k^2}{N\tilde{\gamma_s}\tilde{\sigma}_k^{3/2}}\tau_{\epsilon_9,\delta_9})$.

I. Some Useful Identities and Theorems

Identity 37 (Square of Sum)**.**

$$\Big(\sum_{i=1}^{N} a_i\Big)^2 = \sum_{i=1}^{N} a_i^2 + \sum_{\substack{i=1,j=1\\ i\neq j}}^{N,N} a_i a_j \tag{142}$$

Identity 38 (Cube of Sum)**.**

$$\Big(\sum_{i=1}^{N} a_i\Big)^3 = \sum_{i=1}^{N} a_i^3 + 3 \sum_{\substack{i=1,j=1\\ i\neq j}}^{N,N} a_i^2 a_j + \sum_{\substack{i=1,j=1,k=1\\ i\neq j, j\neq k, k\neq i}}^{N,N,N} a_j a_j a_k \tag{143}$$

Theorem 39 (Weyl's theorem; Theorem 4.11, p. 204 in (Stewart, 1990))**.** . *Let A, E be given $m \times n$ matrices with $m \geq n$, then*

$$\max_{i\in[n]} |\sigma_i(A) - \sigma_i(A+E)| \leq \|E\|_2 \tag{144}$$

Theorem 40 (Bound on the norm of a Gaussian Random Variable)**.** *Let Z be a Gaussian $\mathcal{N}(0,\sigma)$. Then $\mathbb{P}[|Z| \leq t] \geq 1 - 2e^{\frac{-t^2}{2\sigma^2}}$ for all $t > 0$' or alternatively, $\mathbb{P}[|Z| > \sigma\sqrt{2\log(1/\delta)}] \leq \delta$ for all $0 < \delta \leq 1$.*

Theorem 41 (Bound on the norm of a Gaussian Vector)**.** *Let $Y \sim \mathcal{N}(0, \sigma I_k)$, then $\mathbb{P}[\|Y\|_2^2 \geq \sigma^2(k + 2\sqrt{kt} + 2t)] \leq e^{-t}$.*

Proof. The proof is immediate from Theorem 2.1 in (Hsu et al., 2012) with $A = I, \mu = 0$. □

Theorem 42 (Bound on the spectral norm of a Gaussian Matrix (Tao, 2012))**.** *Let $E \in \mathbb{R}^{d\times d}$ be a symmetric Gaussian matrix with elements sampled iid from $\mathcal{N}(0,\sigma)$, then $\mathbb{P}[\|E\|_2 = O(\sqrt{d}\sigma)] \geq 1 - negl(d)$.*

Theorem 43 (Bound on the spectral norm of a Gaussian Tensor (Tomioka & Suzuki, 2014))**.** *Let E be a K^{th} order tensor with each $E_{i_1,\ldots,i_K}$ be sampled i.i.d. from a Gaussian $\mathcal{N}(0,\sigma)$, then $\mathbb{P}[\|E\|_2 \leq \sqrt{8\sigma^2(\sum_{i=1}^{K} d_i)\ln(2K/K_0) + \ln(2/\delta)}] \geq 1 - \delta$, where $K_0 = \ln(3/2)$. Note by extension the bound also holds if the tensor is symmetric as well.*

Lemma 44 (Laplace tail bound)**.** *Let Z be drawn from a Laplace distribution with density $\frac{1}{2b}e^{-\frac{|z|}{b}}$, then $\mathbb{P}(Z \geq t) = \frac{1}{2}e^{-\frac{t}{b}}$ for all $t > 0$, or equivalently $Z \leq b\log(1/(2\delta))$ with probability at least $1 - \delta$ for all $0 < \delta \leq 1$.*