The equivalence of Single-Topic and LDA topic reconstruction

Flavio Chierichetti*
Sapienza University
Rome, Italy
flavio@di.uniroma1.it

Alessandro Panconesi†
Sapienza University
Rome, Italy
ale@di.uniroma1.it

Andrea Vattani
Spiketrap
San Francisco, CA, USA
avattani@cs.ucsd.edu

Abstract

In this paper we show the equivalence between two very different document generative models for the task of topic reconstruction. The first is the very influential Latent Dirichlet Allocation (LDA) model while the second is the seemingly much simpler Single Topic model. This is achieved via a reduction between these two generative models that shows that the set of latent topics is identifiable under LDA if and only if it is identifiable under Single Topic.

We then explore some of the consequences of this equivalence by giving tight impossibility results for LDA topic reconstruction. When documents are generated under LDA, the set of latent topics is identifiable if and only if the length of the documents is at least the number of topics. To our knowledge, these are the first impossibility results for LDA topic reconstruction.

*Supported in part by the ERC Starting Grant DMAP 680153, by a Google Focused Research Award, and by the SIR Grant RBSI14Q743.

†Supported in part by the ERC Starting Grant DMAP 680153, and by a Google Focused Research Award.

1 Introduction

This paper is about the problem of topic reconstruction in the context of Latent Dirichlet Allocation (henceforth LDA). Loosely speaking, the goal of topic reconstruction, sometimes also called topic identifiability, is to algorithmically reconstruct the topics of a given a corpus of documents. LDA is a very influential paradigm for topic reconstruction that was introduced by Blei et al [9, 12]. The main motivation behind it was the need to confront the typical situation in which a document touches upon many topics at once. As such, LDA appears to be more versatile and realistic than previous models such as Single Topic Allocation (henceforth STA), another popular model that assumes documents to be mono-topic (see, for instance, [20, 27]). The main result of this paper is to show that LDA is no more general than STA in the following sense: topic reconstruction is possible under LDA if and only if it is possible under STA. This equivalence is established by exhibiting a reduction from LDA to STA, and viceversa. The equivalence is computational: if we have an efficient algorithm for STA topic reconstruction we can compose it with our reduction to obtain an efficient algorithm for LDA topic reconstruction, and viceversa. This is somewhat surprising, since at first blush LDA appears to be a much richer paradigm than STA.

In order to explain our results with a sufficient degree of technical accuracy, let us briefly review the two generative models. There is a fixed vocabulary $\mathcal{V} = [m]$ and a fixed set of K topics $\mathcal{T} := \{p_1, \ldots, p_K\}$. Each topic is a probability distribution p_k over the words of the vocabulary, *i.e.* for every topic index k, $\sum_{w \in \mathcal{V}} p_k(w) = 1$. Under LDA , to generate a document we first pick a probability distribution $\theta := (\theta_1, \ldots, \theta_K)$ over the topics, the so called *admixture*. LDA specifies the admixture to be drawn from a Dirichlet distribution. It would certainly make sense to consider other types of distributions, but in this paper we will confine ourselves to the original LDA formulation. To select admixtures, we will consider only the case of symmetric Dirichlet distributions, that are characterized by a single parameter α. The symmetric case is usually assumed in practical applications and it is arguably the most interesting one. Under LDA, documents are generated one word at the time in sequence, by first picking a topic p_k with probability θ_k and then a word $w \in \mathcal{V}$ with probability $p_k(w)$. A more concise way to describe this process is to define a "cocktail" topic $p_\theta := \sum_{k=1}^{K} \theta_k p_k$ and use it to pick ℓ words $w_i \in \mathcal{V}, i \in [\ell]$, with replacement, each with probability $p_\theta(w_i)$. To generate another document a new admixture θ is picked, and so on.

The above description does not specify how topics are chosen. Under the original LDA definition, topics too are randomly drawn from a Dirichlet distribution, but it is useful to consider other models, as well. In particular, it is certainly of interest to consider the case when topics are chosen adversarially [7]. Loosely speaking, the problem of LDA topic reconstruction is, given a corpus of documents generated according to the LDA mechanism, reconstruct the hidden topics.

It is apparent that LDA generates documents that are inherently multi-topic. This is in contrast with single-topic models. In this paper we will consider a particularly simple single-topic model that we call *Uniform Single Topic Allocation*, or USTA. In USTA, we have K topics over a vocabulary $\mathcal{V}$ in the background, selected in some fashion. To generate a document, first a topic is selected with uniform probability $1/K$, and then all the words of the document are drawn from this topic. To generate another document, a topic is again selected uniformly at random, and so on.

USTA generates mono-topic documents and it appears to be much more rudimentary than LDA, which makes the main result of this paper surprising. Namely, there is a reduction between LDA and USTA showing that topic reconstruction in the two models is algorithmically the same problem. More precisely, let $\mathcal{D}_\ell^{\mathcal{T},\alpha}$ be the distribution over document of length ℓ induced by LDA when the set of topics is $\mathcal{T}$ and the Dirichlet parameter is α. Similarly, let $\mathcal{S}_\ell^{\mathcal{T},\alpha}$ be the distribution over document of length ℓ induced by USTA with the same set of topics $\mathcal{T}$. We give an algorithm such that, given α and $\mathcal{D}_{\mathcal{T},\alpha}^{\ell}$, it outputs $\mathcal{S}_{\mathcal{T},\alpha}^{\ell}$. And viceversa, given α and $\mathcal{S}_{\mathcal{T},\alpha}^{\ell}$, it outputs $\mathcal{D}_{\mathcal{T},\alpha}^{\ell}$.

While we believe this equivalence to be an interesting conceptual contribution in its own right, in this paper we also explore some of its rather stark consequences for LDA topic reconstruction. We show that if the length of the documents ℓ is less than the number of topics K, then LDA topic reconstruction is impossible. To the best of our knowledge these are the first impossibility results that have been proven about LDA topic reconstruction — that is, for topic reconstruction when documents are generated by a Dirichlet admixture.

The impossibility is established as follows. We show that, except for a set of measure zero, no matter how we pick a set $\mathcal{T}$ of K topics (*i.e.* with probability 1) on a vocabulary of $m = 2$ words, we can always find another set $\mathcal{T}'$ of K topics (in fact, uncountably many such sets) such that they induce exactly the same distribution on uniformly single-topic documents of length $\ell < K$. By the above reduction, it follows that topic identifiability under the LDA framework is impossible unless $\ell \geq K$, and this bound is tight. This result holds for any "reasonable distribution" with which topics are chosen, and in particular it holds for Dirichlet-sampled topics which is the original LDA formulation. For the case of adversarial topics, we extend the above impossibility result to vocabularies of any number of words $m \geq 2$.

A look at the rich literature for LDA topic reconstruction puts these impossibility results in perspective. As a general rule, exact algorithms for LDA topic reconstruction invariably make additional assumptions on top of the LDA framework. For instance, Arora et al. [6, 7] introduce the concept of p-separable topics. A set of topics is p-separable if, for each topic T, there is a word w_T such that *(i)* T assigns to w_T probability at least p, and *(ii)* every other topic T' assigns it probability zero. Under these assumptions algorithms with provable guarantee for LDA topic reconstruction are given. In a similar vein, the authors of [2] assume the topics by word matrix to be full rank. Similar assumptions are used even in the single-topic setting [4, 21].

Our impossibility results give an indication of why such choices are useful— without them LDA topic reconstruction is impossible, unless one explicitly assumes (i) that documents to be long enough, (ii) that the number of words in the vocabulary is large enough, and/or (iii) that topics satisfy certain properties (e.g., that they induce a full-rank matrix). Recently, several authors [1, 13, 15, 18, 22, 24, 26] have reported that topic identifiability within the LDA framework appears to be especially challenging when the corpus consists of short documents, as it is the case of tweets and texting, to cite a couple of familiar examples. Our results show that there are fundamental reasons for this difficulty.

The paper is organized as follows. In Section 2 we discuss the most relevant literature. In Section 3 we give the necessary technical preliminaries. Section 4 presents the equivalence between LDA and the uniformly single-topic scenario. Building on the reduction of the previous section, Sections 5 and 6 presents lower and upper bounds for topic identifiability. Finally, in Section 7, we study an adversarial construction for the case of non-Uniform Single Topic reconstruction.

2 Related Work

The literature on topic reconstruction and LDA in particular is rather huge and we limit ourselves to what seems to be most relevant to the aims of our paper.

After the introduction of LDA [9] in 2003, considerable effort has been invested in circumventing its posterior inference intractability [7] by using approximate local-search and inference techniques, such as Gibbs sampling [12] and variational inference [9, 23]. These algorithms aim to optimize a maximum likelihood objective, and while efficient, have no provable guarantees.

A body of work [6, 7, 16] applies matrix-factorization techniques, relying on the fact that the observed word-document matrix can be viewed as the result of a factorization model. Provable

algorithms have been developed in this respect, but they require unique matrix factors (that is, identifiability) in order to "invert" the model. By and large, these algorithms require a separable non-negative matrix factorization model [10], and achieve this by assuming that every topic has an "anchor word" that only appears in that particular topic [6,7], or similar separability conditions [16].

Another line of work frames topic reconstruction into a tensor factorization problem and uses third and higher-order statistical moments to recover the topics [2, 3, 5]. In this framework, the model becomes a Tucker model which is in general not identifiable. Identifiability is then obtained using assumptions on the rank, or the singular values, of the matrices (or tensors) supporting the model. Essentially, this translates to assuming either a PARAFAC model [2] (that is, topics are uncorrelated), or sparsity of the topics probability mass functions [5].

To summarize, theoretical work looking for algorithms with provable guarantees has relied on additional assumptions on top of the basic LDA framework, while in this work we study LDA in its original form, without any additional assumptions. While past literature focused mainly on designing identifiable models and efficient heuristics, the identifiability of topic models is still poorly understood in terms of basic properties of the model. To best of our knowledge, this paper is the first attempt to relate the identifiability of LDA (and other topic-models) to a key parameter: the length of the observed documents.

3 Preliminaries

A topic is represented as a categorical distribution p over a vocabulary $\mathcal{V}$ of m words, where $p(w) \geq 0$ corresponds to the probability of word $w \in \mathcal{V}$ for the topic, and $\sum_{w \in \mathcal{V}} p(w) = 1$.

Given a hidden set $\mathcal{T}$ of K topics and observations in the form of words collected into documents, the general problem is that of recovering (or learning) the hidden topics. A *topic model* is a parametric statistical model prescribing how the words in the documents are generated from $\mathcal{T}$, thus restricting restricts the space of possible solutions to those compatible with the model.

The core question we study in this paper is that of *identifiability*: when and whether the topics can be *uniquely* recovered given the observations. In particular, we explore the identifiability question in terms of the length ℓ of the observed documents and provide tight bounds for single-topic models and LDA.

Topic models can broadly categorized into those that allow documents to cover multiple topics and those that do not.

Single-topic models Under these models[1], each document is generated by first picking a topic p_i from some distribution over topics, and then generating ℓ words independently from Multinomial(p_i). In general, there will be some distribution $\theta \in \mathbf{R}_+^K$ that will be used to randomly pick the topics.

When θ is arbitrary and hidden, the problem is equivalent to learning a mixture of multinomials. We will refer to the case of (known) $\theta_i = 1/K$ $(1 \leq i \leq K)$ as a *uniform single-topic* model.

Topic admixture models Under these models, the observed documents are generated by an *admixture*[2] over hidden topics. In particular, a document is generated by sampling *i.i.d.* the words

[1]The *single-topic model* is also referred to as *mixture of unigrams model* [20] or *K-cluster model* (each document falls into one of K clusters).

[2]Topic models that allow a document to cover multiple topics are usually referred to as *admixtures* (in chemistry, a mixture is a composition of one or more substances, and an admixture a composition of one or more mixtures). Notice that a collection of documents generated by a single-topic model does cover a multitude of topics. In an admixture, each single document is generated by a mixture of topics.

from a mixture of topics Multinomial($\sum_{i=1}^{K} \theta_i p_i$), with $\theta_i \geq 0$ and $\sum_{i=1}^{K} \theta_i = 1$. Under the PLSA model [14], no assumption is made on the mixture weights θ, causing the number of parameters in the model to grow linearly with the size of the corpus. In the LDA model [9], a mixture weight vector θ is sampled independently for each document from a Dirichlet distribution $\mathrm{Dir}_K(\bar{\alpha})$, $\bar{\alpha} \in \mathbf{R}_+^K$. In this paper we only consider the typical case of a symmetric Dirichlet distribution[3] $\mathrm{Dir}_K(\alpha) = \mathrm{Dir}((\alpha, \ldots, \alpha))$, $\alpha \in \mathbf{R}_+$. In the fully-Bayesian version of LDA, the hidden topics are not adversarial but are sampled *i.i.d.* from a symmetric Dirichlet distribution $\mathrm{Dir}_m(\beta)$, $\beta \in \mathbf{R}_+$.

3.1 Notation

All topic models considered in this paper can be framed within the following framework. In each model, *(i)* there exists some (unknown) set of topics $\mathcal{T}$ over a vocabulary $\mathcal{V}$ of m words; *(ii)* whenever a document is to be generated, the model samples independently a categorical distribution over words $p = (p(1), \ldots, p(m))$ for that document by applying some random function to $\mathcal{T}$. For a model $\mathcal{A}$, we will write $p \sim \mathcal{A}$ to denote that p is randomly sampled from $\mathcal{A}$. Finally, *(iii)* the document is generated by sampling words independently from p, and p is forgotten. We call such an $\mathcal{A}$ a *bag-of-words* model. Observe that the probability given by such a model $\mathcal{A}$ to a document d of length ℓ, containing the word $d(j) \in [m]$ in its jth position, $\forall j \in [\ell]$, is equal to

$$E_{(p(1),\ldots,p(m))\sim\mathcal{A}} \left[\prod_{j=1}^{\ell} p(d(j)) \right].$$

We use $\mathcal{A}_\ell$ to denote the probability distribution induced by model $\mathcal{A}$ on the documents in $\mathcal{V}^\ell$ that is, on the sequences of ℓ words from $\mathcal{V} = [m]$. We write $\mathcal{A}_\ell(d) := \Pr_{X\sim\mathcal{A}_\ell}(X = d)$ to denote the probability of a specific document in $d \in \mathcal{V}^\ell$. By the bag-of-words property, it follows that if d contains exactly c_1 words 1, c_2 words 2, ..., and c_m words m, then its probability is a function of the vector $(c_1, c_2, \ldots, c_m)$. As a shorthand, we will then use $\mathcal{A}_\ell(c_1, \ldots, c_m)$, with $\sum_{i=1}^{m} c_i = \ell$, to denote the probability of a *given*[4] document $d \in \mathcal{V}^\ell$ that contains exactly c_1 words 1, c_2 words 2, ..., and c_m words m.

Given a set $\mathcal{T}$ of K topics, we denote by $\mathcal{S}^{\mathcal{T}}$ the uniform single-topic model on $\mathcal{T}$ (with document distribution $\mathcal{S}_\ell^{\mathcal{T}}$, for document length ℓ). Similarly, we use $\mathcal{D}^{\mathcal{T},\alpha}$ and $\mathcal{D}_\ell^{\mathcal{T},\alpha}$ to denote the LDA model on $\mathcal{T}$ that uses a Dirichlet of parameter α to mix the topics.

We let $\mathbf{N}$ be the set of the non-negative integers, $\mathbf{R}$ be the set of real numbers, and $\mathbf{R}_+$ be the set of positive real numbers. For any $n \in \mathbf{N}$, we define $[n] := \{1, 2, \ldots, n\}$. For a boolean predicate B, we use $[B]$ to denote the indicator function of B (e.g., $[x = 3]$ is 1 if $x = 3$, and 0 otherwise). We will denote the set of permutations over the set S as $\mathrm{Sym}(S)$. We will use $\left[{n \atop k}\right]$ to denote the Stirling numbers of the first kind (see, e.g., [11, §6.1]), that is, the number of permutations of n distinct elements containing exactly k cycles.

3.2 A useful fact

We end this section showing the following intuitive fact — used extensively throughout the paper — stating that if we know the probability distribution over documents of a given length we can

[3] Observe that as $\alpha \to 0$, the resulting document distributions converge to the uniform single-topic document distributions over the same topics; on the other hand, as $\alpha \to \infty$, the resulting document distributions converge to those of a model over the single topic that can be obtained by averaging the original topics.

[4] This is different from the probability that a random document of length ℓ contains exactly c_1 words 1, c_2 words 2, ..., and c_m words m — the latter is equal to $\binom{\ell}{c_1, c_2, \ldots, c_m} \cdot \mathcal{A}_\ell(c_1, \ldots, c_m)$

recover the probability distributions over shorter documents.

Observation 1. *Let $\mathcal{A}$ be a bag-of-words model over the vocabulary $\mathcal{V}$, let $\ell \geq s \geq 0$ be integers, and let $d \in \mathcal{V}^s$ be a document. Then,*

$$\mathcal{A}_s(d) = \sum_{\substack{d' \in \mathcal{V}^\ell \\ d \text{ is a prefix of } d'}} \mathcal{A}_\ell(d').$$

Therefore, $\mathcal{A}_s$ is determined by $\mathcal{A}_\ell$.

Proof. Suppose, wlog, that $\mathcal{V} = [m]$ for some positive integer m. Moreover, suppose that d contains word $d(j) \in [m]$ in its jth position. Then,

$$\begin{aligned}
\mathcal{A}_s(d) &= E_{(p(1),\ldots,p(m))\sim\mathcal{A}}\left[\prod_{j=1}^{s} p(d(j))\right] = E_{(p(1),\ldots,p(m))\sim\mathcal{A}}\left[\prod_{j=1}^{s} p(d(j)) \cdot \prod_{j=1}^{\ell-s} 1\right] \\
&= E_{(p(1),\ldots,p(m))\sim\mathcal{A}}\left[\prod_{j=1}^{s} p(d(j)) \cdot \prod_{j=1}^{\ell-s}\sum_{i=1}^{m} p(i)\right] \\
&= E_{(p(1),\ldots,p(m))\sim\mathcal{A}}\left[\prod_{j=1}^{s} p(d(j)) \cdot \sum_{i_1,\ldots,i_{\ell-s}\in[m]^{\ell-s}} \prod_{j=1}^{\ell-s} p(i_j)\right] \\
&= E_{(p(1),\ldots,p(m))\sim\mathcal{A}}\left[\sum_{\substack{d' \in [m]^\ell \\ d \text{ is a prefix of } d'}} \prod_{j=1}^{\ell} p(d'(j))\right] \\
&= \sum_{\substack{d' \in [m]^\ell \\ d \text{ is a prefix of } d'}} E_{(p(1),\ldots,p(m))\sim\mathcal{A}}\left[\prod_{j=1}^{\ell} p(d'(j))\right] = \sum_{\substack{d' \in [m]^\ell \\ d \text{ is a prefix of } d'}} \mathcal{A}_\ell(d').
\end{aligned}$$

□

4 Equivalence between Uniform Single-Topic and LDA models

In this section we give our main result: a set of topics is identifiable under the uniform LDA model with Dirichlet parameter α at a given document length, if and only if the same set is identifiable under the uniform single-topic model at the same document length.

This result allows us to establish impossibility results for topic identification within the LDA framework by focusing on the simpler uniform, single-topic case. Indeed, later in the paper we will provide a sequence of results of the latter type which, by virtue of this reduction, will translate automatically to the LDA framework. In fact, this reduction is actually an equivalence and this will allows us in many instances to prove the tightness of our impossibility results.

In Appendix A, we discuss the role of the knowledge of α in the reduction.

We now proceed to establishing the equivalence. We begin by writing down the probability of a document d of length ℓ that is given by LDA. The resulting expression is very complicated, but fortunately it exhibits a rich combinatorial structure that we are able to exploit. The expression can be reorganized and written down in an equivalent way as a polynomial combination of terms, each of which depends only on the uniform, single-topic probabilities of sub-documents of d. Hence, if

we have two sets of topics $\mathcal{T}$ and $\mathcal{T}'$ inducing equal distributions on uniform single-topic documents up to length ℓ, then the same topics will induce equal distributions on documents up to length ℓ when generated by LDA!

We begin by computing the probability of a generic document d in the LDA model.

Lemma 2. *Let $\alpha > 0$ and let $\mathcal{T}$ be any set of K topics on $\mathcal{V}$. Then, for any $d \in \mathcal{V}^\ell$,*

$$\mathcal{D}_\ell^{\mathcal{T},\alpha}(d) = \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \sum_{\sigma\in[K]^\ell} \left(\prod_{i=1}^{K} \frac{\Gamma\left(\alpha + \sum_{j=1}^{\ell}[\sigma(j)=i]\right)}{\Gamma(\alpha)} \cdot \prod_{j=1}^{\ell} p_{\sigma(j)}(d(j)) \right). \tag{1}$$

Proof. Let $p_k(i)$ be the probability given by the k-th topic of $\mathcal{T}$ to the word $i \in \mathcal{V}$. Then,

$$\begin{aligned}
\mathcal{D}_\ell^{\mathcal{T},\alpha}(d) &= E_{(x_1,\ldots,x_K)\sim \mathrm{Dir}_K(\alpha)} \left[\prod_{j=1}^{\ell} \sum_{k=1}^{K} (x_k \cdot p_k(d(j))) \right] \\
&= E_{(x_1,\ldots,x_K)\sim \mathrm{Dir}_K(\alpha)} \left[\sum_{\sigma\in[K]^\ell} \prod_{j=1}^{\ell} \left(x_{\sigma(j)} \cdot p_{\sigma(j)}(d(j))\right) \right] \\
&= \sum_{\sigma\in[K]^\ell} E_{(x_1,\ldots,x_K)\sim \mathrm{Dir}_K(\alpha)} \left[\prod_{j=1}^{\ell} \left(x_{\sigma(j)} \cdot p_{\sigma(j)}(d(j))\right) \right] \\
&= \sum_{\sigma\in[K]^\ell} \left(E_{(x_1,\ldots,x_K)\sim \mathrm{Dir}_K(\alpha)} \left[\prod_{j=1}^{\ell} x_{\sigma(j)} \right] \cdot \prod_{j=1}^{\ell} p_{\sigma(j)}(d(j)) \right) \\
&= \sum_{\sigma\in[K]^\ell} \left(E_{(x_1,\ldots,x_K)\sim \mathrm{Dir}_K(\alpha)} \left[\prod_{i=1}^{K} x_i^{\sum_{j=1}^{\ell}[\sigma(j)=i]} \right] \cdot \prod_{j=1}^{\ell} p_{\sigma(j)}(d(j)) \right) = (\star)
\end{aligned}$$

We now apply the following identity for the product moments of the Dirichlet distribution (see, e.g., [8, §27.6]):

Lemma 3 (Moments of the Dirichlet Distribution [8])**.** *Let $\alpha > 0$ and $(a_1,\ldots,a_K) \in \mathbf{N}^K$. Then,*

$$E_{(x_1,\ldots,x_K)\sim \mathrm{Dir}_K(\alpha)} \left[\prod_{i=1}^{K} x_i^{a_i} \right] = \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\sum_{i=1}^{K} a_i)} \cdot \prod_{i=1}^{K} \frac{\Gamma(\alpha+a_i)}{\Gamma(\alpha)}.$$

Then,

$$(\star) = \sum_{\sigma\in[K]^\ell} \left(\frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \prod_{i=1}^{K} \frac{\Gamma\left(\alpha + \sum_{j=1}^{\ell}[\sigma(j)=i]\right)}{\Gamma(\alpha)} \cdot \prod_{j=1}^{\ell} p_{\sigma(j)}(d(j)) \right). \qquad \square$$

We give next our main technical theorem of the section: the probability given to a generic document d by an LDA model is a function of the probabilities given to the subdocuments of d by the uniform Single-Topic model. The next definition lays the groundwork.

Definition 4. *Given a permutation $\pi \in \mathrm{Sym}([\ell])$, let $\mathcal{C}_\pi$ be the partition of $[\ell]$ into the cycles of π:*

$$\mathcal{C}_\pi = \{S \mid S \subseteq [\ell] \text{ and the elements of } S \text{ form a cycle in } \pi\}.$$

Furthermore, for $d \in \mathcal{V}^\ell$ and $S = \{i_1, i_2, \ldots, i_{|S|}\} \subseteq [\ell]$ with $i_1 < i_2 < \ldots < i_{|S|}$, let $d_{|S}$ be the document containing the words $d(i_1), \ldots, d(i_{|S|})$ in this order (that is, let it be the document that is obtained by removing from d the words whose positions in d are not in S).

For example, if $\pi = (163)(25)(4)$ then $\mathcal{C}_\pi = \{\{1,3,6\},\{2,5\},\{4\}\}$. And, if $d = w_1w_2w_3w_4w_5w_6$ and $S = \{1,3,6\}$ then $d_{|S} = w_1w_3w_6$.

Theorem 5. *Let $\mathcal{T}$ be any set of K topics on $\mathcal{V}$ and consider any $d \in \mathcal{V}^\ell$. Then, for any $\alpha > 0$,*

$$\mathcal{D}_\ell^{\mathcal{T},\alpha}(d) = \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \sum_{\pi\in\mathrm{Sym}([\ell])} \prod_{S\in C_\pi} \left(K\cdot\alpha\cdot\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})\right). \tag{2}$$

Proof. We start from Equation 2 and exploit its combinatorial structure to transform it into Equation 1.

For $\sigma \in [K]^\ell$, let C_σ be the (unlabeled) partition of $[\ell]$ according to the σ labeling — that is, $S \in C_\sigma$ iff S is a maximal subset of $[\ell]$ such that $\sigma(i) = \sigma(j)$ for all $i, j \in S$. Given two partitions C_π and C_σ of $[\ell]$, we write $C_\pi \preceq C_\sigma$ (C_π is finer than C_σ) iff for each $S \in C_\pi$ there exists $S' \in C_\sigma$ such that $S \subseteq S'$. Then, the right-hand side of the Equation 2 is equal to

$$\begin{aligned}
&\frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \sum_{\pi\in\mathrm{Sym}([\ell])} \prod_{S\in C_\pi} \left(K\cdot\alpha\cdot\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})\right) \\
&= \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \sum_{\pi\in\mathrm{Sym}([\ell])} \prod_{S\in C_\pi} \left(\alpha\cdot\sum_{t=1}^{K}\prod_{j\in S} p_t(d(j))\right) \\
&= \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \sum_{\pi\in\mathrm{Sym}([\ell])} \left(\alpha^{|C_\pi|}\cdot\prod_{S\in C_\pi}\sum_{t=1}^{K}\prod_{j\in S} p_t(d(j))\right) \\
&= \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \sum_{\sigma\in[K]^\ell} \left(\left(\prod_{j=1}^{\ell} p_{\sigma(j)}\left(d(j)\right)\right)\cdot \sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ C_\pi\preceq C_\sigma}} \alpha^{|C_\pi|}\right) = (\star\star)
\end{aligned}$$

where the last step follows from the fact that each permutation π having a cycle structure C_π finer than C_σ contributes with an $\alpha^{|C_\pi|}$ addend to the coefficient of $\prod_{j=1}^{\ell} p_{\sigma(j)}\left(d(j)\right)$, and no other permutation contributes to that coefficient.

Now, each permutation $\pi \in \mathrm{Sym}([\ell])$ in the inner sum has a cycle structure C_π that is finer than C_σ — that is, each such permutation can be obtained by permuting arbitrarily the elements of each sets $S \in C_\sigma$. In other words, if, wlog, $C_\sigma = \{S_1, \ldots, S_{|C_\sigma|}\}$, we have that

$$\sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ C_\pi\preceq C_\sigma}} \alpha^{|C_\pi|} = \sum_{\pi_1\in\mathrm{Sym}(S_1)} \sum_{\pi_2\in\mathrm{Sym}(S_2)} \cdots \sum_{\pi_{|C_\sigma|}\in\mathrm{Sym}(S_{|C_\sigma|})} \alpha^{\sum_{i=1}^{|C_\sigma|}|C_{\pi_i}|}.$$

Recall that the number of permutations of a ground set of n elements containing exactly t cycles

is the Stirling number of the first kind $\begin{bmatrix} n \\ t \end{bmatrix}$. Then,

$$\begin{aligned}
\sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ C_\pi\preceq C_\sigma}} \alpha^{|C_\pi|} &= \sum_{\pi_1\in\mathrm{Sym}(S_1)}\sum_{\pi_2\in\mathrm{Sym}(S_2)}\cdots\sum_{\pi_{|C_\sigma|}\in\mathrm{Sym}(S_{|C_\sigma|})} \alpha^{\sum_{i=1}^{|C_\sigma|}|C_{\pi_i}|}\\
&= \sum_{\pi_1\in\mathrm{Sym}(S_1)}\cdots\sum_{\pi_{|C_\sigma|-1}\in\mathrm{Sym}(S_{|C_\sigma|-1})}\left(\alpha^{\sum_{i=1}^{|C_\sigma|-1}|C_{\pi_i}|}\cdot\sum_{\pi_{|C_\sigma|}\in\mathrm{Sym}(S_{|C_\sigma|})}\alpha^{|C_{\pi_{|C_\sigma|}}|}\right)\\
&= \sum_{\pi_1\in\mathrm{Sym}(S_1)}\cdots\sum_{\pi_{|C_\sigma|-1}\in\mathrm{Sym}(S_{|C_\sigma|-1})}\left(\alpha^{\sum_{i=1}^{|C_\sigma|-1}|C_{\pi_i}|}\cdot\sum_{t=1}^{|S_{|C_\sigma|}|}\left(\begin{bmatrix}|S_{|C_\sigma|}|\\ t\end{bmatrix}\alpha^t\right)\right)\\
&= \sum_{t=1}^{|S_{|C_\sigma|}|}\left(\begin{bmatrix}|S_{|C_\sigma|}|\\ t\end{bmatrix}\alpha^t\right)\cdot\sum_{\pi_1\in\mathrm{Sym}(S_1)}\cdots\sum_{\pi_{|C_\sigma|-1}\in\mathrm{Sym}(S_{|C_\sigma|-1})}\alpha^{\sum_{i=1}^{|C_\sigma|-1}|C_{\pi_i}|}\\
&= \sum_{t=1}^{|S_{|C_\sigma|}|}\left(\begin{bmatrix}|S_{|C_\sigma|}|\\ t\end{bmatrix}\alpha^t\right)\cdot\sum_{t=1}^{|S_{|C_\sigma|-1}|}\left(\begin{bmatrix}|S_{|C_\sigma|-1}|\\ t\end{bmatrix}\alpha^t\right)\cdot\sum_{\pi_1\in\mathrm{Sym}(S_1)}\cdots\sum_{\pi_{|C_\sigma|-2}\in\mathrm{Sym}(S_{|C_\sigma|-2})}\alpha^{\sum_{i=1}^{|C_\sigma|-2}|C_{\pi_i}|}\\
&\ \ \vdots\\
&= \prod_{S\in C_\sigma}\sum_{t=1}^{|S|}\left(\begin{bmatrix}|S|\\ t\end{bmatrix}\alpha^t\right).
\end{aligned}$$

Recall also that, for an integer $n\geq 1$ and for $x>0$, the rising factorial $x^{\overline{n}} := x\cdot(x+1)\cdots(x+n-1)$ equals $x^{\overline{n}} = \sum_{t=1}^{n}\left(\begin{bmatrix} n \\ t \end{bmatrix}\cdot x^t\right)$ (see, e.g., [11, §6.11]). Thus,

$$\sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ C_\pi\preceq C_\sigma}}\alpha^{|C_\pi|} = \prod_{S\in C_\sigma}\sum_{t=1}^{|S|}\left(\begin{bmatrix}|S|\\ t\end{bmatrix}\alpha^t\right) = \prod_{S\in C_\sigma}\prod_{j=0}^{|S|-1}(\alpha+j) = \prod_{S\in C_\sigma}\frac{\Gamma(\alpha+|S|)}{\Gamma(\alpha)},$$

where $\Gamma(x)$ is the Γ function, defined as $\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt$, and which satisfies $\Gamma(x+1) = x\Gamma(x)$ (see, for instance, [25, §12.1]). Going back to $(\star\star)$, we then get,

$$\begin{aligned}
(\star\star) &= \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)}\cdot\sum_{\sigma\in[K]^\ell}\left(\left(\prod_{j=1}^{\ell}p_{\sigma(j)}\left(d(j)\right)\right)\cdot\prod_{S\in C_\sigma}\frac{\Gamma(\alpha+|S|)}{\Gamma(\alpha)}\right)\\
&= \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)}\cdot\sum_{\sigma\in[K]^\ell}\left(\left(\prod_{j=1}^{\ell}p_{\sigma(j)}\left(d(j)\right)\right)\cdot\prod_{i=1}^{K}\frac{\Gamma(\alpha+[\sum_{j=1}^{\ell}\sigma(j)=i])}{\Gamma(\alpha)}\right)\\
&= \mathcal{D}_\ell^{\mathcal{T},\alpha}(d).
\end{aligned}$$

□

Corollary 6. *Let $\alpha>0$ and $\mathcal{T}$ be any set of K topics. Then, the distribution $\mathcal{D}_\ell^{\mathcal{T},\alpha}$ is a function of the distribution $\mathcal{S}_\ell^{\mathcal{T}}$ and of the real number $K\cdot\alpha$. Analogously, the distribution $\mathcal{S}_\ell^{\mathcal{T}}$ is a function of the distribution $\mathcal{D}_\ell^{\mathcal{T},\alpha}$ and of the real number $K\cdot\alpha$.*

Proof. Theorem 5 states that $\mathcal{D}_{|d|}^{\mathcal{T},\alpha}(d)$ is a function of $K\cdot\alpha$ and of $\mathcal{S}_{|d'|}^{\mathcal{T}}(d')$ for each d' that can be obtained by removing zero or more words from d. By Observation 1, $\mathcal{S}_{|d'|}^{\mathcal{T}}$ is a function of $\mathcal{S}_{|d|}^{\mathcal{T}}$. The first claim, then, follows directly from Theorem 5 and Observation 1.

We prove the second claim by induction. Observe that, for the only document d of length $\ell = 0$, we clearly have $\mathcal{D}_0^{\mathcal{T},\alpha}(d) = \mathcal{S}_0^{\mathcal{T}}(d) = 1$. Now, suppose that, for $\ell \geq 1$, the claim holds for documents of length $\ell - 1$ — we will prove it for documents of length ℓ. Theorem 5 states that,

$$\mathcal{D}_\ell^{\mathcal{T},\alpha}(d) = \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \sum_{\pi\in\mathrm{Sym}([\ell])} \prod_{S\in C_\pi} \left(K\cdot\alpha\cdot\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})\right).$$

We rewrite the expression,

$$\begin{aligned}
\mathcal{D}_\ell^{\mathcal{T},\alpha}(d) &= \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \left(\sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ |C_\pi|=1}} \prod_{S\in C_\pi} \left(K\cdot\alpha\cdot\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})\right) + \sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ |C_\pi|\geq 2}} \prod_{S\in C_\pi} \left(K\cdot\alpha\cdot\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})\right) \right) \\
&= \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \left(\sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ |C_\pi|=1}} \left(K\cdot\alpha\cdot\mathcal{S}_{\ell}^{\mathcal{T}}(d)\right) + \sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ |C_\pi|\geq 2}} \prod_{S\in C_\pi} \left(K\cdot\alpha\cdot\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})\right) \right) \\
&= \frac{\Gamma(K\cdot\alpha)}{\Gamma(K\cdot\alpha+\ell)} \cdot \left((\ell-1)!\cdot\left(K\cdot\alpha\cdot\mathcal{S}_{\ell}^{\mathcal{T}}(d)\right) + \sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ |C_\pi|\geq 2}} \prod_{S\in C_\pi} \left(K\cdot\alpha\cdot\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})\right) \right).
\end{aligned}$$

By rearranging, we get

$$\mathcal{S}_\ell^{\mathcal{T}}(d) = \frac{\Gamma(K\cdot\alpha+\ell)}{\Gamma(K\cdot\alpha+1)\cdot\Gamma(\ell)} \cdot \mathcal{D}_\ell^{\mathcal{T},\alpha}(d) - \frac{1}{K\cdot\alpha\cdot\Gamma(\ell)} \cdot \sum_{\substack{\pi\in\mathrm{Sym}([\ell])\\ |C_\pi|\geq 2}} \prod_{S\in C_\pi} \left(K\cdot\alpha\cdot\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})\right). \tag{3}$$

If $|C_\pi| \geq 2$, and $S \in C_\pi$, then $|S| \subsetneq [\ell]$ — and, for each $S \subsetneq [\ell]$, we have that $\left|d_{|S}\right| = |S| \leq \ell - 1$. By the inductive hypothesis all the terms $\mathcal{S}_{|S|}^{\mathcal{T}}(d_{|S})$ in the latter expression of $\mathcal{S}_\ell^{\mathcal{T}}$ are functions of $\mathcal{D}_{|S|}^{\mathcal{T},\alpha}$ and thus, by Observation 1, also of $\mathcal{D}_\ell^{\mathcal{T},\alpha}$. Since the only remaining terms in the equation are functions of $K\cdot\alpha$, $\mathcal{D}_\ell^{\mathcal{T},\alpha}(d)$ and ℓ, both the inductive step and the Corollary have been proved. □

The following theorem now follows immediately.

Theorem 7. *Let $\alpha > 0$ and suppose that $|\mathcal{T}| = |\mathcal{T}'|$. Then, $\mathcal{S}_\ell^{\mathcal{T}} = \mathcal{S}_\ell^{\mathcal{T}'}$ if and only if $\mathcal{D}_\ell^{\mathcal{T},\alpha} = \mathcal{D}_\ell^{\mathcal{T}',\alpha}$.*

5 Topic Unidentifiability

The results of the previous section tell us that impossibility results for topic identifiability for the uniform, single-topic model transfer automatically to the LDA case. In this section we provide such a result. We provide a construction of sets $\mathcal{T}$ of K topics which are not identifiable under the single-topic model: specifically, for any such set $\mathcal{T}$ of K topics, there is another set $\mathcal{T}'$ of K topics such that the distributions they induce on documents of length $\ell < K$ are the same, *i.e.* $\mathcal{S}_\ell^{\mathcal{T}} = \mathcal{S}_\ell^{\mathcal{T}'}$, $\forall \ell < K$. Since $\mathcal{T}$ and $\mathcal{T}'$ induce exactly the same distribution over all documents of length $\ell < K$, they cannot be distinguished starting from documents of that length. It follows that topics are not identifiable when $\ell < K$. In fact, there are uncountably many such sets $\mathcal{T}$ of topics, each being indistinguishable from uncountably many other topic sets $\mathcal{T}'$.

To prove the result we first give a construction for a vocabulary with just two words, and then show how to extend it to general vocabularies.

5.1 Vocabulary with just two words

We begin by exhibiting pairs of topic sets indistinguishable from each other in the binary vocabulary setting, $\mathcal{V} = \{1, 2\}$. In fact, in this setting we show that a randomly picked set of topics will do. Except for a set of measure zero, if $\mathcal{T}$ is a random set of K topics (say, from a Dirichlet distribution, or any "reasonable" distribution) then there exist uncountably many $\mathcal{T}'$ inducing the same distribution over all documents of length $\ell < K$.

To prove the result we focus on documents of a very simple nature. We say that a document d of length ℓ is *pure* if it contains the same word ℓ times. The first step in the construction is to show that the probability distribution induced on pure documents uniquely determines the probabilities of all documents. In the binary vocabulary case, topics are just "coins", *i.e.* they are of the form $(p_i, 1 - p_i)$, $i \in [K]$. Recall that the bag-of-words property was defined in Section 3 and that c_i is the number of times word i appears in a document.

Observation 8. *Let $\mathcal{V} = \{1, 2\}$ and $\mathcal{A}$ be any model on $\mathcal{V}$ satisfying the bag-of-words property. Then, the probability of any given document of length $\ell = c_1 + c_2$ can be expressed as a linear combination of the probabilities of pure documents of length at most ℓ:*

$$\mathcal{A}_\ell(c_1, c_2) = \sum_{i=0}^{c_2} \binom{c_2}{i} (-1)^i \cdot \mathcal{A}_{c_1+i}(c_1 + i, 0).$$

Analogously,

$$\mathcal{A}_\ell(c_1, c_2) = \sum_{i=0}^{c_1} \binom{c_1}{i} (-1)^i \cdot \mathcal{A}_{c_2+i}(0, c_2 + i).$$

Proof. Let p be the probability of word 1 under model $\mathcal{A}$. Then,

$$\begin{aligned}\mathcal{A}_\ell(c_1, c_2) &= E_{(p,1-p)\sim\mathcal{A}}\left[p^{c_1} \cdot (1-p)^{c_2}\right] = E_{(p,1-p)\sim\mathcal{A}}\left[\sum_{i=0}^{c_2} \binom{c_2}{i} (-1)^i p^{c_1+i}\right] \\ &= \sum_{i=0}^{c_2} \binom{c_2}{i} (-1)^i \cdot E_{(p,1-p)\sim\mathcal{A}}\left[p^{c_1+i}\right] = \sum_{i=0}^{c_2} \binom{c_2}{i} (-1)^i \cdot \mathcal{A}_{c_1+i}(c_1 + i, 0).\end{aligned}$$

The proof of the second claim is analogous. □

Observation 8 tells us that if two sets of topics induce the same distribution on pure documents of a given length, they also induce the same distribution on all documents of that length. The next lemma says that finding a set of topics $\mathcal{T}'$ that is indistinguishable from $\mathcal{T}$ on the basis of pure documents is equivalent to solving a certain system of polynomial equations, and that this system admits uncountably many solutions.

Lemma 9. *Let $0 < p_1 < p_2 < \cdots < p_n < 1$. Consider the system of equations*

$$\sum_{i=1}^{n} x_i^j = \sum_{i=1}^{n} p_i^j,$$

for $j = 1, \ldots, n-1$. Then, there exist uncountably many solutions $\boldsymbol{x} = (x_1, \ldots, x_n) \in (0, 1)^n$ to the system.

Proof. Let $f_j(\boldsymbol{x}) := \sum_{i=1}^{n} x_i^j - \sum_{i=1}^{n} p_i^j$, so that our system becomes $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$, where $\boldsymbol{f} : \mathbf{R}^n \to \mathbf{R}^{n-1}$ and $\boldsymbol{0} \in \mathbf{R}^{n-1}$. The Jacobian of $\boldsymbol{f}$ w.r.t. $x_1, x_2, \ldots, x_{n-1}$ at $\boldsymbol{x} = \boldsymbol{p}$ is

$$J_{\boldsymbol{f},\boldsymbol{x_1},\ldots,\boldsymbol{x_{n-1}}}(\boldsymbol{p}) = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 2p_1 & 2p_2 & \ldots & 2p_{n-1} \\ 3p_1^2 & 3p_2^2 & \ldots & 3p_{n-1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ (n-1)p_1^{n-2} & (n-1)p_2^{n-2} & \ldots & (n-1)p_{n-1}^{n-2} \end{bmatrix}$$

$$= \operatorname{diag}\left(\begin{bmatrix} 1 \\ 2 \\ \vdots \\ n-1 \end{bmatrix} \right) \cdot \begin{bmatrix} 1 & 1 & \ldots & 1 \\ p_1 & p_2 & \ldots & p_{n-1} \\ p_1^2 & p_2^2 & \ldots & p_{n-1}^2 \\ \vdots & \vdots & \ddots & \vdots \\ p_1^{n-2} & p_2^{n-2} & \ldots & p_{n-1}^{n-2} \end{bmatrix},$$

which is invertible because the last matrix is a Vandermonde matrix of rank $n-1$, since the p_i's are all distinct.

Since Jacobian is invertible and $\boldsymbol{f}(\boldsymbol{p}) = \boldsymbol{0}$, we can apply the implicit function theorem (see, e.g., [17]) to conclude that there exist an open disk D in $\mathbf{R}^{n-1}$ containing $(p_1, \ldots, p_{n-1})$, and a unique continuously differentiable function $g : \mathbf{R}^{n-1} \to \mathbf{R}$ such that

$$\boldsymbol{f}(x_1, \ldots, x_{n-1}, g(x_1, \ldots, x_{n-1})) = \boldsymbol{0},$$

for all $(x_1, \ldots, x_{n-1}) \in D$. □

We can now prove the main impossibility result of this section. Section 6 shows the converse of Theorem 10 by demonstrating that the topics can be recovered if we have access to documents of length $\ell \geq K$.

Theorem 10 (Binary Uniform Single-Topic Unidentifiability)**.** *Let $m = 2$, $\mathcal{V} = [m]$, $\beta > 0$, and let the documents be generated by a uniform single-topic model $\mathcal{S}^{\mathcal{T}}$ on a set $\mathcal{T}$ of K topics, each sampled independently from* $\mathrm{Dir}_m(\beta)$*. Then, no algorithm can reconstruct $\mathcal{T}$ from m, K, β, and $\mathcal{S}_1^{\mathcal{T}}, \ldots, \mathcal{S}_{K-1}^{\mathcal{T}}$.*

Proof. Let $\mathcal{T} = \{(p_1, 1-p_1), \ldots, (p_K, 1-p_K)\}$ where $(p_k, 1-p_k) \sim \mathrm{Dir}_m(\beta)$. Since $\mathrm{Dir}_m(\beta)$ is a continuous random variable, the probability that there exists $i \neq j$ such that $p_i = p_j$ is 0. By Theorem 9, there exists a continuum of sequences $p'_1 < \ldots < p'_K$, and of $\mathcal{T}' = \{(p'_1, 1-p'_1), \ldots, (p'_K, 1-p'_K)\}$, such that $\mathcal{S}_\ell^{\mathcal{T}}(d) = \mathcal{S}_\ell^{\mathcal{T}'}(d')$ for all pure documents d of length $\ell \leq K-1$ containing ℓ copies of the word 1. Observation 8, in turn, implies that $\mathcal{S}_{K-1}^{\mathcal{T}} = \mathcal{S}_{K-1}^{\mathcal{T}'}$.

Now, for each $\epsilon > 0$, there exists a convex set $A \subseteq \{(x, y) \mid x + y = 1 \wedge x, y \geq 0\}$ such that the probability that an open ball centered in $\mathrm{Dir}_m(\beta)$ lies inside A is at least $1 - \epsilon$, and such that the ratio between the maximum and the minimum density of the distribution $\mathrm{Dir}_m(\beta)$ in S is a constant. It follows that the likelihoods of each of a continuum of worlds compatible (up to length $K-1$) with $\mathcal{S}^{\mathcal{T}}$ are within a constant of each other. Hence, no algorithm can identify the topics. □

Thanks to our reduction, the lower bound above transfers automatically to the LDA case.

Corollary 11 (Binary LDA Unidentifiability)**.** *Let $m = 2$, $\alpha, \beta > 0$, and let the documents be generated by a LDA model $\mathcal{D}^{\mathcal{T},\alpha}$ on a set $\mathcal{T}$ of K topics, each sampled independently from* $\mathrm{Dir}_m(\beta)$*. Then, no algorithm can reconstruct $\mathcal{T}$ from m, K, α, β, and $\mathcal{D}_1^{\mathcal{T},\alpha}, \ldots, \mathcal{D}_{K-1}^{\mathcal{T},\alpha}$.*

Proof. Apply Theorem 7 to Theorem 10. □

5.2 General vocabularies

We now give a simple reduction from lower bounds for vocabularies of $m = 2$ words, to lower bounds for general vocabularies. Let $\mathcal{S}^{\mathcal{T},\theta}$ denote the document probability function of a single-topic model on topics $\mathcal{T}$ with weights $\theta \in \mathbf{R}^K, \sum_k \theta_k = 1$. Now, suppose that $\mathcal{T} = \{(p_1, 1-p_1), \ldots, (p_K, 1-p_K)\}$ and $\mathcal{T}' = \{(p'_1, 1-p'_1), \ldots, (p'_K, 1-p'_K)\}$ are two distinct families of K topics on the vocabulary $\mathcal{V} = \{1, 2\}$, such that $\mathcal{S}_\ell^{\mathcal{T},\theta} = \mathcal{S}_\ell^{\mathcal{T}',\theta'}$ and $\mathcal{S}_{\ell+1}^{\mathcal{T},\theta} \neq \mathcal{S}_{\ell+1}^{\mathcal{T}',\theta'}$ — possibly, with $\theta = \theta'$.

We will give a reduction that increases the number of words while guaranteeing that the two worlds remain unidentifiable at length ℓ and identifiable at length $\ell + 1$.

Choose some $x \in (0, 1)$, and let $m > 2$ be the desired number of words. We define two sets of topics $\hat{\mathcal{T}}$ and $\hat{\mathcal{T}}'$ on the m words as follows:

$$\hat{\mathcal{T}} = \left\{ \left(xp_1, x(1-p_1), \frac{1-x}{m-2}, \ldots, \frac{1-x}{m-2} \right), \ldots, \left(xp_K, x(1-p_K), \frac{1-x}{m-2}, \ldots, \frac{1-x}{m-2} \right) \right\},$$

and

$$\hat{\mathcal{T}}' = \left\{ \left(xq_1, x(1-q_1), \frac{1-x}{m-2}, \ldots, \frac{1-x}{m-2} \right), \ldots, \left(xq_K, x(1-q_K), \frac{1-x}{m-2}, \ldots, \frac{1-x}{m-2} \right) \right\}.$$

Theorem 12. *If $\mathcal{S}_\ell^{\mathcal{T},\theta} = \mathcal{S}_\ell^{\mathcal{T}',\theta'}$ then $\mathcal{S}_\ell^{\hat{\mathcal{T}},\theta} = \mathcal{S}_\ell^{\hat{\mathcal{T}}',\theta'}$.*

Proof. Let $c_1, \ldots, c_m \in \mathbf{N}$ be given, with $\sum_{i=1}^m c_i = \ell$. Then,

$$\begin{aligned}
\mathcal{S}_\ell^{\hat{\mathcal{T}},\theta}(c_1, \ldots, c_m) &= \sum_{k=1}^K \left(\theta_k \cdot (xp_k)^{c_1} \cdot (x(1-p_k))^{c_2} \cdot \prod_{i=3}^m \left(\frac{1-x}{m-2} \right)^{c_i} \right) \\
&= \sum_{k=1}^K \left(\theta_k \cdot p_k^{c_1} \cdot (1-p_k)^{c_2} \right) \cdot x^{c_1+c_2} \cdot \prod_{i=3}^m \left(\frac{1-x}{m-2} \right)^{c_i} \\
&= \mathcal{S}_{c_1+c_2}^{\mathcal{T},\theta}(c_1, c_2) \cdot x^{c_1+c_2} \cdot \prod_{i=3}^m \left(\frac{1-x}{m-2} \right)^{c_i} \\
&= \mathcal{S}_{c_1+c_2}^{\mathcal{T}',\theta'}(c_1, c_2) \cdot x^{c_1+c_2} \cdot \prod_{i=3}^m \left(\frac{1-x}{m-2} \right)^{c_i} \\
&= \sum_{k=1}^K \left(\theta'_k \cdot (p'_k)^{c_1} \cdot (1-p'_k)^{c_2} \right) \cdot x^{c_1+c_2} \cdot \prod_{i=3}^m \left(\frac{1-x}{m-2} \right)^{c_i} \\
&= \sum_{k=1}^K \left(\theta'_k \cdot \left(xp'_k\right)^{c_1} \cdot \left(x(1-p'_k)\right)^{c_2} \right) \cdot \prod_{i=3}^m \left(\frac{1-x}{m-2} \right)^{c_i} = \mathcal{S}_\ell^{\hat{\mathcal{T}}',\theta'}(c_1, \ldots, c_m).
\end{aligned}$$

□

Theorem 12 together with the construction supporting Corollary 10 gives us our general case unidentifiability results.

Corollary 13 (Uniform Single-Topic Unidentifiability)**.** *Let $m \geq 2$, $\mathcal{V} = [m]$ and let the documents be generated by a uniform single-topic model $\mathcal{S}^{\mathcal{T}}$ on a set $\mathcal{T}$ of K adversarial topics. Then, no algorithm can reconstruct $\mathcal{T}$ from m, K and $\mathcal{S}_1^{\mathcal{T}}, \ldots, \mathcal{S}_{K-1}^{\mathcal{T}}$.*

Corollary 14 (LDA Unidentifiability)**.** *Let $m \geq 2$, $\alpha > 0$ and let the documents be generated by a LDA model $\mathcal{D}^{\mathcal{T},\alpha}$ on a set $\mathcal{T}$ of K adversarial topics. Then, no algorithm can reconstruct $\mathcal{T}$ from m, K and α, and $\mathcal{D}_1^{\mathcal{T},\alpha}, \ldots, \mathcal{D}_{K-1}^{\mathcal{T},\alpha}$.*

Proof. Apply Theorem 7 to Corollary 13. □

6 Identifiability with documents of length K

We now proceed to show that if $\mathcal{S}_K^{\mathcal{T}}$ is known then the set $\mathcal{T}$ of K topics can be identified, except for a set of measure zero. We make use of the following known result [19].

Lemma 15 (Newton's identities)**.** *Let $f(x)$: $\mathbf{R} \to \mathbf{R}$ be a n-degree polynomial with roots $x_1, x_2, \ldots, x_n$. Let $a_j = \sum_{i=1}^{n} x_i^j$, $1 \le j \le n$. Then, $f(x) = \sum_{j=0}^{n} (-1)^j e_j x^{n-j}$, where*

$$e_j = \frac{1}{j!} \det \begin{bmatrix} a_1 & 1 & 0 & \cdots & \\ a_2 & a_1 & 2 & 0 & \cdots \\ \vdots & & \ddots & \ddots & \\ a_{j-1} & a_{j-2} & \cdots & a_1 & j-1 \\ a_j & a_{j-1} & \cdots & a_2 & a_1 \end{bmatrix}$$

Corollary 16 (Binary Uniform Single-Topic Identifiability)**.** *Let $m = 2$, $\mathcal{V} = [m]$, and let the documents be generated by a uniform single-topic model $\mathcal{S}^{\mathcal{T}}$ on a arbitrary multiset $\mathcal{T}$ of K topics. Then, there exists an algorithm that can reconstruct $\mathcal{T}$ from K and $\mathcal{S}_K^{\mathcal{T}}$.*

Proof. For $1 \le i \le K$, let p_i be the (unknown) probability of word 1 under topic i. Then, $a_j = \sum_{i=1}^{K} p_i^j$ is K times the (known) probability that a document of length j contains j copies of the word 1, for $1 \le j \le K$. By Observation 8, the a_j determine the whole document distribution $\mathcal{S}_K^{\mathcal{T}}$. Then, by Lemma 15, the unknown p_i's are simply the (unique) roots of a polynomial whose coefficients are functions of the a_j's, and can be computed in polynomial time. □

We now extends the upper bound above to work for random topics over any vocabulary.

Corollary 17 (Uniform Single-Topic Identifiability)**.** *Let $\mathcal{V} = [m]$, $\beta > 0$, and let the documents be generated by a uniform single-topic model $\mathcal{S}^{\mathcal{T}}$ on a set $\mathcal{T}$ of K topics, each sampled independently from $\mathrm{Dir}_m(\beta)$. Then, there exists an algorithm that can reconstruct $\mathcal{T}$ from K and $\mathcal{S}_K^{\mathcal{T}}$.*

Proof. Let $M_k = \{p_k(1), \ldots, p_k(m)\}$ for $k \in [K]$ be the multiset of the probabilities of the kth topic, and let $M = \bigcup_{k \in [K]} M_k$ be the multiset of the probabilities of the K topics. Then, $|M| = m \cdot K$. Since $\mathrm{Dir}_m(\beta)$ is a continuous probability distribution, and since $p_1, \ldots, p_K$ are *i.i.d.* samples distributed like $\mathrm{Dir}_m(\beta)$, the probability that there exists a subset $S \subseteq M$, with $S \notin \{M_1, \ldots, M_K\}$ and $\sum_{x \in S} x = 1$, is 0.

Now, let $\mathcal{S}_K^{\mathcal{T}_i}$, $i \in [m]$, be the document distribution that is obtained by sampling a document d from $\mathcal{S}_K^{\mathcal{T}}$, and then substituting each word different from i in d with $\star$. Then, $\mathcal{S}_K^{\mathcal{T}_i}$ is a distribution over documents of length K, on a vocabulary of two words (i and $\star$), and supported on K topics; therefore, by Corollary 16, we can recover its topics — in particular, we can recover the multiset $M(i) = \{p_1(i), p_2(i), \ldots, p_K(i)\}$, that is, the multiset of the probabilities of the word i in the K topics.

We have that $M = \bigcup_{i \in [m]} M(i)$. Since, with probability 1, each subset of M that sums up to 1 contains exactly the elements of a topic vector, and since we can label each element of M with its word, we can reconstruct the set of the K topics. □

Again, our reduction allows to port the above upper bounds for single-topics models to the LDA case.

Corollary 18 (LDA Indentifiability)**.** *Let $\mathcal{V} = [m]$, $\alpha, \beta > 0$, and let the documents be generated by a LDA model $\mathcal{D}^{\mathcal{T},\alpha}$ on a set $\mathcal{T}$ of K topics, each sampled independently from $\mathrm{Dir}_m(\beta)$. Then, there exists an algorithm that can reconstruct $\mathcal{T}$ from $\alpha \cdot K$ and $\mathcal{D}_K^{\mathcal{T},\alpha}$.*

Proof. Apply Corollary 6 to Corollary 17. □

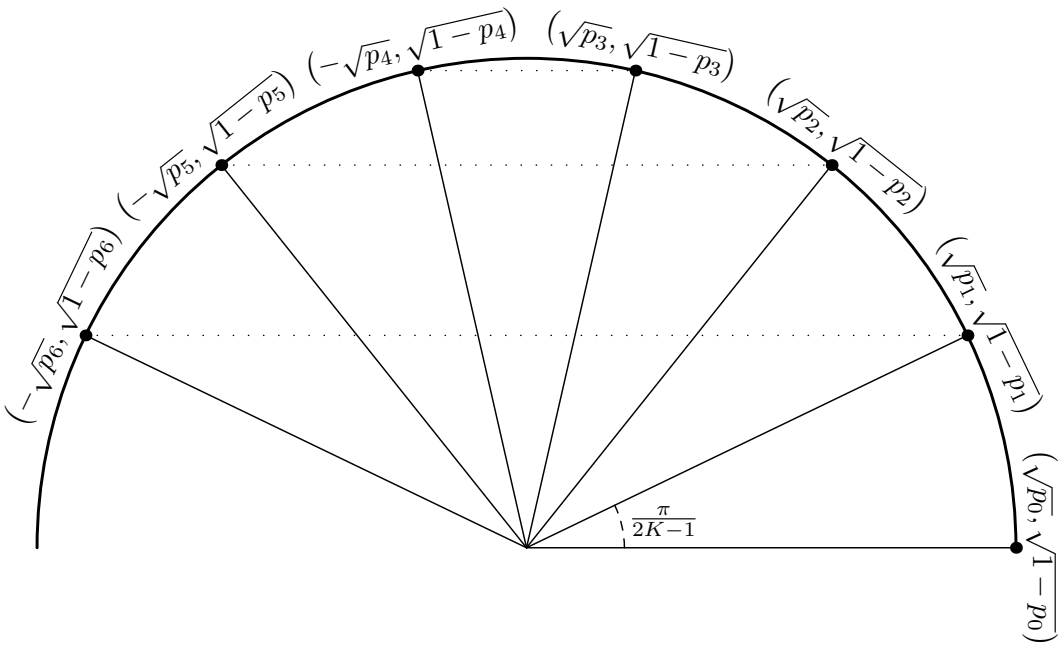

Figure 1: An illustration of the construction of Section 7, with $K = 4$.

7 An Extremal Construction

In this section we study the adversarial (non-uniform) Single Topic case, when topics are worst-case. We will exhibit two sets of K topics $\mathcal{T}$ and $\mathcal{T}'$, together with topic selection probabilities θ, such that they induce the same distribution over documents of length $\ell \leq 2K-2$. It is known that single-topic model identifiability requires $\ell \geq 2K-1$ when the topic selection θ is arbitrary and unknown [**?**, **?**, **?**]. Here we show that the same bound holds even when θ is known and almost uniform — all topics but one have the same probability of being selected. Afterwards we will explore other interesting properties of our construction. In particular, we show that our construction is extremal in the sense that our construction achieves the maximum total variation distance on documents of length $\ell = 2K-1$, while maintaining indistinguishability for $\ell < 2K-1$.

We now describe the construction. Let $i = 0, \ldots, 2K-2$, and define $p_i = \cos^2 \gamma_i$, where $\gamma_i = \pi \cdot i/2K-1$. The two families of topics are defined as $\mathcal{T} = \{(p_0, 1-p_0), \ldots, (p_{2K-2}, 1-p_{2K-2})\}$ and $\mathcal{T}' = \{(1-p_0, p_0), \ldots, (1-p_{2K-2}, p_{2K-2})\}$. We give an illustration of the construction in Figure 1. Note that $\mathcal{T}$ and $\mathcal{T}'$ are distinct (e.g., $(1,0) \in \mathcal{T}$ and $(1,0) \notin \mathcal{T}'$). Moreover, $\mathcal{T}$ and $\mathcal{T}'$ are multisets — in fact, since $p_i = p_{2K-1-i}$, for $i = 1, \ldots, K-1$, and $p_0 > p_1 > \cdots > p_{K-1}$, each of the multi-sets $\mathcal{T}$ and $\mathcal{T}'$ has K distinct elements: in each multi-set, all elements have multiplicity 2, except for one of multiplicity 1. Hence, the single-topic model $\mathcal{S}^{\mathcal{T}}$ on topics $\mathcal{T}$ (equivalently, $\mathcal{S}^{\mathcal{T}'}$ on $\mathcal{T}'$) can be either seen as a *uniform* single-topic model on $2K-1$ (non-distinct) topics, or as a *quasi-uniform* single-topic model on K distinct topics where all topics but one have the same weight, and the remaining has half of that weight.

By Observation 1, in order to show that $\mathcal{S}^{\mathcal{T}}_\ell = \mathcal{S}^{\mathcal{T}'}_\ell$ for all $\ell < 2K-1$, it is enough to do so for $\ell = 2K-2$. To this aim, we define a "limit" uniform single-topic model $\mathcal{S}^\infty$ by which documents are generated as follows: an angle γ is chosen uniformly at random in $[0, \pi)$, after which words are picked from $\mathcal{V} = [2]$ by the topic $(\cos^2(\gamma), \sin^2(\gamma))$. More precisely, the probability under $\mathcal{S}^\infty$ of a

given document of length ℓ containing $0 \leq z \leq \ell$ occurrences of word 1 is equal to:

$$\mathcal{S}_\ell^\infty(z, \ell - z) = \int_0^\pi \frac{1}{\pi} \cos^{2z}(\gamma) \sin^{2(\ell-z)}(\gamma)\, d\gamma.$$

We then have

$$\mathcal{S}_\ell^\infty(z, \ell - z) = \frac{2}{\pi} \cdot \int_0^{\pi/2} \cos^{2z}(\gamma) \sin^{2(\ell-z)}(\gamma)\, d\gamma = \frac{1}{\pi} \cdot \mathrm{B}\left(z + \frac{1}{2}, \ell - z + \frac{1}{2}\right),$$

where $\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$ is the Beta function [25, §12.4], with the last step following from one of its properties [25, §12.42]. Since it will be useful later, we decompose the Beta function [25, §12.41] in the $\mathcal{S}_\ell^\infty(z, \ell - z)$ expression, to get

$$\mathcal{S}_\ell^\infty(z, \ell - z) = \frac{1}{\pi} \cdot \mathrm{B}\left(z + \frac{1}{2}, \ell - z + \frac{1}{2}\right) = \frac{1}{\pi} \cdot \frac{\Gamma\left(z + \frac{1}{2}\right) \cdot \Gamma\left(\ell - z + \frac{1}{2}\right)}{\Gamma\left(\ell + 1\right)},$$

where Γ is the Gamma function [25, §12.1], defined as $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$. By symmetry we then have $\mathcal{S}_\ell^\infty(z, \ell - z) = \mathcal{S}_\ell^\infty(\ell - z, z)$.

We aim to show that the three worlds $\mathcal{S}_\ell^\infty$, $\mathcal{S}_\ell^{\mathcal{T}}$, $\mathcal{S}_\ell^{\mathcal{T}'}$ induce the very same document distribution at length $\ell = 2K - 2$, and suddenly become pairwise distinct at length $2K - 1$. We will make use of the following trigonometric identity by Merca [**?**] (who observes that the first identity follows from a result of Ramus [**?**], and was also proved independently by Quoniam and Greening [**?**]):

Lemma 19 ([**?**, **?**, **?**])**.** *Let $p < n$ be non-negative integers. Then,*

$$\sum_{i=0}^{n-1} \cos^{2p}\left(\frac{i}{n} \cdot \pi\right) = \frac{n}{4^p} \cdot \binom{2p}{p}.$$

If $p = n$, then

$$\sum_{i=0}^{n-1} \cos^{2p}\left(\frac{i}{n} \cdot \pi\right) = \frac{n}{4^p} \cdot \binom{2p}{p} + \frac{2n}{4^n}.$$

We are now ready to prove the equivalence of the three document distributions.

Theorem 20. *For any $1 \leq \ell \leq 2K - 2$, $\mathcal{S}_\ell^{\mathcal{T}} = \mathcal{S}_\ell^\infty = \mathcal{S}_\ell^{\mathcal{T}'}$.*

Proof. Let z be any integer in $\{0, \ldots, \ell\}$ and consider the probability that $\mathcal{S}_\ell^{\mathcal{T}}$ gives to a given document of ℓ words containing z occurrences of the word 1. We have,

$$\mathcal{S}_\ell^{\mathcal{T}}(z, \ell - z) = \frac{1}{2K - 1} \sum_{i=0}^{2K-2} \left(p_i^z \cdot (1 - p_i)^{\ell - z}\right).$$

Observe that the right-hand side is also equal to the probability that $\mathcal{S}_\ell^{\mathcal{T}'}$ gives to a given document of ℓ words that contains $\ell - z$ occurrences of the word 1. That is, $\mathcal{S}_\ell^{\mathcal{T}}(z, \ell - z) = \mathcal{S}_\ell^{\mathcal{T}'}(\ell - z, z)$ for every $0 \leq z \leq \ell$. The crux of the proof is to show that $\mathcal{S}_\ell^{\mathcal{T}}(z, \ell - z) = \mathcal{S}_\ell^\infty(z, \ell - z)$ which, by the previous equation and the symmetry property $\mathcal{S}_\ell^\infty(z, \ell - z) = \mathcal{S}_\ell^\infty(\ell - z, z)$, implies that

$\mathcal{S}_\ell^{\mathcal{T}}(z,\ell-z) = \mathcal{S}_\ell^{\mathcal{T}'}(z,\ell-z)$. We start by expanding $\mathcal{S}_\ell^{\mathcal{T}}(z,\ell-z)$.

$$\begin{aligned}
\mathcal{S}_\ell^{\mathcal{T}}(z,\ell-z) &= \frac{1}{2K-1}\sum_{i=0}^{2K-2}\left(\cos^{2z}\gamma_i\cdot\left(1-\cos^2\gamma_i\right)^{\ell-z}\right)\\
&= \frac{1}{2K-1}\sum_{i=0}^{2K-2}\left(\cos^{2z}\gamma_i\cdot\sum_{j=0}^{\ell-z}\left(\binom{\ell-z}{j}(-1)^j\cos^{2j}\gamma_i\right)\right)\\
&= \frac{1}{2K-1}\sum_{j=0}^{\ell-z}\left(\binom{\ell-z}{j}(-1)^j\sum_{i=0}^{2K-2}\cos^{2z+2j}\gamma_i\right).
\end{aligned}$$

Observe that the angles γ_i in the internal sum of our expression are of the form $\gamma_i = \pi\cdot i/2K-1$ and that i ranges from 0 to $2K-2$. Moreover, the exponents of the cosines are all the even integers from $2z$ to $2\ell < 2(2K-1) = 2n$. We can then apply Lemma 19 with $n = 2K-1$ and $p = z+j < n$, to obtain

$$\mathcal{S}_\ell^{\mathcal{T}}(z,\ell-z) = \frac{1}{2K-1}\sum_{j=0}^{\ell-z}\binom{\ell-z}{j}(-1)^j\frac{2K-1}{4^{z+j}}\binom{2z+2j}{z+j} = \sum_{j=0}^{\ell-z}\binom{\ell-z}{j}(-1)^j\frac{\binom{2z+2j}{z+j}}{4^{z+j}}.$$

We now need the following fact.

Observation 21. *If k is a non-negative integer then $4^{-k}\cdot\binom{2k}{k} = \pi^{-1/2}\cdot\frac{\Gamma\left(k+\frac{1}{2}\right)}{\Gamma(k+1)}$.*

Proof. We proceed by induction. Recall that $\Gamma(k+1/2) = (k-1/2)\cdot(k-3/2)\cdots(1/2)\cdot\sqrt{\pi}$. Then, the claim is true for $k=0$. Now, suppose it holds for k, that is:

$$4^{-k}\cdot\frac{(2k)\cdot(2k-1)\cdots(k+1)}{k!} = \frac{(k-1/2)\cdot(k-3/2)\cdots(1/2)}{k!},$$

we multiply both sides by $\frac{2k+1}{2\cdot(k+1)}$, to get:

$$\begin{aligned}
\frac{2k+1}{2\cdot(k+1)}\cdot 4^{-k}\cdot\frac{(2k)\cdot(2k-1)\cdots(k+1)}{k!} &= \frac{2k+1}{2\cdot(k+1)}\cdot\frac{(k-1/2)\cdot(k-3/2)\cdots(1/2)}{k!}\\
4^{-k-1}\cdot\frac{(2k+2)\cdot(2k+1)\cdots(k+2)}{(k+1)!} &= \frac{(k+1/2)\cdot(k-1/2)\cdot(k-3/2)\cdots(1/2)}{(k+1)!},
\end{aligned}$$

so the claim holds for $k+1$, as well. □

Using the last observation, we get,

$$\mathcal{S}_\ell^{\mathcal{T}}(z,\ell-z) = \pi^{-1/2}\cdot\sum_{j=0}^{\ell-z}\binom{\ell-z}{j}(-1)^j\cdot\frac{\Gamma\left(z+j+\frac{1}{2}\right)}{\Gamma(z+j+1)}.$$

Since $\Gamma(x+1) = x\cdot\Gamma(x)$ (see, for instance, [25, §12.12]), we can collect the common terms of the

addends,

$$\begin{aligned}\mathcal{S}^{\mathcal{T}}_{\ell}(z,\ell-z) &= \pi^{-1/2}\cdot\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma\left(z+1\right)}\cdot\sum_{j=0}^{\ell-z}\binom{\ell-z}{j}(-1)^j\cdot\frac{\prod_{k=0}^{j-1}\left(z+\frac{1}{2}+k\right)}{\prod_{k=0}^{j-1}\left(z+1+k\right)}\\ &= \pi^{-1/2}\cdot\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma\left(z+1\right)}\cdot\sum_{j=0}^{\ell-z}\frac{\prod_{k=0}^{j-1}\left(\ell-z-k\right)}{j!}\cdot(-1)^j\cdot\frac{\prod_{k=0}^{j-1}\left(z+\frac{1}{2}+k\right)}{\prod_{k=0}^{j-1}\left(z+1+k\right)}\\ &= \pi^{-1/2}\cdot\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma\left(z+1\right)}\cdot\sum_{j=0}^{\ell-z}\frac{\prod_{k=0}^{j-1}\left(z-\ell+k\right)\cdot\prod_{k=0}^{j-1}\left(z+\frac{1}{2}+k\right)}{\prod_{k=0}^{j-1}\left(z+1+k\right)}\cdot\frac{1}{j!}.\end{aligned}$$

The value of the first product of the last expression is equal to 0, for every $j \geq \ell - z + 1$, which implies that the terms of the sum are equal to 0 for $j \geq \ell - z + 1$. Thus, we can then extend the sum to all the non-negative integers, as follows,

$$\begin{aligned}\mathcal{S}^{\mathcal{T}}_{\ell}(z,\ell-z) &= \pi^{-1/2}\cdot\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma\left(z+1\right)}\cdot\sum_{j=0}^{\infty}\frac{\prod_{k=0}^{j-1}\left(z-\ell+k\right)\cdot\prod_{k=0}^{j-1}\left(z+\frac{1}{2}+k\right)}{\prod_{k=0}^{j-1}\left(z+1+k\right)}\cdot\frac{1}{j!}\\ &= \pi^{-1/2}\cdot\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma\left(z+1\right)}\cdot\ {}_2F_1\left(z-\ell,z+\frac{1}{2};z+1;1\right)\end{aligned}$$

where ${}_2F_1(a,b;c;z)$ is the hypergeometric series [25, §14.1]. We now apply Gauss's Hypergeometric Theorem (see, for instance, [25, §14.11]), which states

Theorem 22 (Gauss's Hypergeometric Theorem)**.** *Let $c > a + b$. Then,*

$${}_2F_1(a,b;c;1) = \frac{\Gamma(c)\cdot\Gamma(c-a-b)}{\Gamma(c-a)\cdot\Gamma(c-b)}.$$

In our case the condition of the theorem is satisfied since $z - \ell \leq 0$, and thus $z + 1 > z + \frac{1}{2} \geq (z-\ell) + (z+\frac{1}{2})$. Then,

$$\mathcal{S}^{\mathcal{T}}_{\ell}(z,\ell-z) = \pi^{-1/2}\cdot\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma\left(z+1\right)}\cdot\frac{\Gamma(z+1)\cdot\Gamma(\ell-z+\frac{1}{2})}{\Gamma(\ell+1)\cdot\Gamma(\frac{1}{2})}$$

By using the identity $\Gamma(1/2) = \sqrt{\pi}$, and by canceling out the terms $\Gamma(z+1)$, we obtain

$$\mathcal{S}^{\mathcal{T}}_{\ell}(z,\ell-z) = \frac{\Gamma(z+\frac{1}{2})\cdot\Gamma(\ell-z+\frac{1}{2})}{\pi\cdot\Gamma(\ell+1)} = \mathcal{S}^{\infty}_{\ell}(z,\ell-z),$$

which holds for any $0 \leq z \leq \ell$, and for every $1 \leq \ell \leq 2K-2$. The claim follows. □

We now show that the three sets of topics become distinguishable as soon as the document length reaches $2K-1$.

Theorem 23. *$\mathcal{S}^{\mathcal{T}}_{2K-1}$, $\mathcal{S}^{\mathcal{T}'}_{2K-1}$, and $\mathcal{S}^{\infty}_{2K-1}$ are pairwise different. In particular, for $0 \leq z \leq 2K-1$,*

$$\begin{aligned}\mathcal{S}^{\mathcal{T}}_{2K-1}(z,\ell-z) &= \mathcal{S}^{\infty}_{2K-1}(z,\ell-z) - (-1)^z\cdot 2^{3-4K},\\ \mathcal{S}^{\mathcal{T}'}_{2K-1}(z,\ell-z) &= \mathcal{S}^{\infty}_{2K-1}(z,\ell-z) + (-1)^z\cdot 2^{3-4K}.\end{aligned}$$

Thus, the ℓ_1-distance between $\mathcal{S}^{\mathcal{T}}_{2K-1}$ and $\mathcal{S}^{\mathcal{T}'}_{2K-1}$ is equal to $\left|\mathcal{S}^{\mathcal{T}}_{2K-1} - \mathcal{S}^{\mathcal{T}'}_{2K-1}\right|_1 = 2^{3-2K}$.

Proof. We proceed along the lines of the proof of Theorem 20 except that, this time, $\ell = 2K-1$. We get,

$$\begin{aligned}
\mathcal{S}_\ell^{\mathcal{T}}(z,\ell-z) &= \mathcal{S}_\ell^{\mathcal{T}'}(\ell-z,z) \\
&= \frac{1}{2K-1}\sum_{i=0}^{2K-2}\left(\cos^{2z}\gamma_i\cdot\left(1-\cos^2\gamma_i\right)^{\ell-z}\right) \\
&= \frac{1}{2K-1}\sum_{j=0}^{\ell-z}\left(\binom{\ell-z}{j}(-1)^j\sum_{i=0}^{2K-2}\cos^{2z+2j}\gamma_i\right) \\
&= \frac{1}{2K-1}\sum_{j=0}^{\ell-z}\left(\binom{\ell-z}{j}(-1)^j\frac{2K-1}{4^{z+j}}\binom{2z+2j}{z+j}\right)+\frac{(-1)^{\ell-z}}{2K-1}\cdot\frac{2\cdot(2K-1)}{4^\ell} \\
&= \sum_{j=0}^{\ell-z}\left(\binom{\ell-z}{j}(-1)^j\frac{1}{4^{z+j}}\binom{2z+2j}{z+j}\right)+(-1)^{\ell-z}\cdot\frac{2}{4^\ell} \\
&= \pi^{-1/2}\cdot\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma(z+1)}\cdot{}_2F_1\left(z-\ell,z+\frac{1}{2};z+1;1\right)+(-1)^{\ell-z}\cdot\frac{2}{4^\ell} \\
&= \pi^{-1}\cdot\frac{\Gamma\left(z+\frac{1}{2}\right)\cdot\Gamma\left(\ell-z+\frac{1}{2}\right)}{\Gamma(\ell+1)}+(-1)^{\ell-z}\cdot\frac{2}{4^\ell} \\
&= \mathcal{S}_\ell^{\infty}(z,\ell-z)+(-1)^{\ell-z}\cdot 2^{3-4K}.
\end{aligned}$$

Since ℓ is odd, we have that $\ell - z$ is odd iff z is even. Then,

$$\mathcal{S}_\ell^{\mathcal{T}}(z,\ell-z) = \mathcal{S}_\ell^{\infty}(z,\ell-z)-(-1)^z\cdot 2^{3-4K}$$

and $\mathcal{S}_\ell^{\mathcal{T}'}(z,\ell-z) = \mathcal{S}_\ell^{\infty}(\ell-z,z)+(-1)^z\cdot 2^{3-4K} = \mathcal{S}_\ell^{\infty}(z,\ell-z)+(-1)^z\cdot 2^{3-4K}$. Thus,

$$\left|\mathcal{S}_\ell^{\mathcal{T}}(z,\ell-z)-\mathcal{S}_\ell^{\mathcal{T}'}(z,\ell-z)\right| = 2^{4-4K},$$

and

$$\left|\mathcal{S}_\ell^{\mathcal{T}}-\mathcal{S}_\ell^{\mathcal{T}'}\right|_1 = 2^\ell\cdot 2^{4-4K} = 2^{3-2K}. \quad \square$$

We observe that the same proof shows that the ℓ_p distance of the two distributions is equal to

$$\left|\mathcal{S}_{2K-1}^{\mathcal{T}}-\mathcal{S}_{2K-1}^{\mathcal{T}'}\right|_p = \sqrt[p]{2^{2K-1}\cdot 2^{(4-4K)\cdot p}} = 2^{(2K-1)/p+(4-4K)},$$

thus, in particular, $\left|\mathcal{S}_{2K-1}^{\mathcal{T}}-\mathcal{S}_{2K-1}^{\mathcal{T}'}\right|_\infty = 16^{1-K}$.

7.1 Bounds for Quasi-Uniform Adversarial Single Topic Identifiability

In this Section, we derive upper and lower identifiability bounds for quasi-uniform Single Topic identifiability, with adversarial topics.

Corollary 24 (Adversarial Binary Quasi-Uniform Single-Topic Unidentifiability)**.** *Let $m = 2$, $\mathcal{V} = [m]$ and let the documents be generated by a quasi-uniform single-topic model $\mathcal{S}^{\mathcal{T}}$ on a set $\mathcal{T}$ of K adversarial topics. Then, no algorithm can reconstruct $\mathcal{T}$ from m, K and $\mathcal{S}_1^{\mathcal{T}},\ldots,\mathcal{S}_{2K-2}^{\mathcal{T}}$.*

Theorem 23 also gives a lower bound on the minimum ℓ_1-error that any algorithm has to incur in guessing the distribution $\mathcal{S}^{\mathcal{T}}_{2K-1}$, while only observing $\mathcal{S}^{\mathcal{T}}_1, \ldots, \mathcal{S}^{\mathcal{T}}_{2K-2}$:

Corollary 25. *Let $m = 2$, $\mathcal{V} = [m]$ and let the documents be generated by a quasi-uniform single-topic model $\mathcal{S}^{\mathcal{T}}$ on a set $\mathcal{T}$ of K adversarial topics. Then, no algorithm can approximate $\mathcal{S}^{\mathcal{T}}_{2K-1}$ from m, K and $\mathcal{S}^{\mathcal{T}}_1, \ldots, \mathcal{S}^{\mathcal{T}}_{2K-2}$, to within an ℓ_1-error smaller than 2^{2-2K}.*

The following corollary states that, if no upper bound of the number of topics is given, then no algorithm is able to prove that a given set of topics corresponds to the hidden one.

Corollary 26. *Let $m = 2$ and let the documents be generated by a quasi-uniform (resp., uniform) single-topic model $\mathcal{S}^{\mathcal{T}}$ on a set $\mathcal{T}$ of some unknown number of adversarial topics. Then, for each $\ell \geq 0$, no algorithm that knows m, and that observes $\mathcal{S}^{\mathcal{T}}_1, \ldots, \mathcal{S}^{\mathcal{T}}_\ell$, can prove that a given set of topics equals the set of unknown topics.*

7.2 Tightness of the Construction

We now prove the extremality of the finite construction in Section 7 — we will show that, if two bag-of-words models on the dictionary $\mathcal{V} = \{1, 2\}$ induce the same document distributions up to length $\ell - 1$, then the maximum ℓ_1 distance between the document distributions of length ℓ that they induce, is $2^{2-\ell}$. In our construction that distance is exactly $2^{2-\ell}$ — the finite construction (and Theorem 23) has $\ell = 2K - 1$, and it is therefore tight.

Recall that the construction of Section 7 was of the quasi-uniform Single-Topic type and contained exactly K topics — the tightness result in this section, instead, holds for arbitrary bag-of-words models (thus, in particular, it does not require a bounded number of topics).

We start with a technical lemma, that bounds the difference of the ℓ-th moments of two random variables supported on $[0, 1]$, and having the same first $\ell - 1$ moments:

Lemma 27. *Let A and B be two numerical random variables satisfying $0 \leq A, B \leq 1$. Let $\ell \geq 1$ and suppose that, for each $i \in \{0, 1, \ldots, \ell - 1\}$, the i-th moments of A and B coincide, that is, $E\left[A^i - B^i\right] = 0$. Then, $E\left[A^\ell - B^\ell\right] \leq 2^{2-2\ell}$.*

Proof. Let $T_{2\ell}(x)$ be the (2ℓ)th Chebyshev polynomial of the first kind [**?**]. Then, (i) $T_{2\ell}(x)$ is a polynomial of degree 2ℓ, (ii) the coefficient of $x^{2\ell}$ in $T_{2\ell}(x)$ is $2^{2\ell-1}$, (iii) the maximum absolute value of $T_{2\ell}(x)$ in $[0, 1]$ is 1, and (iv) all the odd powers of x have coefficient 0 in $T_{2\ell}(x)$ (see [**?**, §1.1-1.2]). Then, $P_\ell(x) = 2^{1-2\ell} \cdot T_{2\ell}\left(\sqrt{x}\right)$ is (i) a polynomial in x having degree ℓ, (ii) having its x^ℓ coefficient equal to 1, and (iii) having maximum absolute value in $[0, 1]$ equal to $2^{1-2\ell}$.

If we define $Q_{\ell-1}(x) = x^\ell - P_\ell(x)$, we have that $Q_{\ell-1}$ is a polynomial in x of degree at most $\ell - 1$, and such that $\left|x^\ell - Q_{\ell-1}(x)\right| = |P_\ell(x)| \leq 2^{1-2\ell}$, for $x \in [0, 1]$. Then,

$$\begin{aligned}
E\left[A^\ell - B^\ell\right] &= E\left[\left(A^\ell - Q_{\ell-1}(A) + Q_{\ell-1}(A)\right) - \left(Q_{\ell-1}(B) + B^\ell - Q_{\ell-1}(B)\right)\right] \\
&\leq E\left[\left|A^\ell - Q_{\ell-1}(A)\right| + Q_{\ell-1}(A) - Q_{\ell-1}(B) + \left|-B^\ell + Q_{\ell-1}(B)\right|\right] \\
&= E\left[Q_{\ell-1}(A) - Q_{\ell-1}(B) + |P_\ell(A)| + |P_\ell(B)|\right] \\
&\leq E\left[Q_{\ell-1}(A) - Q_{\ell-1}(B) + 2^{2-2\ell}\right] \\
&= E\left[Q_{\ell-1}(A) - Q_{\ell-1}(B)\right] + 2^{2-2\ell},
\end{aligned}$$

now, without loss of generality, let $Q_{\ell-1}(x) = \sum_{i=0}^{\ell-1} \left(c_{i,\ell-1} \cdot x^i\right)$. Then,

$$\begin{aligned} E\left[A^\ell - B^\ell\right] &\leq E\left[Q_{\ell-1}(A) - Q_{\ell-1}(B)\right] + 2^{2-2\ell} \\ &= \sum_{i=0}^{\ell-1} \left(c_{i,\ell-1} \cdot E[A^i - B^i]\right) + 2^{2-2\ell} \\ &= 2^{2-2\ell}. \end{aligned}$$

□

In the statement of the main theorem of this Section, we use $\mathcal{A}$ and $\mathcal{B}$ to represent to bag-of-words models on the vocabulary $\mathcal{V}$. As usual, we will use $\mathcal{A}_\ell$ and $\mathcal{B}_\ell$ to denote the probability distributions induced, respectively, by $\mathcal{A}$ and $\mathcal{B}$ on the documents of $\mathcal{V}^\ell$.

Theorem 28. *Let $\mathcal{A}$ and $\mathcal{B}$ be two bag-of-words models on the vocabulary $\mathcal{V} = \{1, 2\}$. If, for a given $\ell \geq 1$, it holds $\mathcal{A}_{\ell-1} = \mathcal{B}_{\ell-1}$, then, for each $0 \leq z \leq \ell$, it must hold $|\mathcal{A}_\ell(z, \ell - z) - \mathcal{B}_\ell(z, \ell - z)| \leq 2^{2-2\ell}$ — thus, $|\mathcal{A}_\ell - \mathcal{B}_\ell|_1 \leq 2^{2-\ell}$.*

Proof. First of all, observe that $\mathcal{A}_{\ell-1} = \mathcal{B}_{\ell-1}$ implies $\mathcal{A}_i = \mathcal{B}_i$ for each $i = 0, 1, \ldots, \ell - 1$. We apply Observation 8 to compute the difference between $\mathcal{A}_\ell(z, \ell - z)$ and $\mathcal{B}_\ell(z, \ell - z)$, for $0 \leq z \leq \ell$,

$$\mathcal{A}_\ell(z, \ell - z) - \mathcal{B}_\ell(z, \ell - z) = \sum_{i=0}^{\ell-z} \left(\binom{\ell - z}{i} (-1)^i \cdot (\mathcal{A}_{z+i}(z + i, 0) - \mathcal{B}_{z+i}(z + i, 0)) \right).$$

If $z + i < \ell$, we have by assumption that $\mathcal{A}_{z+i} = \mathcal{B}_{z+i}$. Thus, all the terms of the sum indexed by $i < \ell - z$ evaluate to zero — we remove them, to get:

$$\mathcal{A}_\ell(z, \ell - z) - \mathcal{B}_\ell(z, \ell - z) = (-1)^{\ell-z} \left(\mathcal{A}_\ell(\ell, 0) - \mathcal{B}_\ell(\ell, 0)\right).$$

That is, for each $0 \leq z \leq \ell$, we have that $|\mathcal{A}_\ell(z, \ell - z) - \mathcal{B}_\ell(z, \ell - z)| = |\mathcal{A}_\ell(\ell, 0) - \mathcal{B}_\ell(\ell, 0)|$.

Let us, again, consider the random variable p (resp., q) that is equal to the probability given to the word 1 by a topic randomly chosen by $\mathcal{A}$ (resp., $\mathcal{B}$). We have that $0 \leq p, q \leq 1$, that $\mathcal{A}_i(i, 0) = E[p^i]$, and that $\mathcal{B}_i(i, 0) = E[q^i]$. Since, for $0 \leq i \leq \ell - 1$, we have $\mathcal{A}_i(i, 0) = \mathcal{B}_i(i, 0)$, we also have, for the same range of i, that $E[p^i - q^i] = 0$. Lemma 27, applied to p and q, guarantees that $E[p^\ell - q^\ell] \leq 2^{2-2\ell}$ and, thus, $|\mathcal{A}_\ell(\ell, 0) - \mathcal{B}_\ell(\ell, 0)| \leq 2^{2-2\ell}$; it follows that $|\mathcal{A}_\ell(z, \ell - z) - \mathcal{B}_\ell(z, \ell - z)| \leq 2^{2-2\ell}$ for each $0 \leq z \leq \ell$, and that $|\mathcal{A}_\ell - \mathcal{B}_\ell|_1 \leq 2^{2-\ell}$. □

References

[1] D. Alvarez-Melis and M. Saveski. Topic modeling in twitter: Aggregating tweets by conversations. In *ICWSM*, 2016.

[2] A. Anandkumar, D. P. Foster, D. J. Hsu, S. M. Kakade, and Y.-K. Liu. A spectral algorithm for latent dirichlet allocation. *Algorithmica*, 72(1):193–214, 2015.

[3] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[4] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pages 33–1, 2012.

[5] A. Anandkumar, D. J. Hsu, M. Janzamin, and S. M. Kakade. When are overcomplete topic models identifiable? uniqueness of tensor tucker decompositions with structured sparsity. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *NIPS*, pages 1986–1994, 2013.

[6] S. Arora, R. Ge, Y. Halpern, D. M. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *ICML (2)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 280–288. JMLR.org, 2013.

[7] S. Arora, R. Ge, and A. Moitra. Learning topic models - going beyond SVD. In *FOCS*, pages 1–10. IEEE Computer Society, 2012.

[8] N. Balakrishnan and V. Nevzorov. *A Primer on Statistical Distributions*. Wiley, 2004.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *Journal of Machine Learning Research*, 2003.

[10] D. L. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. K. Saul, and B. Schlkopf, editors, *NIPS*, pages 1141–1148. MIT Press, 2003.

[11] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2nd edition, 1994.

[12] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April 2004.

[13] M. Hajjem and C. Latiri. Combining IR and LDA topic modeling for filtering microblogs. In C. Zanni-Merk, C. S. Frydman, C. Toro, Y. Hicks, R. J. Howlett, and L. C. Jain, editors, *KES*, volume 112 of *Procedia Computer Science*, pages 761–770. Elsevier, 2017.

[14] T. Hofmann. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In S. A. Solla, T. K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 914–920. MIT Press, 2000.

[15] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.

[16] K. Huang, X. Fu, and N. D. Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *NIPS*, pages 1786–1794, 2016.

[17] S. Krantz and H. Parks. *The Implicit Function Theorem: History, Theory, and Applications*. The Implicit Function Theorem: History, Theory, and Applications. Birkhäuser, 2002.

[18] C. Li, H. Wang, Z. Zhang, A. Sun, and Z. Ma. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 165–174, New York, NY, USA, 2016. ACM.

[19] I. Macdonald. *Symmetric Functions and Hall Polynomials*. Oxford classic texts in the physical sciences. Clarendon Press, 1998.

[20] K. Nigam, A. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134, 2000.

[21] C. H. Papadimitriou, H. Tamaki, P. Raghavan, and S. Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168. ACM, 1998.

[22] V. K. R. Sridhar. Unsupervised topic modeling for short texts using distributed representations of words. In P. Blunsom, S. B. Cohen, P. S. Dhillon, and P. Liang, editors, *VS@HLT-NAACL*, pages 192–200. The Association for Computational Linguistics, 2015.

[23] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

[24] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.

[25] E. Whittaker and G. Watson. *A Course of Modern Analysis*. The University Press, 1920.

[26] X. Yan, J. Guo, Y. Lan, and X. Cheng. A biterm topic model for short texts. In *WWW*, 2013.

[27] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.

A LDA and Single-Topic Equivalence: the Role of α

The reduction we gave in Section 4 assumes knowledge of the LDA parameter α (more precisely, $K \cdot \alpha$ must be known, as stated in Corollary 6). Here we show that this is needed: specifically, we will present two LDA models, each with its own α parameter, on $K = 2$ topics and $m = 2$ words each, that are indistinguishable up to document length 2.

Let $K = m = 2$ with the two topics defined as $\mathcal{T}_x = \{(x, 1-x), (1-x, x)\}$, for some $0 \le x \le 1$. Then, for any $\alpha > 0$, by the Dirichlet symmetry and the topics symmetry, we have $\mathcal{D}_1^{\mathcal{T}_x,\alpha}(1,0) = \mathcal{D}_1^{\mathcal{T}_x,\alpha}(0,1) = \frac{1}{2}$.

Moreover, for any $\alpha > 0$, by Theorem 5, we have

$$\begin{aligned}
\mathcal{D}_2^{\mathcal{T}_x,\alpha}(1,1) &= \frac{\Gamma(2\alpha)}{\Gamma(2\alpha+2)} \cdot \Big((2\alpha)^2 \cdot \mathcal{S}_1^{\mathcal{T}_x}(1,0) \cdot \mathcal{S}_1^{\mathcal{T}_x}(0,1) + (2\alpha) \cdot \mathcal{S}_2^{\mathcal{T}_x}(1,1)\Big) \\
&= \frac{\Gamma(2\alpha)}{\Gamma(2\alpha+2)} \cdot \big(\alpha^2 \cdot (x + (1-x)) \cdot ((1-x) + x) + \alpha \cdot (x \cdot (1-x) + (1-x) \cdot x)\big) \\
&= \frac{1}{2\alpha \cdot (2\alpha+1)} \cdot \big(\alpha^2 + 2\alpha \cdot x \cdot (1-x)\big) \\
&= \frac{\alpha + 2 \cdot x(1-x)}{4\alpha+2}.
\end{aligned}$$

Again by symmetry, we will have,

$$\mathcal{D}_2^{\mathcal{T}_x,\alpha}(2,0) = \mathcal{D}_2^{\mathcal{T}_x,\alpha}(0,2) = \frac{1 - 2 \cdot \mathcal{D}_2^{\mathcal{T}_x,\alpha}(1,1)}{2}.$$

Now, it is easy to see that we can get the same distribution over documents by varying x and α. Fix $\delta \in \left(0, \frac{1}{4}\right)$; then, pick arbitrarily $\alpha \in \left(0, \frac{2\delta}{1-4\delta}\right)$ and set $x = \frac{1}{2} \cdot \left(1 - \sqrt{(1-4\delta) \cdot (1+2\alpha)}\right)$: with this choice, we get $\mathcal{D}_2^{\mathcal{T}_x,\alpha}(1,1) = \delta$, $\alpha > 0$, and $0 < x < \frac{1}{2}$ — in fact, observe that, after having fixed δ, the variables α and x are bijectively related (with x decreasing as α increases).

As a concrete example, fix $\delta = 35/144$ and pick, on the one hand, $\alpha_1 = 4$ and $x_1 = 1/4$; on the other, pick $\alpha_2 = 12$ and $x_2 = 1/12$. Then, $\mathcal{D}_2^{\mathcal{T}_{x_1},\alpha_1}(1,1) = \mathcal{D}_2^{\mathcal{T}_{x_2},\alpha_2}(1,1) = \delta$ and, thus, the two alternatives give exactly the same distributions $\mathcal{D}_\ell^{\mathcal{T}_x,\alpha}$ for each $\ell \leq 2 = K$.

Since the two sets of topics $\mathcal{T}_{1/4}$ and $\mathcal{T}_{1/12}$ are different, one cannot reconstruct the unknown topics by only accessing $\mathcal{D}_K^{\mathcal{T}_x,\alpha}$ and K, with no knowledge of α. On the other hand, by Corollary 16, $\mathcal{S}_K^{\mathcal{T}_x}$ is sufficient to reconstruct the unknown topics.