Hierarchical Mixture Models: a Probabilistic Analysis

Mark Sandler
Google, Inc
sandler@google.com

ABSTRACT

Mixture models form one of the most widely used classes of generative models for describing structured and clustered data. In this paper we develop a new approach for the analysis of hierarchical mixture models. More specifically, using a text clustering problem as a motivation, we describe a natural generative process that creates a hierarchical mixture model for the data. In this process, an adversary starts with an arbitrary base distribution and then builds a topic hierarchy via some evolutionary process, where he controls the parameters of the process. We prove that under our assumptions, given a subset of topics that represent generalizations of one another (such as `baseball` → `sports` → `base`), for any document which was produced via some topic in this hierarchy, we can efficiently determine the most specialized topic in this subset, it still belongs to.

The quality of the classification is independent of the total number of topics in the hierarchy and our algorithm does not need to know the total number of topics in advance. Our approach also yields an algorithm for clustering and unsupervised topical tree reconstruction.

We validate our model by showing that properties predicted by our theoretical results carry over to real data. We then apply our clustering algorithm to two different datasets: (i) "20 newsgroups" [19] and (ii) a snapshot of abstracts of arXiv [2] (15 categories, ≈240,000 abstracts). In both cases our algorithm performs extremely well.

Categories and Subject Descriptors: F.2.2: Nonnumerical Algorithms and Problems, H.1.0: Information Systems: Models and Principles

General Terms: Algorithms, theory

Keywords: Mixture Models, probabilistic analysis, hierarchical clustering

1 Introduction

Mixture models form one of the most widely used classes of generative models for describing structured and clustered data [18]. Various applications of mixture models include problems in computer vision [27, 13, 24], text clustering [22, 4], collaborative filtering [12, 15, 4], and bioinformatics [14]. One of the common formulations of mixture models can be described as follows. The data (e.g., texts, images, user profiles, etc.) is a collection of independent samples (e.g., individual documents, image features, etc.), each created by a sequence of independent draws from some hidden distribution over a feature space (terms, pixels, etc.). For example, in the case of text, each document is modeled as a sequence of independent trials from some underlying term distribution. More precisely, there is a set of topics, where each topic is defined by a hidden distribution over the space of all possible terms. A given document can be related to several topics (e.g., discussing the role of religion in physical sciences), and is modeled as a sample from a linear mixture (hence the name) of corresponding topical distributions. The actual mixing coefficients are defined by the document's quantitative relevance to each of the topics. A similar interpretation is possible for collaborative filtering where each user's profile (e.g. the books he bought) is modeled as a sequence of independent trials from some hidden distributions over the item space.

In this work we will be discussing mixture models in the context of text, however the results apply in any other framework.

Inspired by statistical approaches there has been a vast amount of work on mixture models [18]. Most of it is based on local search techniques, such as different flavors of EM[4, 17, 11],[1] and Naive-Bayes methods[8, 16].

Recently, Kleinberg and Sandler [15] have shown that there is a combinatorial algorithm, that, given a sufficient amount of unlabeled data, reconstructs the underlying term distributions for each document (with a small error and with high probability). Then, given the topical distributions, the algorithm reconstructs accurately the relevances to each of the topics for each document. In order to guarantee fixed accuracy, the required length of a document depends only on the total number of topics, and two special parameters introduced in [15], but not on the total number of terms in the vocabulary.

However, while providing a mathematically clean and appealing framework, plain mixture models are known to be too simplistic [10] for real data with a large number of topics.

KDD'07, August 12-15, 2007, San Jose, California, USA.

[1] It should be noted, that in the case of gaussian mixtures, and under some specific conditions, EM is known to converge to a *global* optimum [6], however no results like that are known for the models that we are interested in here.

Indeed, most, if not all, of the existing algorithms for mixture models need to reconstruct the entire mixture model and/or need to know the total number of topics in the system in advance in order to perform classification. However, even the number of topics is hard to learn in the context of the Web or any other large unmoderated text collection. To deal with this problem, hierarchical mixture models were introduced (see [10, 26, 3] among others). The common assumption behind hierarchical models is that topics form some natural hierarchy based on the level of their specificity. For instance, topics like "physics" and "mathematics" would be under a more general topic "science." The hierarchy itself might have been inferred from labeled data [26, 10], or in the case of Latent Dirichlet Allocation [4], by assuming a specific generative process called the Chinese Restaurant Process[3, 25] and then applying EM to it. To the best of our knowledge, none of the previous results would provide measurable guarantees of the quality of the produced classification and would use flavors of local search techniques to do final estimations of the parameters and/or the hierarchy. In the present work we suggest a new model of the topic hierarchy and prove that under this model we can provide classification guarantees even if only a small part of hierarchy is known.

At a high level we assume the following evolutionary process that builds the topic structure. There is a "proto-topic" (which we also sometimes call the *baseline* topic or the *root* topic) that is the root of the hierarchy tree. Each topic has zero or more children. Each child topic distribution evolves from parent distribution by altering the probabilities of occurrence of some (possibly all) terms. The exact process by which we obtain children topics from a parent topic will be described later. The very important assumption that we make here is that given a parent topic, all children topics are generated independently from each other. This assumption is inspired by the following intuition: related topics such as `mathematics` and `physics` share inherited bias for generic science terms from their parent `science` topic, but otherwise are independent. This independence property is a key assumption that enables us to guarantee classification accuracy without knowing the total number of topics and without ever having to learn the full hierarchy.

Remark: A similar concept of evolution is widely used in studying the origin of the species, and in particular for phylogenentic tree reconstruction[21]. There, the assumption is that there is an evolutionary tree of all the species, where each node contains some aggregate description of a particular species (e.g. number of legs, presence of specific gene, etc), the root contains a description of a common ancestor of all currently existing species, and the transition from the parent is modeled via a mutation process. There has been a lot of work done on phylogenetic trees including several results that guarantee close to optimal reconstruction. We refer to [1, 23, 7] for more details. Our problem is different in several aspects. First, our attributes (terms) are defined on a continuous domain and share the same normalization, whereas in phylogenetic trees they could be drawn from different domains. Second, our samples (documents) do not contain full information about the node they belong to, but rather a small sample drawn from it (whereas a single animal sample would contain all or most of the properties for a given species.) Finally, document distributions could be a mixture of distributions from several topics (in the case of a phylogenetic tree this would correspond to observing a mix of a dog and a mouse). However it would be very interesting to connect these different lines of research.

1.1 Our Contributions. There are several contributions of this paper. First, we introduce a new theoretical model to describe topical hierarchies. Combined with its empirical validation, this is an important step in the understanding of the topical structure of textual data.

Second, we prove that within this model one can compute how a given document relates to a known topic (or several topics that are generalizations of each other) with very little knowledge of the rest of the hierarchy. We are not aware of any other theoretical work that would analyze the possibility of accurate predictions without knowing the full topical structure. We also show that the accuracy is *independent* of the total number of topics in the full hierarchy; this makes the approach particularly useful when the available data contains a very large number of topics.

Third, we present an algorithm that can reconstruct a topical hierarchy by analyzing unlabeled data. One other important property of our algorithm is that it scales extremely well and most of the processing can be readily parallelized. We evaluate our algorithm by clustering two datasets: arXiv [2], and "20 newsgroups" [19], and in both cases our algorithms perform extremely well.

Finally, we demonstrate that the theoretical properties predicted by our analysis carry over to the real data, which further supports the validity of our model.

1.2 Organization of the paper. In the next section we introduce some background on Mixture Models and all the necessary tools we need for analysis. Then, in Section 3 we formally describe our model, prove the main theoretical results of the paper and describe our algorithms. Finally, in Section 4 we present our experimental results.

2 Preliminaries

2.1 Concentration bounds. The main concentration result we use is Hoeffding's inequality, which is a generalization of Chernoff's bound [20].

LEMMA 2.1 (HOEFFDING'S INEQUALITY [9]). *Suppose $\{x_i\}$ is a sequence of independent random variables, such that $x_i \in [a_i, b_i]$. Then for any $t > 0$,*

$$\Pr\left[\left|\sum x_i - \mathrm{E}\left[\sum x_i\right]\right| \geq \tau\right] \leq 2\exp[-\frac{2\tau^2}{\sum_i |b_i - a_i|^2}],$$

where $\mathrm{E}[\sum x_i]$ denote the expected value of the sum of the random variables.

2.2 Mixture models, topic independence and generalized pseudoinverses. In this section we provide a formal definition of mixture models and necessary results from earlier work [15]. We start with a brief overview of generative mixture models.

(i) There is a set $\mathcal{D}$ that contains all possible terms in the vocabulary. We let $n = |\mathcal{D}|$.

(ii) There is a set $\mathcal{T}$ of k topics, where each topic $T^{(i)}$ is defined by a probability distribution over the set

of terms. We denote the $n \times k$ matrix comprised of all topical distributions by $\mathcal{W}$. Thus $\mathcal{W}_{ui}$ denotes the probability of term u in topic i.

(iii) Each document has hidden relevance to one or more topics, and these topics are used to generate the content of the document. More specifically, a document d has a hidden vector $\mathbf{r} \in \Re_+^{\mathbf{k}}$ of non-negative relevances that sum to one. Each word in the document is an independent draw from the distribution defined by a linear combination of topics $\mathcal{W}\mathbf{r}$.

(iv) We represent a document by a count vector $\tilde{D}$ of the actual terms occurring in the document d. The goal is to approximately reconstruct the relevance vector $\mathbf{r}$ given $\tilde{D}$.

It has been shown previously[22, 15] that mixture models allow efficient algorithms to find a document's relevance to each topic. The required length of a document needed to guarantee a fixed accuracy of classification is *independent* of the size of a vocabulary. However, in order to make this guarantee the algorithms in [22, 15] require knowledge of the number of topics in advance, and also need to recover all the topics. At a high level the algorithm of [15] is as follows:

ALGORITHM 2.1 (SEMIOMNISCIENT ALGORITHM OF [15]).

Input: Topic matrix W, document vector $\tilde{D}$.
Output: Vector of approximate relevances to all the topics.

1 Compute a generalized pseudoinverse V of the topic matrix W, by solving the following linear programming problem:

$$\text{minimize } c \text{ such that } V\mathcal{W} = I, \text{ and } -c \leq V_{iu} \leq c \tag{1}$$

2 Compute the approximate relevance vector $\tilde{\mathbf{r}} = V\tilde{D}$

If $\max_{i,u} V_{iu}$ is bounded, it was proven that for any $\varepsilon > 0$ and given a sufficiently long document we have $||\tilde{\mathbf{r}} - \mathbf{r}||_\infty \leq \varepsilon$ with high probability. It was also shown that for an optimal pseudoinverse the maximal element could be upper bounded in terms of the original topic matrix $\mathcal{W}$:

$$\max_{i,u} V_{iu} \leq \frac{1}{\Gamma}, \text{ where } \Gamma = \min_{\mathbf{x} \in R^k, x \neq 0} \frac{||\mathcal{W}\mathbf{x}||_1}{||\mathbf{x}||_1}$$

It was proven in [15] that $\Gamma > 0$ is necessary for any mixture model to be fully learnable. We refer the reader to the original paper [15] for an in-depth discussion and comparison with other ways of obtaining pseudoinverses (such as using Singular Value Decomposition). We conclude this part by a simple concentration lemma.[2]

LEMMA 2.2. *Let D be an arbitrary distribution over terms, and let $R \in \Re^n$ be such that $||R||_\infty \leq c$. Let $\tilde{D}$ be a normalized count vector of a sample of size t from D (e.g. $\tilde{D}_i = \frac{1}{t}$ if term i was sampled exactly once). Then for any $\varepsilon > 0$, we have:*

$$\Pr\left[|\langle R\tilde{D}\rangle - \langle RD\rangle| \geq \varepsilon\right] \leq 2\exp\left[-\frac{t\varepsilon^2}{2c^2}\right]$$

[2]A weaker variant of this based on Chebyshev inequality was proven in [15]

Proof. We represent $\tilde{D}$ as a sum of t independent samples $\tilde{D}_i$, where $\tilde{D}_i$ is an indicator vector for the i-th term in the document. Then, $\langle R\tilde{D}\rangle = \frac{1}{t}\sum_{i=0}^{t}\langle R\tilde{D}_i\rangle$ and it is easy to see that $\mathrm{E}\left[\langle R\tilde{D}_i\rangle\right] = \langle RD\rangle$. Now we use Hoeffding's inequality (Lemma 2.1) and substituting $\langle R\tilde{D}_i\rangle$ for x_i, ε for τ and $-c/t$ and c/t for a_i and b_i respectively, we have:

$$\Pr\left[|\langle R\tilde{D}\rangle - \langle RD\rangle| \geq \varepsilon\right] \leq 2\exp\left[-\frac{t^2\varepsilon^2}{2c^2}\right] \tag{2}$$

■

3 Hierarchical Mixture Models

3.1 Model for topical hierarchy. In this section we formally describe our model. It is important to emphasize that we still use the regular mixture model as a generative model for text. However, in addition to that we introduce a hierarchical generative process on the topical distributions. In other words, we assume that there are two underlying generative processes happening: the first process generates the underlying mixture model, and this model in turn defines the parameters of the second generative process that creates all the documents. We show that with very high probability the produced mixture model will have "good" properties, and then given a "good" model we can produce a proper classification.

The topic generation is a multistep process driven by an adversary – he chooses initial parameters at will and then the random process starts by using the adversary's parameters. At the intermediate steps the adversary can analyze the current hierarchy and choose additional parameters. We show that no matter how the adversary behaves we can still provide guarantees.

The construction proceeds as follows. First the adversary chooses the baseline topic $T^{(0)}$, which will be at the root of our hierarchy tree; we don't restrict the adversary regarding how he chooses the base topic. Given the root of the tree, the adversary starts building the rest of the tree. On every topic he decides how many children it will have, and then chooses the distributions for all immediate children topics at once. Suppose topic T is a parent topic, and the adversary is choosing the parameters of a child topic T'. For each term i in a child topic, he chooses a probability distribution D'_i satisfying some specific feasibility conditions. Then, the random process samples a value $\mathcal{E}'_i$ for each D'_i and sets $T'_i \leftarrow T_i + \mathcal{E}'_i$. After the change is applied, the vector T' is normalized by a factor α' so that it represents a valid distribution. In general, for a topic $T^{(i)}$, the change vector that was applied to its parent to generate $T^{(i)}$ is denoted by $\mathcal{E}^{(i)}$ (we also sometimes refer to $\mathcal{E}^{(i)}$ is as a *mutation vector*) and the normalization constant by α_i.

At each step the adversary can expand any leaf in the constructed hierarchy. The change vector $\mathcal{E}'$ satisfies the following constraints:

(i) The resulting topic cannot contain any negative elements (as otherwise it would not be a distribution) thus for any term i , the change is bounded so that $\mathcal{E}'_i \geq -T_i$.

(ii) For any term we have $\mathrm{E}[\mathcal{E}'_i] = 0$. It is important to note that we *do not* require the change to be symmetric, and in fact it can be arbitrarily skewed toward in-

creasing probabilities as long as the expectation stays zero.

(iii) The difference between child and parent topics should be large enough: $\mathrm{E}[||\mathcal{E}'||_1] \geq \Delta$ for some constant Δ, where $\mathcal{E}'$ – is a resulting vector change to the entire distribution.

(iv) Finally, changes cannot be concentrated on just a few items. Therefore, if each $\mathcal{E}'_i$ has range $[A_i, B_i]$, we define

$$\text{slope}[\mathcal{E}'] = \frac{\mathrm{E}[||\mathcal{E}'||_1]}{||A - B||_2}$$

and require $\text{slope}[\mathcal{E}'] \geq \frac{1}{\varepsilon_2}$, where ε_2 is some constant. Note that from condition (i) we have $A_i \geq -T_i$.

The first two conditions guarantee that the resulting vector contains only positive values and that *in expectation* the resulting norm stays one.

The third condition guarantees that ancestors differ enough from the parents so that it is possible to differentiate between them.

The last condition guarantees that changes are spread among many items and deserves some additional explanation. Note the two different norms used in the condition.[3] This condition captures two properties. First, since we are looking at the maximal possible change in any probability, it guarantees absence of "low probability - very high impact" events. Or, in other words, it says that there cannot be a term that with small probability would dominate the entire vocabulary; however it is perfectly feasible to have many terms whose probability will multiply by a large factor.

Also note that by the condition (iii) the expected L_1 norm of a change is large and thus the condition (iv) essentially requires the maximal possible change in the Euclidean norm to be small. Since norms are monotone, this could potentially lead to two conflicting conditions. However, if the changes are distributed across many items (and that's what we observe in practice), the value of the Euclidean norm is smaller than the L_1 norm by an $\Omega(\sqrt{n})$ factor, and thus has $\text{slope}[\mathcal{E}'] \approx \Omega(\sqrt{n})$, for a uniform change vector.

For example, starting with a uniform parent distribution, the adversary might want to create a topic with support on half of the terms. Then, choosing a probability distribution that is $\pm\frac{1}{n}$ with probability 1/2 would produce the desired result. Note that all constraints above are satisfied and $\mathrm{E}[||\mathcal{E}||_1] = 1$, $\mathrm{E}[||B - A||_2] = \frac{2}{\sqrt{n}}$. Using a similar construction and starting with a uniform parental distribution, the adversary can produce *any* power-law distribution in a child. However, the adversary does not have control over whether the probability of any individual term will increase or decrease—only over the general shape of the distribution.

The random topic generation process might appear counterintuitive at first. However, one can argue that, if we did not know any meaning behind any term, then changes in probabilities for different topics *would* appear random and uncorrelated (given the parent distribution) among different child topics; this is exactly what our process models.

[3]A somewhat similar measure has appeared before in [5], in the context of continuous models where it was called "slope ratio" (hence our name) and it was shown that this quantity is one of the necessary parameters which measure "learnability" of their model, it is interesting that it appears in our context as well.

3.2 Analysis: Overview. Our analysis contains two major parts. In Section 3.3 we prove intuitive albeit fairly technical results. Namely, if we have a sequence of topics such that one is obtained from another by introducing random changes and re-normalizing the resulting topical vector, then these topics will stay independent in linear algebraic sense (as measured by the independence coefficient from the previous section), with high probability, provided that the length of the sequence is small compared to the number of terms in the vocabulary. It can be shown that the linear independence is necessary to be able to uniquely determine the relevance of topics to a document. In the process of doing that we also prove a few auxiliary lemmas which we will reuse later.

Then, in Section 3.4 we analyze our perturbation scheme and in particular we prove that if we consider a sequence of topics lying on the path of the hierarchy starting from the root, and consider it as a standalone mixture problem, then applying it to a document produced by a topic from outside of the path, it will get assigned to the closest node[4]

Finally in Section 3.5 we bring all the results together and present out algorithms

3.3 Connecting probabilistic and linear independence. We start with a simple lemma which provides concentration guarantees for a sum of a fixed and random vector.

LEMMA 3.1 (SUM OF A FIXED AND A RANDOM VECTOR). *Fix some ε and δ. Suppose $Z \in \Re^n$ is an arbitrary vector, and consider a random vector $\mathcal{E} = (\mathcal{E}_1, \ldots, \mathcal{E}_n)$, such that all $\mathcal{E}_i$ are independent random variables, such that $\mathrm{E}[\mathcal{E}_i] = 0$, each variable is bounded $-\beta_i \leq \mathcal{E}_i \leq \beta_i$, and the slope is high:*

$$\text{slope}[\mathcal{E}] \geq 3\frac{\sqrt{|\ln \delta|}}{\varepsilon} \tag{3}$$

then the following conditions are satisfied with probability at least $1 - \delta$:

$$\begin{array}{ll} (i) & ||Z + \mathcal{E}||_1 \geq (1 - \varepsilon)\max(||Z||_1, \frac{\mathrm{E}[||\mathcal{E}||_1]}{2}), \\ (ii) & ||\mathcal{E}||_1 \leq \frac{3}{2}\mathrm{E}[||\mathcal{E}||_1], \\ (iii) & \textit{if } \text{sgn}[Z] = \text{sgn}[Z + \mathcal{E}] \textit{ then } ||Z + \mathcal{E}||_1 \leq (1 + \varepsilon)||Z||_1. \end{array}$$

Proof. First of all we show that for all $i \in [1, n]$ we have:

$$\mathrm{E}[|Z_i + \mathcal{E}_i|] \geq \max(|Z_i|, \frac{\mathrm{E}[\mathcal{E}_i]}{2}).$$

Indeed, we immediately have

$$\mathrm{E}[|Z_i + \mathcal{E}_i|] \geq \mathrm{E}[|Z_i| + \text{sgn}[Z_i]\mathcal{E}_i] = |Z_i| \tag{4}$$

where the first transition uses the triangle inequality, and the second uses the linearity of expectation. Combining triangle inequality in a different way and using the previous lower bound we get:

$$\mathrm{E}[|Z_i + \mathcal{E}_i|] \geq \max(|Z_i|, \mathrm{E}[|\mathcal{E}_i| - |Z_i|]) \geq \mathrm{E}\left[|\frac{\mathcal{E}_i}{2}|\right]. \tag{5}$$

[4]Exactly the same result applies if a document is produced by a mixture of topics – the weight assigned to each of the node on the path would reflect the total relevance weight to the topics closest to the given nodes.

Combining (4) and (5) we immediately have

$$\mathrm{E}\left[||Z+\mathcal{E}||_1\right] \geq \max(||Z||_1, \frac{\mathrm{E}\left[||\mathcal{E}||_1\right]}{2}). \tag{6}$$

To finish the proof we apply Hoeffding's inequality to the sum of the individual components of the L_1 norm, to show desired concentration around the expectation. Indeed, using (6) we have:

$$\begin{aligned}&\Pr\left[||Z+\mathcal{E}||_1 - \max(||Z||_1, \tfrac{\mathrm{E}[||\mathcal{E}||_1]}{2}) < -t\right] \leq \\ &\Pr\left[||Z+\mathcal{E}||_1 - \mathrm{E}\left[||Z+\mathcal{E}||_1\right] < -t\right] \leq \exp - \tfrac{t^2}{2\sum_{i=1}^n \beta_i^2}\end{aligned} \tag{7}$$

choosing $t = \frac{\varepsilon \mathrm{E}[||\mathcal{E}||_1]}{2}$ we immediately have: :

$$\Pr\left[||Z+\mathcal{E}||_1 \leq (1-\varepsilon)M\right] \leq \exp - \frac{\varepsilon^2(\mathrm{E}\left[||\mathcal{E}||_1\right])^2}{8\sum \beta_i^2} \leq \delta$$

where $M = \max(||Z||_1, \frac{\mathrm{E}[||\mathcal{E}||_1]}{2})$, and in the last transition we have used our slope constraint (3) in the last transition.

The part (ii) immediately follows from yet another application of Hoeffding's inequality.

For the part (iii) of the theorem, we just note that if $\mathrm{sgn}[Z_i + \mathcal{E}_i] = \mathrm{sgn}[Z_i]$, then we have $\mathrm{E}\left[|Z_i + \mathcal{E}_i|\right] = \mathrm{E}\left[|Z_i|\right]$, and so we can use both upper and lower bounds provided by Hoeffding's inequality and the result immediately follows. ■

Remark: Somewhat surprisingly, the factor $\frac{1}{2}$ in the norm $||\mathcal{E}||_1$ in the lemma above is tight (which would not be the case if Z was random too). For example, let $Z = (\frac{1}{n}, \ldots, \frac{1}{n})$, and let

$$\mathcal{E}_i = \begin{cases} -1/n \text{ with probability } \alpha, \\ \frac{\alpha}{(1-\alpha)n} \text{ with probability } 1-\alpha \end{cases}$$

for some $\alpha \geq 1/2$ then we have

$$\mathrm{E}\left[||\mathcal{E}||_1\right] = 2\alpha$$

whereas

$$\mathrm{E}\left[||Z+\mathcal{E}||_1\right] = \sum_{i=1}^{n}(1-\alpha)(\frac{\alpha}{(1-\alpha)n} + \frac{1}{n}) = 1$$

thus if $\alpha \approx 1$, then $\mathrm{E}\left[||Z+\mathcal{E}||_1\right] \approx \frac{\mathrm{E}[\mathcal{E}]}{2}$. As n grows (and α stays constant), we can apply concentration bounds, and show that for any δ, one can choose α and n_0, so that $\forall n > n_0$ $||Z+\mathcal{E}||_1 \leq (1+\delta)\frac{||\mathcal{E}||_1}{2}$ with probability at least $1/2$.

An immediate corollary of the previous lemma, is that the normalization coefficients for the topics, are in fact very close to one.

Corollary 3.2 (Bounded normalization constants). *Consider T' with parent topic T, then if the distributions of change vector $\mathcal{E}'$ satisfy the constraints (i)-(iv) of the hierarchical process, and in particular for some ε:*

$$\mathrm{slope}\left[\mathcal{E}'\right] \geq \frac{\sqrt{\ln \delta}}{\varepsilon}$$

Then with probability at least $1-\delta$, the normalization constant $\alpha' \in [1-\varepsilon, 1+\varepsilon]$.

Proof. The result immediately follows from Lemma 3.1, since $\mathcal{E}' + T \geq 0$ and $T \geq 0$, thus $||T^{(j)} + \mathcal{E}^{(i)}||_1 \in [1 \pm \varepsilon]$ with probability at least $1-\delta$ ■

Now, we prove a more technical lemma which connects the notion of linear independence from [15, 22] with the notion of probabilistic independence. In this lemma we assume that we have a sequence of random vectors, and we prove that with high probability they will have high independence.

Lemma 3.3. *Let Z be an arbitrary vector such that $||Z||_1 = 1$, let $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(t)}$ be a sequence of random vectors. The values of $\mathcal{E}^{(i)}$ might depend on $\mathcal{E}^{(j)}$, for $j < i$, however, given the values of $\mathcal{E}^{(j)}$, the distribution of $\mathcal{E}^{(i)}$ satisfies the following constraints:*

(1) $\mathrm{E}\left[\mathcal{E}^{(i)}\right] = (0, 0, \ldots, 0), \quad 2 \geq \mathrm{E}\left[||\mathcal{E}^{(i)}||_1\right] \geq \Delta$

(2) The slope ratio of each random vector is high:

$$\mathrm{slope}\,\mathcal{E}^{(i)} \geq (t+1)6\sqrt{t(5\ln 3t + \ln|\Delta\delta|)} \tag{8}$$

Fix $0 \leq l \leq t$. Let $\mathcal{W}_l$ denote a matrix comprised of columns $[Z, \mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \ldots, \mathcal{E}^{(l)}]$ then with probability $1-\delta$ we have

$$\Gamma(\mathcal{W}_l) = \min_{\mathbf{x}\neq 0} \frac{||\mathcal{W}_l\mathbf{x}||_1}{||\mathbf{x}||_1} \geq \frac{\Delta}{6(l+1)}.$$

Proof. The proof goes by induction on l. The result for the base $l = 0$, is immediate because of the normalization of Z. Suppose we have proved for $l-1$ that $\Gamma(\mathcal{W}_{l-1}) \geq \frac{\Delta}{6l}$ with probability at least $1 - \frac{(l-1)\delta}{t}$ and would like to extend to l. The proof of the induction hypothesis consists of three parts, first we show that for an arbitrary fixed unit vector $\mathbf{p} = (p_0, \ldots, p_l)$, the probability

$$||\mathcal{W}_l\mathbf{p}||_1 \geq \left(1 - \frac{2}{3(l+1)}\right)\frac{\Delta}{6l} \tag{9}$$

is exponentially small, then we use the union bound to extend the result to sufficiently dense discrete subset of all possible vectors $\mathbf{p}$, and finally we will use the continuity of a linear operator to extend it to all unit $\mathbf{p} \in \Re^{l+1}$ to complete the proof.

Consider an arbitrary normalized vector $\mathbf{p} \in \Re^{l+1}$, and let $X = p_0 Z + p_1\mathcal{E}^{(1)} + , \ldots, + p_{l-1}\mathcal{E}^{(l-1)}$, by our induction hypothesis we have $\Gamma(\mathcal{W}_{l-1}) \geq \frac{\Delta}{6l}$ with probability $1 - \frac{l-1}{t}\delta$, thus we have:

$$||X||_1 \geq (1-p_l)\frac{\Delta}{6l}. \tag{10}$$

We also have $\mathrm{E}\left[||p_l\mathcal{E}^{(l)}||_1\right] \geq p_l\Delta$. But $\mathcal{E}^{(l)}$ is independent of X, thus, applying Lemma 3.1 part (i) we have

$$\Pr\left[||X + p_l\mathcal{E}^{(l)}||_1 \geq (1-\varepsilon)\max\left[(1-p_l)\frac{\Delta}{6l}, \frac{p_l\Delta}{2}\right]\right] \geq 1-\delta_0 \tag{11}$$

where we choose $\varepsilon = \frac{1}{6(l+1)}$ and $\delta_0 = \frac{\delta\Delta^l}{t(4l)^{3l}}$.

Now we compute a lower bound for the max in this equation. If $p_l \leq \frac{1}{2(l+1)}$ then we have:

$$\max\left[(1-p_l)\frac{\Delta}{6l}, \frac{p_l\Delta}{2}\right] \geq (1 - \frac{1}{2(l+1)})\frac{\Delta}{6l} \tag{12}$$

and if $p_l \geq \frac{1}{2(l+1)}$ then

$$\max\left[(1-p_l)\frac{\Delta}{6l}, \frac{p_l\Delta}{2}\right] \geq \frac{\Delta}{4(l+1)} \geq (1 - \frac{1}{2(l+1)})\frac{\Delta}{6l} \tag{13}$$

Combining (13) and (12) and substituting $\varepsilon = \frac{1}{6(l+1)}$ we have:

$$\begin{aligned}(1-\varepsilon)\max((1-p_l)\tfrac{\Delta}{6l}, \tfrac{p_l\Delta}{2}) &\geq (1-\tfrac{1}{6(l+1)})(1-\tfrac{1}{2(l+1)})\tfrac{\Delta}{6l} \\ &\geq (1-\tfrac{2}{3(l+1)})\tfrac{\Delta}{6l}\end{aligned}$$

therefore from (11) we have:

$$\Pr\left[||X + p_l\mathcal{E}^{(l)}||_1 \geq (1-\frac{2}{3(l+1)})\frac{\Delta}{6l}\right] \geq 1-\delta_0$$

Now consider a set $\mathcal{P} \subset \Re^l$, such that for any $p \in \Re^l, ||p||_1 = 1$, there exists $p^* \in \mathcal{P}$, such that $||p-p^*||_1 \leq \frac{\Delta}{54l^2}$. One possible choice is to take $\mathcal{P}$ such that it contains all possible vectors of the form $(\frac{i_1\Delta}{54(l+1)^3}, \ldots, \frac{i_l\Delta}{54(l+1)^3}$ for all integer i_j, such that $||p||_1 \leq 1$. An immediate calculation gives that this set would contain at most $(\frac{54(l+1)^3}{\Delta})^l$ elements. Using the fact that $\delta_0 = \frac{\delta\Delta^l}{t(4l)^{3l}}$, and a union bound, with probability at least $1-\frac{\delta}{t}$, we have that for all $p \in \mathcal{P}$,

$$||\mathcal{W}_l||_1 \geq (1-\frac{2}{3(l+1)})\frac{\Delta}{6l}.$$

Finally, for any p, there exists $p' \in \mathcal{P}$, such that $||p-p'||_1 < \frac{\Delta}{54l^2}$ and thus we have:

$$\begin{aligned}||\mathcal{W}_l p||_1 &= ||\mathcal{W}_l(p' + (p-p'))||_1 && (14)\\ &\geq ||\mathcal{W}_l p'||_1 - ||\mathcal{W}_l(p-p')||_1 && (15)\\ &\geq ||\mathcal{W}_l p'||_1 - 3||(p-p')||_1 && (16)\\ &\geq (1-\frac{2}{3(l+1)})\frac{\Delta}{6l} - \frac{\Delta}{18(l+1)^2} && (17)\\ &\geq (1-\frac{1}{(l+1)})\frac{\Delta}{6l} = \frac{\Delta}{6(l+1)} && (18)\end{aligned}$$

where the transition in (16) follows from the fact that all columns in $\mathcal{W}_l$ have norm at most 3 with high probability. Thus given the induction hypothesis the $\Gamma(\mathcal{W}_l) \geq \frac{\Delta}{6(l+1)}$ with probability at least $1-\frac{\delta}{t}$, combining it with the fact that the induction hypothesis holds with probability at least $1-\frac{(l-1)\delta}{t}$, we immediately have that $\Gamma(\mathcal{W}_l) \geq \frac{\Delta}{6(l+1)}$ with probability at least $1-\frac{l\delta}{t}$. This finishes the proof of the induction hypothesis, and in turn completes the proof of the lemma. ■

Now we fulfill the main promise of this section and prove that topics along any path are independent in linear algebraic sense:

THEOREM 3.4. *Suppose $W = [T^{(0)}, \ldots, T^{(l-1)}]$ is a path between the root and some node in the topic hierarchy, and such that topic $T^{(i)}$ was obtained by applying a mutation vector $\mathcal{E}^{(i)}$, satisfying the conditions (i)-(ii) of Lemma 3.3 . Then with probability $1-\delta$, $\Gamma(W) \geq \frac{\Delta}{18l}$*

Proof. Recall $T^{(i)}$ is obtained from $T^{(i-1)}$ by adding a random change vector and re-normalizing:

$$T^{(i)} = \alpha_i(T^{(i-1)} + \mathcal{E}^{(i)}), \tag{19}$$

where α_i is a normalization coefficient introduced to maintain $||T^{(i)}||_1 = 1$. Using lemma 3.1, we immediately have that $\alpha_i \leq 2$ with high probability. Now, consider any linear combination of topics:

$$X = p_0T^{(0)} + \ldots + p_{l-1}T^{(l-1)} \tag{20}$$

and we need to show that if $||\mathbf{p}||_1 = 1$, then $||X||_1$ is large. To prove this we substitute (19) into (20) and regroup:

$$X = \gamma_0T^{(0)} + \gamma_1\mathcal{E}^{(1)} + \ldots + \gamma_{l-1}\mathcal{E}^{(l-1)}$$

where

$$\gamma_i = p_i\alpha_i + p_{i+1}\alpha_i\alpha_{i+1} + \ldots + p_l\alpha_i\cdot\ldots\cdot\alpha_l = \alpha\sum_{j=i}^{l} p_j\prod_{q=i}^{j}\alpha_q \tag{21}$$

where we define $\alpha_0 = 1$. By the lemma 3.3 we have $||X||_1 \geq \frac{\Delta}{6l}||\gamma||_1$. Thus to prove the lemma it suffices to show that $\sum|\gamma_i| \geq 1/3$ and the result would immediately follow. To prove that, we rewrite equation (21) as

$$\gamma_i = p_i\alpha_i + \gamma_{i+1}\alpha_i,$$

but $|\gamma_i| + |\gamma_{i+1}\alpha_i| \geq |\gamma_i - \gamma_{i+1}\alpha_i|$ and thus

$$|\gamma_i| + |\gamma_{i+1}\alpha_i| \geq |\gamma_i - \gamma_{i+1}\alpha_i| = |p_i\alpha_i|,$$

dividing both sides by α_i, summing over i and use Lemma 3.2, to show that $\alpha_i \geq 0.5$ we have:

$$3\sum_{i=0}^{l}|\gamma_i| \geq \sum_{i=0}^{l}\frac{|\gamma_i|}{\alpha_i} + |\gamma_{i+1}| \geq \sum|p_i| = 1.$$

Therefore we have proved that $||\gamma||_1 \geq 1/3$ and the lemma follows. ■

3.4 Hierarchical Models: Analysis.

We start with a simple definition, which quantifies a relationship between document generated by any topic in the hierarchy, and an arbitrary root based path in the hierarchy.

DEFINITION 1 (PROJECTION OF A MIXTURE ON TO A PATH). *Given some path in the hierarchy $T^{(0)}$, $T^{(1)}$, ..., $T^{(l)}$, and a linear combination of topics (not necessarily on the path) $D = p_1T^{(b_1)} + \ldots + p_{l'}T^{(b_{l'})}$. The* projection *of D on to a path is a linear combination of topics D', such that*

$$D' = p_1T^{(\pi(b_1))} + \cdots + p_{l'}T^{(\pi(l_1))}$$

where $\pi(b_i)$ denotes the closest to $T^{(b_i)}$ topic in the path.

Now we prove the key lemma of this section.

LEMMA 3.5. *Let $T^{(0)}$, $T^{(1)}$, ..., $T^{(l)}$ be a path in the hierarchy. Let R be some vector which was only chosen based on $T^{(1)}, \ldots, T^{(l)}$, but not any other topics, and such that its maximal element is bounded by some constant β. Then for any topic $T^{(c)}$, let $T^{(a_j)}$ be the closest topic to $T^{(c)}$ and suppose the path between them has length c nodes, then we have*

$$|\langle RT^{(c)}\rangle - \langle RT^{(j)}\rangle| \leq c\varepsilon\beta \tag{22}$$

with high probability.[5] *Furthermore, if M is a mixture of topics, then*

$$|\langle RM\rangle - \langle RM'\rangle| \leq m\varepsilon\beta$$

where M' is a projection of M on to the path $T^{(0)}, \ldots, T^{(l)}$ and m is the maximal length of the path between any topic in the mixture and a path.

[5]It is also possible to improve the factor of c to $\sqrt{c}$ in (22), using martingales and increasing the length of the proof significantly.

Proof. Let $T^{(j)}$ be the closest node in the path to $T^{(c)}$. The proof is based on the observation that $T^{(j)}$ is an ancestor of $T^{(c)}$. Let $T^{(d_0)}, \ldots, T^{(d_c)}$, be a path from $T^{(j)}$ to $T^{(c)}$. To prove our lemma it is sufficient to prove that if T' is a child of T and R is independent of T' then

$$|\langle RT' \rangle - \langle RT \rangle| \leq \varepsilon\beta$$

and apply triangle inequality.

Recall that $T' = \alpha'(T + \mathcal{E}')$, where α' and $\mathcal{E}'$ are respectively the normalization constant and the mutation vector from the generative process. Therefore we have:

$$\begin{aligned} |(\langle RT' \rangle - \langle RT \rangle| &= |(1-\alpha')\langle RT' \rangle + \alpha' \langle R\mathcal{E}' \rangle| \\ &\leq (1-\alpha')\beta + \alpha' \langle R\mathcal{E}' \rangle, \end{aligned} \quad (23)$$

Now we just need to compute an upper bound for the right hand side. For the first term, using Lemma 3.2, we have $|1-\alpha'| \leq \varepsilon/2$ with high probability. To bound the second term note that $\mathcal{E}'_i$ are chosen independently of R, so we can rewrite $\langle R\mathcal{E}' \rangle = \sum_{i=1}^{n} R_i \mathcal{E}'_i$, where all terms in the sum are independent and have zero expectation, so we can apply Hoeffding's inequality:

$$\Pr\left[\langle R\mathcal{E}' \rangle \geq \frac{\beta\varepsilon}{2}\right] \leq 2\exp\left(\frac{\varepsilon^2}{2||M'||_2}\right) \leq 2\exp\left(\frac{\varepsilon^2}{2\varepsilon_2^2}\right) \leq \frac{\delta}{2}$$

where M' denotes the vector of absolute maximal values that $\mathcal{E}'$ can take and we used ε_2 from the constraint (iv) on the slope ratio of the hierarchical construction.

The generalization to a mixture of topics follows automatically by applying the previous result to each of the term in the linear combination and using the fact that the mixture is normalized. ■

Now we conclude this section by building a connection between documents and topics in the hierarchy. Recall that a document is a sample from a mixture of topic distributions. The next corollary shows that we can use the document to infer the relevance to each of the topics to an arbitrary path in the hierarchy.

COROLLARY 3.6 (DOCUMENT RELEVANCE TO A PATH). *Consider a path starting from the root $T^{(a_0)}$, ..., $T^{(a_l)}$, and the corresponding weight matrix $W = [T^{(0)}, \ldots, T^{(l)}]$ and let V be a pseudoinverse of W. Let a document is sampled from a mixture*

$$D = p_1 T^{(b_1)} + \ldots + p_{l'} T^{(b_{l'})}$$

(not necessarily overlapping with the topics in the path) and let $\tilde{D}$ be the term count vector of this document, then the vector $\tilde{\mathbf{r}} = V\tilde{D}$ would approximate the coefficients of the projection of D onto the path.

Proof. The proof follows immediately from the previous lemma and the lemma 2.2 ■

3.5 Algorithms. In this section we present our algorithms and make some additional remarks on the analysis. There are three algorithms to be discussed here. First, given a path in the hierarchy (such as `machine learning` → `computer science` → `science` → `base`), we would like to compute where in the hierarchy a given document belongs. If, for example, a document is related to `machine learning` and `biology`, when for the path above, we will learn that it is related to both `machine learning` and `science` (as `science` is the closest ancestor of `biology` that is in the path). Or, alternatively, if a document is about `soccer`, it will be assigned to the baseline topic. From Corollary 3.6 it follows that the relevances computed by using pseudoinverse are an accurate approximation of the projection of the real mixture onto the path.

Our second algorithm is concerned with reconstructing the hierarchy given the term distributions for topics.[6] Finally, we present an algorithm that finds topics and builds a topical hierarchy from unlabeled data. Here we use a modification of an algorithm from [22] that extracts topics from the co-occurrence matrix.

Classification along the path. The algorithm is based on the fact that that topics along the path are sufficiently independent, which implies that we can build a pseudoinverse matrix for those topics with bounded maximal element. Corollary 3.6 can then be used to prove that the produced relevances are a projection of the real relevances on that path.

ALGORITHM 3.1 (COMPUTING DOCUMENT RELEVANCE).

Input: A path in the hierarchy $T^{(0)}, T^{(i_1)}, \ldots, T^{(i_k)}$, a document's indicator vector $\tilde{D}$
Output: Relevance to each of the topics along the path

1. *Let $W = [T^{(0)} T^{(i_1)}, \ldots, T^{(i_k)}]$*
2. *Compute the pseudoinverse matrix V such that $VW = I$, and $max_{ij}|V_{ij}|$ is minimized.*
3. *Return $\tilde{\mathbf{r}} = V\tilde{D}$*

3a. *To cluster: return the topic i which maximizes $\tilde{r}_i$.*

Reconstructing the topic hierarchy given the leaf topical distributions. The main idea of this algorithm is the following observation: suppose we have a base topic $T^{(0)}$ and some leaf topic $T^{(c)}$. Let $W = [T^{(0)}; T^{(c)}]$ and let $V = W^{-1}$. From Lemma 3.5 it follows if a topic $T^{(d)}$ is in different subtree than $T^{(c)}$, then we would have

$$VT^{(d)} \approx VT^{(0)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (24)$$

since $T^{(0)}$ is the closest node in the path to $T^{(d)}$. However if $T^{(d)}$ is in the same subtree then we don't have any guarantees for the value $VT^{(d)}$. Indeed, our lemma says that the value would be close to the value of the closest topic which lies on the path between $T^{(c)}$ and $T^{(0)}$. However we don't know that topic. We conjecture,[7] that $VT^{(d)} \approx \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ where $\alpha + \beta = 1$ and their ratio is defined by the ratio of distances (under some proximity measure) to the base topic and the topic $T^{(c)}$. This gives us a foundation for the algorithm: we can approximate baseline topic by computing cumulative

[6] An application for this algorithm would be if we have labeled data and would want to build automatic topic hierarchy.

[7] We can prove this for a special case of gaussian change function.

term distributions across all the documents available. Then for each topic T' we compute all the topics which lie outside of the subtree where T' belongs to, and then we combine all the leafs in the subtree to build a new root for the subtree and iterate. A high level description of the algorithm is below.

ALGORITHM 3.2 (TOPIC HIERARCHY).
Input: A collection of topical distributions, a threshold value $\tau \geq 0$
Output: A topic hierarchy

1. *Estimate baseline topic* $T^{(0)}$ *by computing average distributions across all topics.*
2. *For each leaf topic* $T^{(c)}$ *consider weight matrix* $W = (T^{(c)}, T^{(0)})$ *and compute generalized pseudoinverse* V *using linear programming described in Section 2.2*

2a. *For all other leaves topics* $T^{(d)}$*, compute* $\mathbf{r} = VT^{(d)}$ *and assign* $T^{(d)}$ *and* $T^{(c)}$ *to the same subtree if* $r_1 \geq \tau$.

3. *As a result of step [2], we get a disjoint family of topic sets, each of them will form independent subtree.*
4. *Apply steps 1-3 recursively on each subset.*

Reconstructing the topics from unlabeled data. In this section we give an empirical algorithm on how to construct the topics out of a large collection of unlabeled data. The algorithm could be used for both leaf topic reconstruction and hierarchy reconstruction, and produces classification as a byproduct. Consider a co-occurrence matrix $\mathcal{P}_{ij}$, which for every pair of terms i and j measures how often they occur in the same document across the entire collection. Then we normalize all the columns so that they represent valid probability distributions. Then we approximate the root distribution by the aggregate distribution of terms across all documents. After that, we choose a column such that the L_1 distance between the column and the baseline distribution $T^{(0)}$ is maximal. We treat it as a topic and run Algorithm 3.1 to do binary classification between $T^{(0)}$ and the found column. The promise is that all the documents which are outside of the same tree will be assigned to the $T^{(0)}$ topic, and the documents which are in the same subtree topic will get assigned to the found topic. After that we iterate the entire algorithm on remaining data and/or we can further iterate the algorithm on the constructed cluster to build subtree.[8]

ALGORITHM 3.3 (RECONSTRUCTING THE TOPICS).
Input: Unlabeled Data
Output: Topical hierarchy and classification.

1. *Build the co-occurrence matrix* $\mathcal{P}$ *from the data and normalize (in* L_1 *norm) its columns.*
2. *Estimate the baseline distribution* $T^{(0)}$ *by computing the term histogram across all documents.*
3. *Choose a column* $\mathcal{P}_i$ *of the co-occurrence matrix which maximizes the* L_1 *distance to baseline topic.*

[8]Due to space constraints, we omit a few important details on how to deal with the fact some of the columns might not be well approximated. We refer to the actual code which is available upon request, for more details.

1	2	3	4	5	6	7	8	9	10
1.20	1.13	1.17	1.12	1.25	1.07	1.05	.91	.86	0
-.32	-.11	-.12	-.05	-.37	-.02	-.26	.16	.18	1
11	12	13	14	15	16	17	18	19	20
.46	1.15	1.03	1.03	1.09	1.06	1.05	.99	1.05	1.0
.48	-.17	-.05	-.08	-.02	-.02	-.01	.04	-.07	.02

Table 1: First row contains relevance of each of the 20 newsgroup to `rec.sport.baseball` (topic 10). The second row contains the relevance to the baseline topic.

4. *Use* $\mathcal{P}_i$ *and* $T^{(0)}$ *as two topics and perform the clustering by using Algorithm 3.1:*
   - 4a *Let* $\mathcal{W} = \left[\mathcal{P}_i; T^{(0)}\right]$ *compute the pseudoinverse* $V = \mathcal{W}^{-1}$
   - 4b *For each document* d *compute* $V\tilde{\mathbf{d}}$ *and assign documents which have high relevance to* $\mathcal{P}_i$ *to the new cluster.*
   - 4c *Use all the documents assigned to a cluster to refine the term distribution for the cluster*
5. *Remove clustered documents from the collection and iterate the algorithm until* $(1-\varepsilon)$ *of all documents are clustered*
7. *Apply the algorithm on the each cluster.*

4 Experiments

We perform two kinds of experiments. First we validate our model by performing some experiments on labeled data. In the second part all our experiments are performed in fully unsupervised manner, we reconstruct both the topics and the hierarchy, and perform clustering and compare it with the ground truth. .

4.1 Model validation and reconstructing hierarchy from labeled data. In this section we perform a few experiments on labeled data to reconstruct topic hierarchy given topical distributions themselves. In particular for each topic distribution we create the set of other topics which our algorithm deem related to it.

We use 20 newsgroup[19] dataset which contain 20 following newsgroups:

```
1. graphics, 2. os.ms-windows-misc,
3. ibm.hardware, 4. mac.hardware,
5. windows.x, 6. sci.electronics
7. misc.forsale, 8. rec.autos, 9. rec.motorcycles
10. rec.sport.baseball, 11. rec.sport.hockey
12. sci.crypt, 13. sci.med 14. sci.space
15. soc.religion.christian, 16. alt.atheism
17. religion.misc, 18. politics.guns
19. politics.mideast, 20. politics.misc
```

We use the numbering above consistently throughout the rest of the paper.

Let c be one of the topics (say `rec.sport.hockey`), let $W = [T^{(0)}; T^{(c)}]$ and $V = W^{-1}$ is generalized pseudoinverse matrix with minimal maximal element. Now, for each of the topics (including `rec.sport.baseball`) we compute the

2-dimensional vector $VT^{(c)}$ and the results are presented in Table 1. The most striking result in the table above is that for a given topic $T^{(d)}$ which is semantically unrelated to $T^{(c)}$, the product $VT^{(d)}$ is either very close to $V \times T^{(0)} = \binom{0}{1}$ (topics 6, 7 and 13-20), or has the form $VT^{(d)} = \binom{-\alpha}{1+\alpha}$ for some positive α (topics 1-4). For semantically related topics, such as topic 11 (`rec.sport.baseball`) we have relevance $\binom{0.46}{0.48}$ and somewhat related 8 (`rec.autos`) and 9 (`rec.motorcycles`), we have $\approx \binom{0.1}{0.9}$. It is interesting that for most of the topics, the prediction from the theorem 3.5 carry out almost precisely.

The negative relevance phenomenon can be easily explained in the framework of the model. Recall that we approximate the baseline topic distribution $T^{(0)}$ by averaging the term distribution across all the documents in the data. However, if we have a bias for some topic (or many related topics) in the collection, then our baseline distribution might become biased towards that topic. E.g. if a leaf topic $T^{(c)} = T^{(0)} + \mathcal{E}^{(c)}$ and $T^{(d)} = T^{(0)} + \mathcal{E}^{(d)}$ but instead of finding the true $T^{(0)}$ we found

$$\tilde{T}^{(0)} = T^{(0)} + \frac{1}{3}\mathcal{E}^{(c)}, \tag{25}$$

then topics c and d are no longer probabilistically independent from each other given $\tilde{T}^{(0)}$. Instead we have:

$$T^{(d)} = \frac{3}{2}\tilde{T}^{(0)} - \frac{1}{2}T^{(c)} + \mathcal{E}^{(d)} \tag{26}$$

Let V be a pseudoinverse matrix of $W = [\tilde{T}^{(0)}, T^{(c)}]$, multiplying both parts of (26) by V we immediately have:

$$VT^{(d)} = \binom{3/2}{-1/2} + V\frac{2}{3}\mathcal{E}^{(d)} \approx \begin{pmatrix} 3/2 \\ -1/2 \end{pmatrix}$$

Tuning the constant in the bias formula (25) we get that $VT^{(d)} = \binom{1+\alpha}{-\alpha}$, which is exactly what we observe in our experiments. Furthermore, topics 1-4 in the table 1 that exhibit this behavior are indeed a part of the largest topic in the dataset (computer related documents).

Now we apply Algorithm 3.2 with the threshold parameter $\tau = 0.3$. It produces 8 high level topics which we roughly name as: all computer related, `sci.crypt`, `sci.space`, `sci.med, politics, sports, motor vehicles` and `religion`. Iteration on each component produces a collection of individual topics with the exception of ibm.hardware and mac.hardware and the latter remain grouped and needed one more iteration to split. Note that produced hierarchy is a perfectly reasonable hierarchy on the newsgroups.

4.2 Clustering arXiv. In this section we perform unsupervised clustering of arXiv dataset[2]. The snapshot has approximately 250,000 abstracts on various areas of physics, and also computer science and math. The first level of the hierarchy produced by our algorithm is presented in Table 3. Also the recall/precision table corresponding to the full hierarchy is presented in Table 2. Note that the arXiv contains many heavily overlapping topics, and it is not clear, that there would be a consensus if we did classification manually. For example, many abstracts assigned to quantum-ph and cs (cluster 5), are about quantum computing, and it is probably not possible to differentiate them in a meaningful way. Nevertheless, many topics do get separated cleanly with high precision and recall (in particular `astro-ph`, `hep-ph`, `condensed matter`, `computer science` and `math`). Another interesting and exciting property of the algorithm is that it successfully separated topics of very different sizes. The most striking example is `computer science` (3K documents, 64% recall/78% precision in a single cluster), and one of the largest topics `astro-ph` (45K documents, 58% recall, 93% precision in the largest cluster, or 86%/94% if we combine the 5 clusters where astrophysics is a majority). We also note that for two topics `math-ph` and `physics` we did not succeed finding clusters where they would form a majority – which however, comes hardly as a surprise, as they don't have well defined boundaries (especially `physics`!) and span across many areas of physics and mathematics.

	all clusters		majority	
topic(size)	R/P	NC	R/P	NC
math-ph (2880)	0 / 0	0	0 / 0	0
hep-ph (37926)	79 / 69	18	73 / 77	14
nucl-ex (1373)	46 / 44	2	21 / 58	1
nucl-th (8467)	46 / 61	3	46 / 61	3
gr-qc (10993)	56 / 41	7	14 / 64	2
cond-mat (47536)	87 / 69	10	80 / 78	6
astro-ph (44702)	90 / 83	9	86 / 94	5
hep-lat (5813)	16 / 51	2	8 / 53	1
quant-ph (7273)	43 / 84	1	43 / 84	1
cs (3549)	64 / 78	1	64 / 78	1
hep-ex (4646)	51 / 80	4	47 / 89	3
hep-th (31603)	66 / 64	14	54 / 74	10
nlin (9750)	17 / 45	2	12 / 54	1
physics (6866)	7 / 42	1	0 / 0	0
math (23368)	81 / 82	2	81 / 82	2
total (246745)	70 / 70	76	63 / 80	50

Table 2: Recall/precision tables for arXiv hierarchy. When counting P/R, each cluster (e.g. all documents assigned to a single leaf node in the hierarchy) is assigned to the most expressed topic in that cluster. In the column "all clusters" the P/R is counted over the union of all clusters assigned to a particular topic. NC stands for the total number of clusters assigned to a topic. The second column contains recall/precision table where we disqualify all the clusters where the largest topic is not a majority.

4.3 Clustering of Newsgroups 20. We present the first level topics of the newsgroup hierarchy that our algorithm has reconstructed in table 4.3. Note that for religion/political newsgroups our algorithm produced high level clusters which go across groups boundary, but yet make perfect sense: Cluster 8 contains mostly documents related to wars (politics.mideast and politics.guns). Cluster 4 contains mostly documents related to religion, and cluster 9 contains medical and health related documents contains sci.med and partly some politics.

5 Conclusions and Open Problems

We have proposed a new generative model to describe hierarchical topic structure in discrete mixture models. Our analysis provides a mathematical framework and enables efficient algorithms for text classifications and topic hierarchy reconstruction. One of the key features of our approach, is that it provides guarantees without the assumption that the entire model is reconstructed. Now we outline a few

(size)	major topics (recall)
1 (16K)	85% math (59%)
2 (10K)	99% astro-ph (24%)
3 (12K)	75% hep-ph (24%), 18% hep-ex (48%),
4 (17K)	91% cond-mat (34%),
5 (23K)	20% quant-ph (40%), 15% astro-ph (8%),
6 (19K)	42% hep-ph (21%), 28% hep-th (17%),
7 (39K)	63% cond-mat (50%), 12% nlin (51%)
8 (27K)	93% astro-ph (58%),
9 (25K)	49% hep-th (39%), 24% hep-ph (16%)
10 (19K)	51% hep-ph (26%), 28% nucl-th (66%)
11 (35K)	30% hep-th (33%), 21% math (31.2%)
n/a(5K)	21% astro-ph (2%), 18% hep-ph (2.5%)

Table 3: The clustering or arXiv that only uses the top level topics. The first number before each topic is that topic's precision in that cluster, the number inside the parentheses is the topic's recall in the cluster. For each cluster we show the most represented topic(s) in that cluster. Note that some topics are so small that they don't appear in this top level hierarchy (like cs, which is almost entirely in cluster 5, but only contributed 12% of its size).

(size)	major topics (recall)
1 (6716)	windows.x (96%), ibm.pc (93%), graphics (88%), ms-windows (89%), mac (88%), forsale (79%), electronics (61%)
2 (2156)	hockey (93%), baseball (87%), med (6%)
3 (1433)	mideast (66%), guns (22%), politics.misc (24%) religion.misc (9%)
4 (2649)	christian (78%), 29% atheism (78%), religion.misc (57%)
5 (883)	sci.crypt (73%), politics.misc (4%)
6 (2204)	motorcycles (86%), rec.autos (70 %), sci.med (13%)
7 (1260)	sci.space (70%), sci.electronics (20%), sci.med (17 %)
8 (1463)	guns (56%), politics (31%), religion (20%)
9 (1214)	sci.med (31%), politics (11%), guns (10%), autos (7%)

Table 4: The top part of the hierarchy for Newsgroups. The numbers in parentheses in the second column is the recall of that topic in that cluster. We shortened the name of each newsgroup to be able to fit them into a table. The topic `computer` corresponds to the 5 computer related groups.

open questions and further directions for this framework. We know how to classify documents along the path in the tree. However the algorithm which reconstructs the path (and the tree) is based on a conjecture that says that topics on the path between root and leaf note behave in a continuous manner, and this allows to differentiate between topics which belong to the same subtree. It would be interesting to prove this conjecture. Another related direction would be to develop a connection between proposed model and phylogenetic trees reconstruction problem. We believe that out approach can provide a valuable tool for the analysis of the origin of species. Another interesting direction is to use our algorithm to build hierarchy on terms. In particular we could apply the algorithm which reconstructs topic hierarchy to the co-occurrence matrix, and that would create an hierarchy on terms. Exploring this, would be an interesting and exciting direction.

6 Acknowledgment

Author would like to thank Corinna Cortes, Jon Feldman and S. Muthukrisnhan for useful discussions.

7 References

[1] A. Ambainis, R. Desper, M. Farach-Colton, and S. Kannan. Nearly tight bounds on the learnability of evolution. In *FOCS*, 1997.

[2] Arxiv. http://www.arxiv.org.

[3] D. Blei, T. Gri, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, 2004.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *J. of Machine Learning Research*, 3, 2003.

[5] A. Dasgupta, J. Hopcroft, J. Kleinberg, and M. Sandler. On learning mixtures of heavy-tailed distributions. In *Proc. of FOCS*, 2005.

[6] S. Dasgupta and L. Schulman. A two-round variant of em for gaussian mixtures. In *Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2000.

[7] M. Farach and S. Kannan. Efficient algorithms for inverting evolution. In *STOC*, 1999.

[8] K. Heller and Z. Ghahramani. Bayesian hierarchical clustering. In *ICML*, 2005.

[9] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:301:13–30, March 1963.

[10] T. Hofmann. The cluster-abstraction model: Unsupervised learning of topic hierarchies from text data. In *IJCAI*, 1999.

[11] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[12] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *Proc. of IJCAI*, 1999.

[13] A. Jepson and M. Black. Mixture models for optical flow computation. *Partitioning Data Sets I*, 1993.

[14] L. B.-T. Jones, R. Bean, G. McLachlan, and J. Zhu. *Intelligent Data Engineering and Automated Learning - I*, volume 3578, chapter Application of Mixture Models to Detect Differentially Expressed Genes, pages 422–431. Springer, 2005.

[15] J. Kleinberg and M. Sandler. Using mixture models for collaborative filtering. In *STOC*, 2004.

[16] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[17] A. K. McCallum. Multi-label text classification with a mixture model trained by em. In *Proc. of AAAI'99 workshop on text learning*, 1999.

[18] G. McLachlan and K. Basford. *Mixture Models, inference and applications to clustering*. Marcel Dekker, 1987.

[19] T. Mitchell. 20 newsgroups.

[20] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[21] L. A. Salter. Algorithms for phylogenetic tree reconstruction, 2000.

[22] M. Sandler. On the use of linear programming for unsupervised text classification. In *Proc. of KDD*, 2005.

[23] M. Sharikar and N. Ailon. fitting tree metrics: hierarchical clustering and phylogeny. In *FOCS*, 2005.

[24] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, volume II, pages 246–252, 1999.

[25] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *J. of the American Statistical Association*, 2006.

[26] K. Toutanova, F. Chen, K. Popat, and T. Hofmann. Text classification in a hierarchical mixture model for small training sets. In *CIKM '01*. ACM Press, 2001.

[27] Y. Weiss and E. H. Adelson. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *CVPR*, 1996.