The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Algorithms for Generalized Topic Modeling

Avrim Blum
Toyota Technological Institute at Chicago
avrim@ttic.edu

Nika Haghtalab
Computer Science Department
Carnegie Mellon University
nhaghtal@cs.cmu.edu

Abstract

Recently there has been significant activity in developing algorithms with provable guarantees for topic modeling. In this work we consider a broad generalization of the traditional topic modeling framework, *where we no longer assume that words are drawn i.i.d. and instead view a topic as a complex distribution over sequences of paragraphs.* Since one could not hope to even represent such a distribution in general (even if paragraphs are given using some natural feature representation), we aim instead to directly learn a predictor that given a new document, accurately predicts its topic mixture, without learning the distributions explicitly. We present several natural conditions under which one can do this from unlabeled data only, and give efficient algorithms to do so, also discussing issues such as noise tolerance and sample complexity. More generally, our model can be viewed as a generalization of the multi-view or co-training setting in machine learning.

1 Introduction

Topic modeling is an area with significant recent work in the intersection of algorithms and machine learning (Arora et al. 2012; Arora, Ge, and Moitra 2012; Arora et al. 2013; Anandkumar et al. 2012; 2014; Bansal, Bhattacharyya, and Kannan 2014). In topic modeling, a topic (such as sports, business, or politics) is modeled as a probability distribution over words, expressed as a vector $\mathbf{a}_i$. A document is generated by first selecting a mixture $\mathbf{w}$ over topics, such as 80% sports and 20% business, and then choosing words i.i.d. from the associated mixture distribution, which in this case would be $0.8\mathbf{a}_{sports} + 0.2\mathbf{a}_{business}$. Given a large collection of such documents (and some assumptions about the distributions $\mathbf{a}_i$ as well as the distribution over mixture vectors $\mathbf{w}$) the goal is to recover the topic vectors $\mathbf{a}_i$ and then to use the $\mathbf{a}_i$ to correctly classify new documents according to their topic mixtures. Algorithms for this problem have been developed with strong provable guarantees even when documents consist of only two or three words each (Arora, Ge, and Moitra 2012; Anandkumar et al. 2012; Papadimitriou et al. 1998). In addition, algorithms based on this problem formulation perform well empirically on standard datasets (Blei, Ng, and Jordan 2003; Hofmann 1999).

As a theoretical model for document generation, however, an obvious problem with the standard topic modeling framework is that documents are not actually created by independently drawing words from some distribution. Moreover, important words within a topic often have meaningful correlations, like shooting a free throw or kicking a field goal. Better would be a model in which *sentences* are drawn i.i.d. from a distribution over sentences. Even better would be *paragraphs* drawn i.i.d. from a distribution over paragraphs (this would account for the word correlations that exist within a coherent paragraph). Or, even better, how about a model in which paragraphs are drawn non-independently, so that the second paragraph in a document can depend on what the first paragraph was saying, though presumably with some amount of additional entropy as well? This is the type of model we study here.

Note that an immediate problem with considering such a model is that now the task of learning an explicit distribution (over sentences or paragraphs) is hopeless. While a distribution over words can be reasonably viewed as a probability vector, one could not hope to learn or even represent an explicit distribution over sentences or paragraphs. Indeed, except in cases of plagiarism, one would not expect to see the same paragraph twice in the entire corpus. Moreover, this is likely to be true even if we assume paragraphs have some natural feature-vector representation. Instead, we bypass this issue by aiming to directly learn a predictor for documents—that is, a function that given a document, predicts its mixture over topics—without explicitly learning topic distributions. Another way to think of this is that our goal is not to learn a model that could be used to *write* a new document, but instead just a model that could be used to *classify* a document written by others. This is much as in standard supervised learning where algorithms such as SVMs learn a decision boundary (such as a linear separator) for making predictions on the labels of examples without explicitly learning the distributions D_+ and D_- over positive and negative examples respectively. However, our setting is *un*supervised (we are not given labeled data containing the correct classifications of the documents in the training set) and furthermore, rather than each data item belonging to one of the k classes (topics), each data item belongs to a *mixture* of the k topics. Our goal is given a new data item to output what that mixture is.

We begin by describing our high level theoretical formulation. This formulation can be viewed as a generalization both of standard topic modeling and of *multi-view learning* or *co-training* (Blum and Mitchell 1998; Dasgupta, Littman, and McAllester 2002; Chapelle, Schlkopf, and Zien 2010; Balcan, Blum, and Yang 2004; Sun 2013). We then describe several natural assumptions under which we can indeed efficiently solve the problem, learning accurate topic mixture predictors.

2 Preliminaries

We assume that paragraphs are described by n real-valued features and so can be viewed as points $\mathbf{x}$ in an instance space $\mathcal{X} \subseteq \mathbb{R}^n$. We assume that each document consists of at least two paragraphs and denote it by $(\mathbf{x}^1, \mathbf{x}^2)$. Furthermore, we consider k topics and partial membership functions $f_1, \ldots, f_k : \mathcal{X} \to [0, 1]$, such that $f_i(\mathbf{x})$ determines the degree to which paragraph $\mathbf{x}$ belongs to topic i, and, $\sum_{i=1}^{k} f_i(\mathbf{x}) = 1$. For any vector of probabilities $\mathbf{w} \in \mathbb{R}^k$ — which we sometimes refer to as mixture weights — we define $\mathcal{X}^{\mathbf{w}} = \{\mathbf{x} \in \mathbb{R}^n \mid \forall i,\ f_i(\mathbf{x}) = w_i\}$ to be the set of all paragraphs with partial membership values $\mathbf{w}$. We assume that both paragraphs of a document have the same partial membership values, that is $(\mathbf{x}^1, \mathbf{x}^2) \in \bigcup_{\mathbf{w}} \mathcal{X}^{\mathbf{w}} \times \mathcal{X}^{\mathbf{w}}$, although we also allow some noise later on. To better relate to the literature on multi-view learning, we also refer to topics as "classes" and refer to paragraphs as "views" of the document.

Much like the standard topic models, we consider an unlabeled sample set that is generated by a two-step process. First, we consider a distribution $\mathcal{P}$ over vectors of mixture weights and draw $\mathbf{w}$ according to $\mathcal{P}$. Then we consider distribution $\mathcal{D}^{\mathbf{w}}$ over the set $\mathcal{X}^{\mathbf{w}} \times \mathcal{X}^{\mathbf{w}}$ and draw a document $(\mathbf{x}^1, \mathbf{x}^2)$ according to $\mathcal{D}^{\mathbf{w}}$. We consider two settings. In the first setting, which is addressed in Section 3, the learner receives the instance $(\mathbf{x}^1, \mathbf{x}^2)$. In the second setting, discussed in Section 4, the learner receives samples $(\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2)$ that have been perturbed by some noise. In both cases, the goal of the learner is to recover the partial membership functions f_i.

More specifically, in this work we consider partial membership functions of the form $f_i(\mathbf{x}) = f(\mathbf{v}_i \cdot \mathbf{x})$, where $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ are linearly independent and $f : \mathbb{R} \to [0, 1]$ is a monotonic function.[1] For the majority of this work, we consider f to be the identity function, so that $f_i(\mathbf{x}) = \mathbf{v}_i \cdot \mathbf{x}$. Define $\mathbf{a}_i \in \operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ such that $\mathbf{v}_i \cdot \mathbf{a}_i = 1$ and $\mathbf{v}_j \cdot \mathbf{a}_i = 0$ for all $j \neq i$. In other words, the matrix containing $\mathbf{a}_i$s as columns is the pseudoinverse of the matrix containing $\mathbf{v}_i$s as columns, and $\mathbf{a}_i$ can be viewed as the projection of a paragraph that is purely of topic i onto $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. Define $\Delta = \mathrm{CH}(\{\mathbf{a}_1, \ldots, \mathbf{a}_k\})$ to be the convex hull of $\mathbf{a}_1, \ldots, \mathbf{a}_k$.

Throughout this work, we use $\|\cdot\|_2$ to denote the spectral norm of a matrix or the L_2 norm of a vector. When it is clear from the context, we simply use $\|\cdot\|$ to denote these quantities. We denote by $B_r(\mathbf{x})$ the ball of radius r around $\mathbf{x}$. For a M, we use M^+ to denote the pseudoinverse of M.

Generalization of Standard Topic Modeling

Let us briefly discuss how the above model is a generalization of the standard topic modeling framework. In the standard framework, a topic is modeled as a probability distribution over n words, expressed as a vector $\mathbf{a}_i \in [0, 1]^n$, where a_{ij} is the probability of word j in topic i. A document is generated by first selecting a mixture $\mathbf{w} \in [0, 1]^k$ over k topics, and then choosing words i.i.d. from the associated mixture distribution $\sum_{i=1}^{k} w_i \mathbf{a}_i$. The document vector $\hat{\mathbf{x}}$ is then the vector of word counts, normalized by dividing by the number of words in the document so that $\|\hat{\mathbf{x}}\|_1 = 1$.

As a thought experiment, consider infinitely long documents. In the standard framework, all infinitely long documents of a mixture weight $\mathbf{w}$ have the same representation $\mathbf{x} = \sum_{i=1}^{k} w_i \mathbf{a}_i$. This representation implies $\mathbf{x} \cdot \mathbf{v}_i = w_i$ for all $i \in [k]$, where $V = (\mathbf{v}_1, \ldots, \mathbf{v}_k)$ is the pseudo-inverse of $A = (\mathbf{a}_1, \ldots, \mathbf{a}_k)$. Thus, by partitioning the document into two halves (views) $\mathbf{x}^1$ and $\mathbf{x}^2$, our *noise-free model* with $f_i(\mathbf{x}) = \mathbf{v}_i \cdot \mathbf{x}$ generalizes the standard topic model for long documents. However, our model is substantially more general: features within a view can be arbitrarily correlated, the views themselves can also be correlated, and even in the zero-noise case, documents of the same mixture can look very different so long as they have the same projection to the span of the $\mathbf{a}_1, \ldots, \mathbf{a}_k$.

For a shorter document $\hat{\mathbf{x}}$, each feature $\hat{x}_i$ is drawn according to a distribution with mean x_i, where $\mathbf{x} = \sum_{i=1}^{k} w_i \mathbf{a}_i$. Therefore, $\hat{\mathbf{x}}$ can be thought of as a noisy measurement of $\mathbf{x}$. The fewer the words in a document, the larger is the noise in $\hat{\mathbf{x}}$. Existing work in topic modeling, such as (Arora, Ge, and Moitra 2012; Anandkumar et al. 2014), provide elegant procedures for handling large noise that is caused by drawing only 2 or 3 words according to the distribution induced by $\mathbf{x}$. As we show in Section 4, our method can also tolerate large amounts of noise under some conditions. While our method cannot deal with documents that are only 2- or 3-words long, the benefit is a model that is much more general in many other respects.

Generalization of Co-training Framework

Here, we briefly discuss how our model is a generalization of the *co-training* framework. The standard co-training framework of Blum and Mitchell considers learning a binary classifier from primarily unlabaled instances, where each instance $(\mathbf{x}^1, \mathbf{x}^2)$ is a pair of *views* that have the same classification. For example, Blum and Mitchell and Balcan and Blum show that if views are independent of each other given the classification, then one can efficiently learn a halfspace from primarily unlabeled data. In the language of our model, this corresponds to a setting with $k = 2$ classes, unknown class vectors $\mathbf{v}_1 = -\mathbf{v}_2$, where each view of an instance belongs to one class *fully* using membership function $f_i(\mathbf{x}) = \operatorname{sign}(\mathbf{v}_i \cdot \mathbf{x})$. Our work generalizes co-training by extending it to multi-class settings where each instance belongs to one or more classes *partially*, using a partial membership function $f_i(\cdot)$.

[1] We emphasize that linear independence is a much milder assumption than the assumption that topic vectors are orthogonal.

3 An Easier Case with Simplifying Assumptions

We make two main simplifying assumptions in this section, both of which will be relaxed in Section 4: 1) The documents are not noisy, i.e., $\mathbf{x}^1 \cdot \mathbf{v}_i = \mathbf{x}^2 \cdot \mathbf{v}_i$; 2) There is non-negligible probability density on instances that belong purely to one class. In this section we demonstrate ideas and techniques.

The Setting: We make the following assumptions. The documents are not noisy, that is for any document $(\mathbf{x}^1, \mathbf{x}^2)$ and for all $i \in [k]$, $\mathbf{x}^1 \cdot \mathbf{v}_i = \mathbf{x}^2 \cdot \mathbf{v}_i$. Regarding distribution $\mathcal{P}$, we assume that a non-negligible probability density is assigned to pure documents for each class. More formally, for some $\xi > 0$, for all $i \in [k]$, $\Pr_{\mathbf{w}\sim\mathcal{P}}[\mathbf{w} = \mathbf{e}_i] \geq \xi$. Regarding distribution $\mathcal{D}^{\mathbf{w}}$, we allow the two paragraphs in a document, i.e., the two views $(\mathbf{x}^1, \mathbf{x}^2)$ drawn from $\mathcal{D}^{\mathbf{w}}$, to be correlated as long as for any subspace $Z \subset \operatorname{null}\{\mathbf{v}_1 \ldots, \mathbf{v}_k\}$ of dimension strictly less than $n - k$, $\Pr_{(\mathbf{x}^1,\mathbf{x}^2)\sim\mathcal{D}^{\mathbf{w}}}[(\mathbf{x}^1 - \mathbf{x}^2) \notin Z] \geq \zeta$ for some non-negligible ζ. One way to view this in the context of topic modeling is that if, say, "sports" is a topic, then it should not be the case that the second paragraph always talks about the exact same sport as the first paragraph; else "sports" would really be a union of several separate but closely-related topics. Thus, while we do not require independence we do require some non-correlation between the paragraphs.

Algorithm and Analysis: The main idea behind our approach is to use the consistency of the two views of the samples to first recover the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_k$ (Phase 1). Once this subspace is recovered, we show that a projection of a sample on this space corresponds to the convex combination of class vectors using the appropriate mixture weight that was used for that sample. Therefore, we find vectors $\mathbf{a}_1, \ldots, \mathbf{a}_k$ that purely belong to each class by taking the extreme points of the projected samples (Phase 2). The class vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are the unique vectors (up to permutations) that classify $\mathbf{a}_1, \ldots, \mathbf{a}_k$ as pure samples. Phase 2 is similar to that of (Arora, Ge, and Moitra 2012). Algorithm 1 formalizes the details of this approach.

Algorithm 1 ALGORITHM FOR GENERALIZED TOPIC MODELS — NO NOISE

Input: A sample set $S = \{(\mathbf{x}_i^1, \mathbf{x}_i^2) \mid i \in [m]\}$ such that for each i, first a vector $\mathbf{w}$ is drawn from $\mathcal{P}$ and then $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ is drawn from $\mathcal{D}^{\mathbf{w}}$.

Phase 1: Let X^1 and X^2 be matrices where the i^{th} column is $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$, respectively. Let P be the projection matrix on the last k left singular vectors of $(X^1 - X^2)$.

Phase 2: Let $S_\parallel = \{P\mathbf{x}_i^j \mid i \in [m], j \in \{1, 2\}\}$. Let A be a matrix whose columns are the extreme points of the convex hull of $S_\parallel$. (This can be found using farthest traversal or linear programming.)[2]

Output: Return columns of A^+ as $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

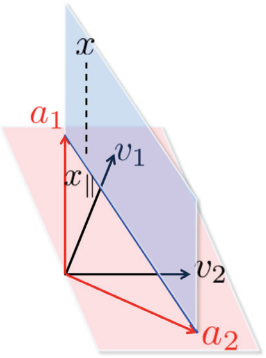

Figure 1: $\mathbf{v}_1, \mathbf{v}_2$ correspond to class 1 and 2, and $\mathbf{a}_1$ and $\mathbf{a}_2$ correspond to canonical vectors purely of class 1 and 2, respectively.

In Phase 1 for recovering $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, note that for any sample $(\mathbf{x}^1, \mathbf{x}^2)$ drawn from $\mathcal{D}^{\mathbf{w}}$, we have that $\mathbf{v}_i \cdot \mathbf{x}^1 = \mathbf{v}_i \cdot \mathbf{x}^2 = w_i$. Therefore, regardless of what $\mathbf{w}$ was used to produce the sample, we have that $\mathbf{v}_i \cdot (\mathbf{x}^1 - \mathbf{x}^2) = 0$ for all $i \in [k]$. That is, $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are in the null-space of all such $(\mathbf{x}^1 - \mathbf{x}^2)$. The assumptions on $\mathcal{D}^{\mathbf{w}}$ show that after seeing sufficiently many samples, $(\mathbf{x}_i^1 - \mathbf{x}_i^2)$ span a $n-k$ dimensional subspace. So, $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ can be recovered by taking $\operatorname{null}\{(\mathbf{x}^1 - \mathbf{x}^2) \mid (\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{X}^{\mathbf{w}} \times \mathcal{X}^{\mathbf{w}},\ \forall \mathbf{w} \in \mathbb{R}^k\}$. This null space is spanned by the last k singular vectors of $X^1 - X^2$, where X^1 and X^2 are matrices with columns $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$, respectively. The next lemma, whose proof appears in the full version of this paper (Blum and Haghtalab 2016), formalizes this discussion.

Lemma 3.1. *Let* $Z = \operatorname{span}\{(\mathbf{x}_i^1 - \mathbf{x}_i^2) \mid i \in [m]\}$. *Then,* $m = O(\frac{n-k}{\zeta} \log(\frac{1}{\delta}))$ *is sufficient such that with probability* $1 - \delta$, $\operatorname{rank}(Z) = n - k$.

Using Lemma 3.1, Phase 1 of Algorithm 1 recovers $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. Next, we show that pure samples are the extreme points of the convex hull of all samples when projected on the subspace $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. Figure 1 demonstrates the relation between the class vectors, $\mathbf{v}_i$, projection of samples, and the projection of pure samples $\mathbf{a}_i$. The next lemma, whose proof appears in the full version of this paper (Blum and Haghtalab 2016), formalizes this claim.

Lemma 3.2. *For any* $\mathbf{x}$, *let* $\mathbf{x}_\parallel$ *represent the projection of* $\mathbf{x}$ *on* $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. *Then,* $x_\parallel = \sum_{i\in[k]} (\mathbf{v}_i \cdot \mathbf{x})\mathbf{a}_i$.

With $\sum_{i\in[k]} (\mathbf{v}_i \cdot \mathbf{x})\mathbf{a}_i$ representing the projection of $\mathbf{x}$ on $\operatorname{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, it is clear that the extreme points of the set of all projected instances that belong to $\mathcal{X}^{\mathbf{w}}$ for all $\mathbf{w}$ are $\mathbf{a}_1, \ldots, \mathbf{a}_k$. Since in a large enough sample set, with high probability for all $i \in [k]$, there is a pure sample of type i, taking the extreme points of the set of projected samples is also $\mathbf{a}_1, \ldots, \mathbf{a}_k$. The following lemma, whose proof appears in the full version of this paper (Blum and Haghtalab 2016), formalizes this discussion.

Lemma 3.3. *Let* $m = c(\frac{1}{\xi} \log(\frac{k}{\delta}))$ *for a large enough constant* $c > 0$. *Let* P *be the projection matrix for*

$\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ *and* $S_{\parallel} = \{P\mathbf{x}_i^j \mid i \in [m], j \in \{1, 2\}\}$ *be the set of projected samples. With probability* $1 - \delta$, $\{\mathbf{a}_1, \ldots, \mathbf{a}_k\}$ *is the set of extreme points of* $\mathrm{CH}(S_{\parallel})$.

Therefore, $\mathbf{a}_1, \ldots, \mathbf{a}_k$ can be learned by taking the extreme points of the convex hull of all samples projected on $\mathrm{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})$. Furthermore, $V = A^+$ is unique, therefore $\mathbf{v}_1, \ldots, \mathbf{v}_k$ can be easily found by taking the pseudoinverse of matrix A. Together with Lemma 3.1 and 3.3 this proves the next theorem regarding learning class vectors in the absence of noise.

Theorem 3.4 (No Noise)**.** *There is a polynomial time algorithm for which* $m = O\left(\frac{n-k}{\zeta}\ln(\frac{1}{\delta}) + \frac{1}{\xi}\ln(\frac{k}{\delta})\right)$ *is sufficient to recover* $\mathbf{v}_i$ *exactly for all* $i \in [k]$, *with probability* $1-\delta$.

4 Relaxing the Assumptions

In this section, we relax the two main simplifying assumptions from Section 3. We relax the assumption on non-noisy documents and allow a large fraction of the documents to not satisfy $\mathbf{v}_i \cdot \mathbf{x}^1 = \mathbf{v}_i \cdot \mathbf{x}^2$. In the standard topic model, this corresponds to having a large fraction of short documents. Furthermore, we relax the assumption on the existence of pure documents to an assumption on the existence of "almost-pure" documents.

The Setting: We assume that any sampled document has a non-negligible probability of being non-noisy and with the remaining probability, the two views of the document are perturbed by additive Gaussian noise, independently. More formally, for a given sample $(\mathbf{x}^1, \mathbf{x}^2)$, with probability $p_0 > 0$ the algorithm receives $(\mathbf{x}^1, \mathbf{x}^2)$ and with the remaining probability $1 - p_0$, the algorithm receives $(\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2)$, such that $\hat{\mathbf{x}}^j = \mathbf{x}^j + \mathbf{e}^j$, where $\mathbf{e}^j \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$.

We assume that for each topic, the probability that a document is mostly about that topic is non-negligible. More formally, for any topic $i \in [k]$, $\Pr_{\mathbf{w}\sim\mathcal{P}}[\|\mathbf{e}_i - \mathbf{w}\|_1 \leq \epsilon\|] \geq g(\epsilon)$, where g is a polynomial function of its input. A stronger form of this assumption, better known as the *dominant admixture assumption*, assumes that every document is mostly about one topic and has been empirically shown to hold on several real world data sets (Bansal, Bhattacharyya, and Kannan 2014). Furthermore, in the Latent Dirichlet Allocation model, $\Pr_{\mathbf{w}\sim\mathcal{P}}[\max_{i\in[k]} w_i \geq 1-\epsilon] \geq O(\epsilon^2)$ for typical values of the concentration parameter.

We also make assumptions on the distribution over instances. We assume that the covariance of the distribution over $(\mathbf{x}_i^1 - \mathbf{x}_i^2)(\mathbf{x}_i^1 - \mathbf{x}_i^2)^\top$ is larger than the noise covariance σ^2.[3] That is, for some $\delta_0 > 0$, the least significant non-zero eigen value of $\mathbb{E}_{(\mathbf{x}_i^1, \mathbf{x}_i^2)}[(\mathbf{x}_i^1 - \mathbf{x}_i^2)(\mathbf{x}_i^1 - \mathbf{x}_i^2)^\top]$, equivalently its $(n-k)^{th}$ eigen value, is greater than $6\sigma^2 + \delta_0$. Moreover, we assume that the L_2 norm of each view of a sample is bounded by some $M > 0$. We also assume that for all $i \in [k]$, $\|\mathbf{a}_i\| \leq \alpha$ for some $\alpha > 0$. At a high level, $\|\mathbf{a}_i\|$s are inversely proportional to the non-zero singular values of $V = (\mathbf{v}_1, \ldots, \mathbf{v}_k)$. Therefore, $\|\mathbf{a}_i\| \leq \alpha$ implies that the k topic vectors are sufficiently different.

Algorithm and Results: Our approach follows the general theme of the previous section: First, recover $\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and then recover $\mathbf{a}_1, \ldots, \mathbf{a}_k$ by taking the extreme points of the projected samples. In this case, in the first phase we recover $\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ approximately, by finding a projection matrix $\hat{P}$ such that $\|P - \hat{P}\| \leq \epsilon$ for an arbitrarily small ϵ, where P is the projection matrix on $\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. At this point in the algorithm, the projection of samples on $\hat{P}$ can include points that are arbitrarily far from Δ. This is due to the fact that the noisy samples are perturbed by $\mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, so, for large values of σ some noisy samples map to points that are quite far from Δ. Therefore, we have to detect and remove these samples before continuing to the second phase. For this purpose, we show that the low density regions of the projected samples can safely be removed such that the convex hull of the remaining points is close to Δ. In the second phase, we consider projections of each sample using $\hat{P}$. To approximately recover $\mathbf{a}_1, \ldots, \mathbf{a}_k$, we recover samples, $\mathbf{x}$, that are far from the convex hull of the remaining points, when $\mathbf{x}$ and a ball of points close to it are removed. We then show that such points are close to one of the pure class vectors, $\mathbf{a}_i$. Algorithm 2 and the details of the above approach and its performance are as follows.

Theorem 4.1. *Consider any small enough* $\epsilon > 0$ *and any* $\delta > 0$, *there is an efficient algorithm for which an unlabeled sample set of size*

$$m = O\left(\frac{n-k}{\zeta}\ln(\frac{1}{\delta}) + \frac{n\sigma^2 M^4 r^2}{\delta_0^2 \epsilon^2}\mathrm{polylog}(\frac{nrM}{\epsilon\delta}) + \frac{k\ \ln(1/\delta)}{p_0\ g(\epsilon/(kr\alpha))}\right)$$

is sufficient to recover $\hat{\mathbf{a}}_i$ *such that* $\|\hat{\mathbf{a}}_i - \mathbf{a}_i\|_2 \leq \epsilon$ *for all* $i \in [k]$, *with probability* $1 - \delta$. *Where, r is a parameter that depends on the geometry of the simplex* Δ *and will be defined in section 4.3.*

The proof of Theorem 4.1 relies on the next lemmas regarding the performance of each phase of the algorithm. We formally state them here, but defer their proofs to Sections 4.1, 4.2 and 4.3.

Lemma 4.2 (Phase 1)**.** *For any* $\sigma, \epsilon > 0$, *it is sufficient to have an unlabeled sample set of size*

$$m = O\left(\frac{n-k}{\zeta}\ln(\frac{1}{\delta}) + \frac{n\sigma^2 M^2}{\delta_0^2 \epsilon^2}\mathrm{polylog}(\frac{n}{\epsilon\delta})\right).$$

so with probability $1-\delta$, *Phase 1 of Algorithm 2 returns a matrix* $\hat{P}$, *such that* $\|P - \hat{P}\|_2 \leq \epsilon$.

Lemma 4.3 (Denoising)**.** *Let* $\epsilon' \leq \frac{1}{3}\sigma\sqrt{k}$, $\|P - \hat{P}\| \leq \epsilon'/8M$, *and* $\gamma = g\left(\frac{\epsilon'}{8k\alpha}\right)$. *An unlabeled sample size of* $m =$

[3] This assumption is only used in Phase 1. One can assure that this assumption holds by taking the average of several documents in phase 1, where the average of documents $(\hat{\mathbf{x}}_1^1, \hat{\mathbf{x}}_1^2), \ldots, (\hat{\mathbf{x}}_m^1, \hat{\mathbf{x}}_m^2)$ is $\left(\sum_{i=1}^m \hat{\mathbf{x}}_i^1/m, \sum_{i=1}^m \hat{\mathbf{x}}_i^2/m\right)$. Since the noise shrinks in the averaged documents, the noise level falls under the required level. This would mildly increase the sample complexity.

Algorithm 2 ALGORITHM FOR GENERALIZED TOPIC MODELS — WITH NOISE

```
Input: A sample set $\{(\hat{\mathbf{x}}_i^1, \hat{\mathbf{x}}_i^2) \mid i \in [m]\}$ such that for each $i$, first a vector $\mathbf{w}$ is drawn from $\mathcal{P}$, then $(\mathbf{x}_i^1, \mathbf{x}_i^2)$ is drawn from $\mathcal{D}^{\mathbf{w}}$, then with probability $p_0$, $\hat{\mathbf{x}}_i^j = \mathbf{x}_i^j$, else with probability $1 - p_0$, $\hat{\mathbf{x}}_i^j = \mathbf{x}_i^j + \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ for $i \in [m]$ and $j \in \{1, 2\}$.
Phase 1:
1. Take $m_1 = \Omega\left(\frac{n-k}{\zeta}\ln(\frac{1}{\delta}) + \frac{n\sigma^2 M^4 r^2}{\delta_0^2 \epsilon^2}\text{polylog}(\frac{nrM}{\epsilon\delta})\right)$ samples.
2. Let $\hat{X}^1$ and $\hat{X}^2$ be matrices where the $i^{th}$ column is $\hat{\mathbf{x}}_i^1$ and $\hat{\mathbf{x}}_i^2$, respectively.
3. Let $\hat{P}$ be the projection matrix on the last $k$ left singular vectors of $\hat{X}^1 - \hat{X}^2$.
Denoising Phase:
4. Let $\epsilon' = \epsilon/(8r)$ and $\gamma = g\left(\epsilon'/(8k\alpha)\right)$.
5. Take $m_2 = \Omega\left(\frac{k}{p_0\gamma}\ln\frac{1}{\delta}\right)$ fresh samples and let $\hat{S}_{\parallel} = \left\{\hat{P}\hat{\mathbf{x}}_i^1 \mid \forall i \in [m_2]\right\}$.
6. Remove $\hat{\mathbf{x}}_{\parallel}$ from $\hat{S}_{\parallel}$, for which there are less than $\frac{p_0\gamma m_2}{2}$ points within distance of $\frac{\epsilon'}{2}$ in $\hat{S}_{\parallel}$.
Phase 2:
7. For all $\hat{\mathbf{x}}_{\parallel}$ in $\hat{S}_{\parallel}$, if $\text{dist}(\mathbf{x}_{\parallel}, \text{CH}(\hat{S}_{\parallel} \setminus B_{6r\epsilon'}(\hat{\mathbf{x}}))) \geq 2\epsilon'$ add $\hat{\mathbf{x}}_{\parallel}$ to $C$.
8. Cluster $C$ using single linkage with threshold $16r\epsilon'$. Assign any point from cluster $i$ as $\hat{\mathbf{a}}_i$.
Output: Return $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_k$.
```

$O\left(\frac{k}{p_0\gamma}\ln(\frac{1}{\delta})\right)$ *is sufficient such that for* $\hat{S}_{\parallel}$ *defined in Step 6 of Algorithm 2 the following holds with probability* $1 - \delta$*: For any* $\mathbf{x} \in \hat{S}_{\parallel}$, $\text{dist}(\mathbf{x}, \Delta) \leq \epsilon'$, *and, for all* $i \in [k]$, *there exists* $\hat{\mathbf{a}}_i \in \hat{S}_{\parallel}$ *such that* $\|\hat{\mathbf{a}}_i - \mathbf{a}_i\| \leq \epsilon'$.

Lemma 4.4 (Phase 2)**.** *Let* $\hat{S}_{\parallel}$ *be a set for which the conclusion of Lemma 4.3 holds with the value of* $\epsilon' = \epsilon/8r$. *Then, Phase 2 of Algorithm 2 returns* $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_k$ *such that for all* $i \in [k]$, $\|\mathbf{a}_i - \hat{\mathbf{a}}_i\| \leq \epsilon$.

We now prove our main Theorem 4.1 by directly leveraging the three lemmas we just stated.

Proof of Theorem 4.1. By Lemma 4.2, sample set of size m_1 is sufficient such that Phase 1 of Algorithm 2 leads to $\|P - \hat{P}\| \leq \frac{\epsilon}{32Mr}$, with probability $1 - \delta/2$. Let $\epsilon' = \frac{\epsilon}{8r}$ and take a fresh sample of size m_2. By Lemma 4.3, with probability $1 - \delta/2$, for any $\mathbf{x} \in \hat{S}_{\parallel}$, $\text{dist}(\mathbf{x}, \Delta) \leq \epsilon'$, and, for all $i \in [k]$, there exists $\hat{\mathbf{a}}_i \in \hat{S}_{\parallel}$ such that $\|\hat{\mathbf{a}}_i - \mathbf{a}_i\| \leq \epsilon'$. Finally, by Lemma 4.4 we have that Phase 2 of Algorithm 2 returns $\hat{\mathbf{a}}_i$, such that for all $i \in [k]$, $\|\mathbf{a}_i - \hat{\mathbf{a}}_i\| \leq \epsilon$. □

Theorem 4.1 discusses the approximation of $\mathbf{a}_i$ for all $i \in [k]$. It is not hard to see that such an approximation also translates to the approximation of class vectors, $\mathbf{v}_i$ for all $i \in [k]$. That is, using the properties of perturbation of pseudoinverse matrices (see the full version of the paper for details) one can show that $\|\hat{A}^+ - V\| \leq O(\|\hat{A} - A\|)$. Therefore, $\hat{V} = \hat{A}^+$ is a good approximation for V.

4.1 Proof of Lemma 4.2 — Phase 1

For $j \in \{1, 2\}$, let X^j and $\hat{X}^j$ be $n \times m$ matrices with the i^{th} column being $\mathbf{x}_i^j$ and $\hat{\mathbf{x}}_i^j$, respectively. As we demonstrated in Lemma 3.1, with high probability $\text{rank}(X^1 - X^2) = n - k$. Note that the nullspace of columns of $X^1 - X^2$ is spanned by the left singular vectors of $X^1 - X^2$ that correspond to its k zero singular values. We show that the nullspace of columns of $X^1 - X^2$ can be approximated within any desirable accuracy by the space spanned by the k least significant left singular vectors of $\hat{X}^1 - \hat{X}^2$, given a sufficiently large number of samples.

Let $D = X^1 - X^2$ and $\hat{D} = \hat{X}^1 - \hat{X}^2$. For ease of exposition, assume that all samples are perturbed by Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma^2 I_n)$.[4] Since each view of a sample is perturbed by an independent draw from a Gaussian noise distribution, we can view $\hat{D} = D + E$, where each column of E is drawn i.i.d from distribution $\mathcal{N}(\mathbf{0}, 2\sigma^2 I_n)$. Then, $\frac{1}{m}\hat{D}\hat{D}^\top = \frac{1}{m}DD^\top + \frac{1}{m}DE^\top + \frac{1}{m}ED^\top + \frac{1}{m}EE^\top$. As a thought experiment, consider this equation in expectation. Since $\mathbb{E}[\frac{1}{m}EE^\top] = 2\sigma^2 I_n$ is the covariance matrix of the noise and $\mathbb{E}[DE^\top + ED^\top] = 0$, we have

$$\frac{1}{m}\mathbb{E}\left[\hat{D}\hat{D}^\top\right] - 2\sigma^2 I_n = \frac{1}{m}\mathbb{E}\left[DD^\top\right]. \qquad (1)$$

Moreover, the eigen vectors and their order are the same in $\frac{1}{m}\mathbb{E}[\hat{D}\hat{D}^\top]$ and $\frac{1}{m}\mathbb{E}[\hat{D}\hat{D}^\top] - 2\sigma^2 I_n$. Therefore, one can recover the nullspace of $\frac{1}{m}\mathbb{E}[DD^\top]$ by taking the space of the smallest k eigen vectors of $\frac{1}{m}\mathbb{E}[\hat{D}\hat{D}^\top]$. Next, we show how to recover the nullspace using $\hat{D}\hat{D}^\top$, rather than $\mathbb{E}[\hat{D}\hat{D}^\top]$. Assume that the following properties hold:

1. Equation 1 holds not only in expectation, but also with high probability. That is, with high probability, $\|\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n - \frac{1}{m}DD^\top\|_2 \leq \epsilon$.
2. With high probability $\lambda_{n-k}(\frac{1}{m}\hat{D}\hat{D}^\top) > 4\sigma^2 + \delta_0/2$, where $\lambda_i(\cdot)$ denotes the i^{th} most significant eigen value.

Let $D = U\Sigma V^\top$ and $\hat{D} = \hat{U}\hat{\Sigma}\hat{V}^\top$ be SVD representations. We have that $\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n = \hat{U}(\frac{1}{m}\hat{\Sigma}^2 - 2\sigma^2 I_n)\hat{U}^\top$. By property 2, $\lambda_{n-k}(\frac{1}{m}\hat{\Sigma}^2) > 4\sigma^2 + \delta_0/2$. That is, the eigen vectors and their order are the same in $\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n$ and $\frac{1}{m}\hat{D}\hat{D}^\top$. As a result the projection matrix, $\hat{P}$, on the least significant k eigen vectors of $\frac{1}{m}\hat{D}\hat{D}^\top$, is the same as the projection matrix, Q, on the least significant k eigen vectors of $\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n$.

[4] The assumption that with a non-negligible probability a sample is non-noisy is not needed for the analysis and correctness of Phase 1 of Algorithm 2. This assumption only comes into play in the denoising phase.

Recall that $\hat{P}$ and P and Q are the projection matrices on the least significant k eigen vectors of $\frac{1}{m}\hat{D}\hat{D}^\top$, $\frac{1}{m}DD^\top$, and $\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I$, respectively. As we discussed, $\hat{P} = Q$. Now, using the Wedin $\sin\theta$ theorem (Davis and Kahan 1970; Wedin 1972) (see the full version of the paper for details) from matrix perturbation theory, we have,

$$\begin{aligned}\|P - \hat{P}\|_2 &= \|P - Q\| \\ &\le \frac{\|\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n - \frac{1}{m}DD^\top\|_2}{\left|\lambda_{n-k}(\frac{1}{m}\hat{D}\hat{D}^\top) - 2\sigma^2 - \lambda_{n-k+1}(\frac{1}{m}DD^\top)\right|} \le \frac{2\epsilon}{\delta_0},\end{aligned}$$

where we use Properties 1 and 2 and the fact that $\lambda_{n-k+1}(\frac{1}{m}DD^\top) = 0$, in the last transition.

Concentration It remains to prove Properties 1 and 2. We briefly describe our proof that when m is large, with high probability $\|\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n - \frac{1}{m}DD^\top\|_2 \le \epsilon$ and $\lambda_{n-k}(\frac{1}{m}\hat{D}\hat{D}^\top) > 4\sigma^2 + \delta_0/2$. Let us first describe $\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n - \frac{1}{m}DD^\top$ in terms of the error matrices. We have

$$\begin{aligned}\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n - \frac{1}{m}DD^\top &= \left(\frac{1}{m}EE^\top - 2\sigma^2 I_n\right) \\ &+ \left(\frac{1}{m}DE^\top + \frac{1}{m}ED^\top\right). \qquad (2)\end{aligned}$$

It suffices to show that for large enough $m > m_{\epsilon,\delta}$, $\Pr[\|\frac{1}{m}EE^\top - 2\sigma^2 I_n\|_2 \ge \epsilon] \le \delta$ and $\Pr[\|\frac{1}{m}DE^\top + \frac{1}{m}ED^\top\|_2 \ge \epsilon] \le \delta$. In the former, note that $\frac{1}{m}EE^\top$ is the sample covariance of the Gaussian noise matrix and $2\sigma^2 I_n$ is the true covariance matrix of the noise distribution. The next two claims follow by the convergence properties of sample covariance of the Gaussians and the use of Matrix Bernstein inequality (Tropp 2015). See the full version of this paper (Blum and Haghtalab 2016) for more details.

Claim 1. $m = O(\frac{n\sigma^4}{\epsilon^2}\log(\frac{1}{\delta}))$ *is sufficient to get* $\|\frac{1}{m}EE^\top - 2\sigma^2 I_n\|_2 \le \epsilon$, *with probability* $1-\delta$.

Claim 2. $m = O(\frac{n\sigma^2 M^2}{\epsilon^2}\text{polylog}\frac{n}{\epsilon\delta})$ *is sufficient to get* $\left\|\frac{1}{m}DE^\top + \frac{1}{m}ED^\top\right\|_2 \le \epsilon$, *with probability* $1-\delta$.

We prove that $\lambda_{n-k}(\frac{1}{m}\hat{D}\hat{D}^\top) > 4\sigma^2 + \delta_0/2$. Since for any two matrices, the difference in λ_{n-k} can be bounded by the spectral norm of their difference (see the full version of the paper for details), by Equation 2, we have

$$\begin{aligned}&\left|\lambda_{n-k}\left(\frac{1}{m}\hat{D}\hat{D}^\top\right) - \lambda_{n-k}\left(\frac{1}{m}DD^\top\right)\right| \\ &\le \left\|2\sigma^2 I + \left(\frac{1}{m}EE^\top - 2\sigma^2 I_n\right) - \left(\frac{1}{m}DE^\top + \frac{1}{m}ED^\top\right)\right\| \\ &\le 2\sigma^2 + \frac{\delta_0}{4},\end{aligned}$$

where in the last transition we use Claims 1 and 2 with the value of $\delta_0/8$ to bound the last two terms by a total of $\delta_0/4$. Since $\lambda_{n-k}(\mathbb{E}[\frac{1}{m}DD^\top]) \ge 6\sigma^2 + \delta_0$, it is sufficient to show that $|\lambda_{n-k}(\mathbb{E}[\frac{1}{m}DD^\top]) - \lambda_{n-k}([\frac{1}{m}DD^\top])| \le \delta_0/4$. Similarly as before, this is bounded by $\|\frac{1}{m}DD^\top - \mathbb{E}[\frac{1}{m}DD^\top]\|$. We use the Matrix Bernstein inequality to prove this concentration result; see the full version of this paper (Blum and Haghtalab 2016) for a proof.

Claim 3. $m = O\left(\frac{M^4}{\delta_0^2}\log\frac{n}{\delta}\right)$ *is sufficient to get* $\left\|\frac{1}{m}DD^\top - \mathbb{E}\left[\frac{1}{m}DD^\top\right]\right\|_2 \le \frac{\delta_0}{4}$, *with probability* $1-\delta$.

This completes the analysis of Phase 1 of our algorithm and the proof of Lemma 4.2 follows directly from the above analysis and the application of Claims 1 and 2 with the error of $\epsilon\delta_0$, and Claim 3.

4.2 Proof of Lemma 4.3 — Denoising Step

We use projection matrix $\hat{P}$ to partially denoise the samples while approximately preserving $\Delta = \mathrm{CH}(\{\mathbf{a}_1, \dots, \mathbf{a}_k\})$. At a high level we show that, in the projection of samples on $\hat{P}$, 1) the regions around $\mathbf{a}_i$ have sufficiently high density, and, 2) the regions that are far from Δ have low density.

We claim that if $\hat{x}_\parallel \in \hat{S}_\parallel$ is *non-noisy and corresponds almost purely to one class* then $\hat{S}_\parallel$ also includes a non-negligible number of points within $O(\epsilon')$ distance of $\hat{x}_\parallel$. This is due to the fact that a non-negligible number of points (about $p_0\gamma m$ points) correspond to non-noisy and almost-pure samples that using P would get projected to points within a distance of $O(\epsilon')$ of each other. Furthermore, the inaccuracy in $\hat{P}$ can only perturb the projections up to $O(\epsilon')$ distance. So, the projections of all non-noisy samples that are almost purely of class i fall within $O(\epsilon')$ of $\mathbf{a}_i$. The following claim, whose proof appears in the full version of this paper (Blum and Haghtalab 2016), formalizes this discussion.

In the following lemmas, let D denote the flattened distribution of the first paragraphs. That is, the distribution over $\hat{\mathbf{x}}^1$ where we first take $\mathbf{w} \sim \mathcal{P}$, then take $(\mathbf{x}^1, \mathbf{x}^2) \sim \mathcal{D}^{\mathbf{w}}$, and finally take $\hat{\mathbf{x}}^1$.

Claim 4. *For all* $i \in [k]$, $\Pr_{\mathbf{x}\sim D}\left[\hat{P}\mathbf{x} \in B_{\epsilon'/4}(\mathbf{a}_i)\right] \ge p_0\gamma$.

On the other hand, any projected point that is far from the convex hull of $\mathbf{a}_1, \dots, \mathbf{a}_k$ has to be noisy, and as a result, has been generated by a Gaussian distribution with variance σ^2. For a choice of ϵ' that is small with respect to σ, such points do not concentrate well within any ball of radius ϵ'. In the next claim, we show that the regions that are far from the convex hull have low density.

Claim 5. *For any* $\mathbf{z}$ *such that* $\mathrm{dist}(\mathbf{z}, \Delta) \ge \epsilon'$, *we have* $\Pr_{\mathbf{x}\sim D}\left[\hat{P}\mathbf{x} \in B_{\epsilon'/2}(\mathbf{z})\right] \le \frac{p_0\gamma}{4}$.

The next claim shows that in a large sample set, the fraction of samples that fall within any of the described regions in Claims 4 and 5 is close to the density of that region. The proof of this claim follows from VC dimension of the set of balls.

Claim 6. *Let* D *be any distribution over* $\mathbb{R}^k$ *and* $\mathbf{x}_1, \dots, \mathbf{x}_m$ *be* m *points drawn i.i.d from* D. *Then* $m = O(\frac{k}{\gamma}\ln\frac{1}{\delta})$ *is sufficient so that with probability* $1-\delta$, *for any ball* $B \subseteq \mathbb{R}^k$ *such that* $\Pr_{\mathbf{x}\sim D}[\mathbf{x} \in B] \ge 2\gamma$, $|\{\mathbf{x}_i \mid \mathbf{x}_i \in B\}| > \gamma m$ *and for any ball* $B \subseteq \mathbb{R}^k$ *such that* $\Pr_{\mathbf{x}\sim D}[\mathbf{x} \in B] \le \gamma/2$, $|\{\mathbf{x}_i \mid \mathbf{x}_i \in B\}| < \gamma m$.

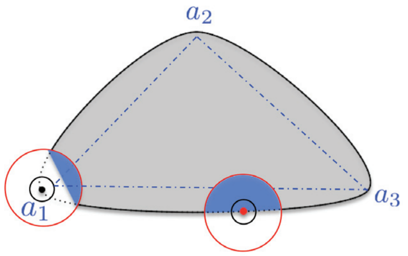

Figure 2: Demonstrating the distinction between points close to $\mathbf{a}_i$ and far from $\mathbf{a}_i$. The convex hull of $CH(\hat{S}_{||} \setminus B_{r_2}(\hat{\mathbf{x}}))$, which is a subset of the blue and gray region, intersects $B_{r_1}(\hat{\mathbf{x}})$ only for $\hat{\mathbf{x}}$ that is sufficiently far from $\mathbf{a}_i$'s.

Therefore, upon seeing $\Omega(\frac{k}{p_0\gamma} \ln \frac{1}{\delta})$ samples, with probability $1-\delta$, for all $i \in [k]$ there are more than $p_0\gamma m/2$ projected points within distance $\epsilon'/4$ of $\mathbf{a}_i$ (by Claims 4 and 6), and, no point that is ϵ' far from Δ has more than $p_0\gamma m/2$ points in its $\epsilon'/2$-neighborhood (by Claims 5 and 6). Phase 2 of Algorithm 2 leverages these properties of the set of projected points for denoising the samples while preserving Δ: Remove any point from $\hat{S}_\parallel$ with fewer than $p_0\gamma m/2$ neighbors within distance $\epsilon'/2$.

We conclude the proof of Lemma 4.3 by noting that the remaining points in $\hat{S}_\parallel$ are all within distance ϵ' of Δ. Furthermore, any point in $B_{\epsilon'/4}(\mathbf{a}_i)$ has more than $p_0\gamma m/2$ points within distance of $\epsilon'/2$. Therefore, such points remain in $\hat{S}_\parallel$ and any one of them can serve as $\hat{\mathbf{a}}_i$ for which $\|\mathbf{a}_i - \hat{\mathbf{a}}_i\| \leq \epsilon'/4$.

4.3 Proof of Lemma 4.4 — Phase 2

At a high level, we consider two balls around each projected sample point $\hat{\mathbf{x}} \in \hat{S}_\parallel$ with appropriate choice of radii $r_1 < r_2$ (see Figure 2). Consider the set of projections $\hat{S}_\parallel$ when points in $B_{r_2}(\mathbf{x})$ are removed from it. For points that are far from all $\mathbf{a}_i$, this set still includes points that are close to $\mathbf{a}_i$ for all topics $i \in [k]$. So, the convex hull of $\hat{S}_\parallel \setminus B_{r_2}(\mathbf{x})$ is close to Δ, and in particular, intersects $B_{r_1}(\mathbf{x})$. On the other hand, for $\mathbf{x}$ that is close to $\mathbf{a}_i$, $\hat{S}_\parallel \setminus B_{r_2}(\mathbf{x})$ does not include an extreme point of Δ or points close to it. So, the convex hull of $\hat{S}_\parallel \setminus B_{r_2}(\mathbf{x})$ is considerably smaller than Δ, and in particular, does not intersect $B_{r_1}(\mathbf{x})$.

The geometry of the simplex and the angles between $\mathbf{a}_1, \dots, \mathbf{a}_k$ play an important role in choosing the appropriate r_1 and r_2. Note that when the samples are perturbed by noise, $\mathbf{a}_1, \dots, \mathbf{a}_k$ can only be approximately recovered if they are sufficiently far apart and the angles of the simplex at each $\mathbf{a}_i$ is far from being flat. That is, we assume that for all $i \neq j$, $\|\mathbf{a}_i - \mathbf{a}_j\| \geq 3\epsilon$. Furthermore, define $r \geq 1$ to be the smallest value such that the distance between $\mathbf{a}_i$ and $\mathrm{CH}(\Delta \setminus B_{r\epsilon}(\mathbf{a}_i))$ is at least ϵ. Note that such a value of r always exists and depends entirely on the angles of the simplex defined by the class vectors. Therefore, the number of samples needed for our method depends on the value of r. The smaller the value of r, the larger is the separation between the topic vectors and the easier it is to identify them (See Figure 3). The next claim, whose proof appears in the full version of this paper (Blum and Haghtalab 2016), demonstrates this concept.

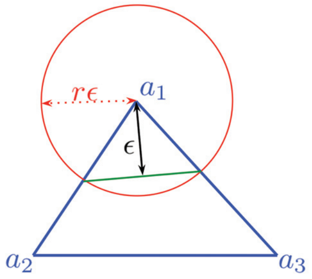

Figure 3: Parameter r is determined by the geometry of Δ.

Claim 7. *Let $\epsilon' = \epsilon/8r$. Let $\hat{S}_\parallel$ be the set of denoised projections, as in step 6 of Algorithm 2. For any $\hat{\mathbf{x}} \in \hat{S}_\parallel$ such that for all i, $\|\hat{\mathbf{x}} - \mathbf{a}_i\| > 8r\epsilon'$, $\mathrm{dist}(\hat{\mathbf{x}}, \mathrm{CH}(\hat{S}_\parallel \setminus B_{6r\epsilon'}(\hat{\mathbf{x}}))) \leq 2\epsilon'$. Furthermore, for all $i \in [k]$ there exists $\hat{\mathbf{a}}_i \in \hat{S}_\parallel$ such that $\|\hat{\mathbf{a}}_i - \mathbf{a}_i\| < \epsilon'$ and $\mathrm{dist}(\hat{\mathbf{a}}_i, \mathrm{CH}(\hat{S}_\parallel \setminus B_{6r\epsilon'}(\hat{\mathbf{a}}_i))) > 2\epsilon'$.*

Given the above structure, it is clear that set of points in C are all within ϵ of one of the $\mathbf{a}_i$'s. So, we can cluster C using single linkage with threshold ϵ to recover $\mathbf{a}_i$ upto accuracy ϵ.

5 Additional Results and Extensions

In this section, we briefly mention some additional results and extensions. We explain these and discuss other extensions (such as alternative noise models) in more detail in the full version of this paper (Blum and Haghtalab 2016).

Sample Complexity Lower bound As we observed the number of samples required by our method is $\mathrm{poly}(n)$. However, as the number of classes can be much smaller than the number of features, one might hope to recover $\mathbf{v}_1, \dots, \mathbf{v}_k$, with a number of samples that is polynomial in k rather than n. Here, we show that in the general case $\Omega(n)$ samples are needed to learn $\mathbf{v}_1, \dots, \mathbf{v}_k$ regardless of the value of k. See, the full version of this paper (Blum and Haghtalab 2016) for more information.

General function $f(\cdot)$ We also consider the general model described in Section 2, where $f_i(x) = f(\mathbf{v}_i \cdot \mathbf{x})$ for an unknown strictly increasing function $f : \mathbb{R}^+ \rightarrow [0, 1]$ such that $f(0) = 0$. We describe how variations of the techniques discussed up to now can extend to this more general setting. See, the full version of this paper (Blum and Haghtalab 2016) for more information.

Alternative Noise Models We also discuss two additional noise models and interesting open problems that arise in these settings. In the first model, we consider the problem of recovering $\mathbf{v}_1, \dots, \mathbf{v}_k$ in the presence of agnostic noise, where for an ϵ fraction of the samples $(\mathbf{x}^1, \mathbf{x}^2)$, $\mathbf{x}^1$ and $\mathbf{x}^2$ correspond to different mixture weights. In the second model, we consider the case of $p_0 = 0$. That is, when *every document* is affected by Gaussian noise $\mathcal{N}(0, \sigma^2 I_n)$, for $\sigma \gg \epsilon$.

Acknowledgments

This work was partially supported by NSF grants CCF-1525971, CCF-1535967, CCF-1800317; a Microsoft Research Ph.D. Fellowship and an IBM Ph.D. Fellowship.

References

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

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

Arora, S.; Ge, R.; Kannan, R.; and Moitra, A. 2012. Computing a nonnegative matrix factorization–provably. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC)*, 145–162.

Arora, S.; Ge, R.; Halpern, Y.; Mimno, D. M.; Moitra, A.; Sontag, D.; Wu, Y.; and Zhu, M. 2013. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 280–288.

Arora, S.; Ge, R.; and Moitra, A. 2012. Learning topic models – going beyond svd. In *Proceedings of the 53rd Symposium on Foundations of Computer Science (FOCS)*, 1–10.

Balcan, M.-F., and Blum, A. 2010. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)* 57(3):19.

Balcan, M.-F.; Blum, A.; and Yang, K. 2004. Co-training and expansion: Towards bridging theory and practice. In *Advances in Neural Information Processing Systems (NIPS)*, 89–96.

Bansal, T.; Bhattacharyya, C.; and Kannan, R. 2014. A provable svd-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems (NIPS)*, 1997–2005.

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

Blum, A., and Haghtalab, N. 2016. Generalized topic modeling. *arXiv preprint arXiv:1611.01259*.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Conference on Computational Learning Theory (COLT)*, 92–100.

Chapelle, O.; Schlkopf, B.; and Zien, A. 2010. *Semi-Supervised Learning*. The MIT Press, 1st edition.

Dasgupta, S.; Littman, M. L.; and McAllester, D. 2002. Pac generalization bounds for co-training. *Advances in Neural Information Processing Systems (NIPS)* 1:375–382.

Davis, C., and Kahan, W. M. 1970. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Computing* 7(1):1–46.

Hofmann, T. 1999. Probabilistic latent semantic analysis. In *Proceedings of the 15th conference on Uncertainty in artificial intelligence*, 289–296.

Papadimitriou, C. H.; Tamaki, H.; Raghavan, P.; and Vempala, S. 1998. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, 159–168.

Sun, S. 2013. A survey of multi-view machine learning. *Neural computing and applications* 23:2031–2038.

Tropp, J. A. 2015. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*.

Wedin, P.-Å. 1972. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics* 12(1):99–111.