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 [4, 5, 3, 1, 2, 8]. 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 [5, 1, 18]. In addition, algorithms based on this problem formulation perform well empirically on standard datasets [9, 15].

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* [10, 12, 11, 7, 21]. 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 $f_1, \dots, f_k : \mathcal{X} \rightarrow [0, 1]$, 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, the learner receives samples $(\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2)$ that have been perturbed by some noise. We discuss two noise models in Sections 4 and F.2. 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 $\mathbf{v}_1, \dots, \mathbf{v}_k \in \mathbb{R}^n$ are linearly independent and $f : \mathbb{R} \rightarrow [0, 1]$ 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 $\mathbf{a}_i \in \operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$ 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 $\operatorname{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$. Define $\Delta = \operatorname{CH}(\{\mathbf{a}_1, \dots, \mathbf{a}_k\})$ to be the convex hull of $\mathbf{a}_1, \dots, \mathbf{a}_k$.

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 $V = (\mathbf{v}_1, \dots, \mathbf{v}_k)$ is the pseudo-inverse of $A = (\mathbf{a}_1, \dots, \mathbf{a}_k)$. 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 $\mathbf{a}_1, \dots, \mathbf{a}_k$.

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 [5, 2], 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 [10] 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, [10] and [6] 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 $f_i(\cdot)$.

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

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

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 $Z \subset \text{null}\{\mathbf{v}_1 \ldots, \mathbf{v}_k\}$ 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 [5]. 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 $\mathbf{v}_1, \ldots, \mathbf{v}_k$.

In Phase 1 for recovering $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, 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, $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ can be recovered by taking $\text{null}\{(\mathbf{x}^1 - \mathbf{x}^2) \mid (\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{X}^{\mathbf{w}} \times \mathcal{X}^{\mathbf{w}}, \forall \mathbf{w} \in \mathbb{R}^k\}$. 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 (see Appendix A.1 for a proof) formalizes this discussion.

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

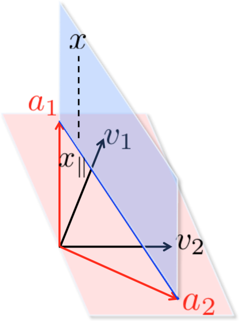

Figure 1: $\mathbf{v}_1, \mathbf{v}_2$ correspond to class 1 and 2, and $\mathbf{a}_1$ and $\mathbf{a}_2$ correspond to canonical vectors purely of class 1 and 2, respectively.

Using Lemma 3.1, Phase 1 of Algorithm 1 recovers $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. Next, we show that pure samples are the extreme points of the convex hull of all samples when projected on the subspace $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. 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 Appendix A.2, formalizes this claim.

Lemma 3.2. *For any $\mathbf{x}$, let $\mathbf{x}_\parallel$ represent the projection of $\mathbf{x}$ on $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. 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 $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, 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 Appendix A.3, 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 $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ 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 $\text{CH}(S_\parallel)$.*

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 $\text{span}(\{\mathbf{v}_1, \ldots, \mathbf{v}_k\})$. 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 [8]. 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.

```
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 $m_1 = \Omega\left(\frac{n-k}{\zeta}\ln(\frac{1}{\delta}) + \frac{n\sigma^2 M^4 r^2}{\delta_0^2 \epsilon^2}\mathrm{polylog}(\frac{nrM}{\epsilon\delta})\right)$ 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 $\mathrm{dist}(\mathbf{x}_\parallel, \mathrm{CH}(\hat{S}_\parallel \setminus B_{6r\epsilon'}(\hat{\mathbf{x}}))) \geq 2\epsilon'$ 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 $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_k$.
```

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 $\|P - \hat{P}\|_2 \leq \epsilon$.

[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.

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 $m = O\left(\frac{k}{p_0\gamma}\ln(\frac{1}{\delta})\right)$ is sufficient such that for $\hat{S}_{\parallel}$ defined in Step 6 of Algorithm 2 the following holds with probability $1-\delta$: For any $\mathbf{x} \in \hat{S}_{\parallel}$, $\operatorname{dist}(\mathbf{x}, \Delta) \leq \epsilon'$, 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 $\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_k$ 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 $\mathbf{x} \in \hat{S}_{\parallel}$, $\operatorname{dist}(\mathbf{x}, \Delta) \leq \epsilon'$, 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 Proposition B.5) 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 $\operatorname{rank}(X^1 - X^2) = n - k$. 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

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

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 $\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n$.

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 [13, 24] (see Proposition B.1) from matrix perturbation theory, we have,

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

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 $\|\frac{1}{m}\hat{D}\hat{D}^\top - 2\sigma^2 I_n - \frac{1}{m}DD^\top\|_2 \leq \epsilon$ 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

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

It suffices to show that for large enough $m > m_{\epsilon,\delta}$, $\Pr[\|\frac{1}{m}EE^\top - 2\sigma^2 I_n\|_2 \geq \epsilon] \leq \delta$ and $\Pr[\|\frac{1}{m}DE^\top + \frac{1}{m}ED^\top\|_2 \geq \epsilon] \leq \delta$. 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

[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.

covariance of the Gaussians and the use of Matrix Bernstein inequality [22] (Appendix B). See, Appendix C.1 for a proof.

Claim 1. *$m = O(\frac{n\sigma^4}{\epsilon^2}\log(\frac{1}{\delta}))$ is sufficient to get $\|\frac{1}{m}EE^\top - 2\sigma^2 I_n\|_2 \leq \epsilon$, 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 $\left\|\frac{1}{m}DE^\top{+}\frac{1}{m}ED^\top\right\|_2{\leq}\epsilon$, 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 Proposition B.4), by Equation 2, we have

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

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 $\lambda_{n-k}(\mathbb{E}[\frac{1}{m}DD^\top]) \geq 6\sigma^2 + \delta_0$, it is sufficient to show that $|\lambda_{n-k}(\mathbb{E}[\frac{1}{m}DD^\top]) - \lambda_{n-k}([\frac{1}{m}DD^\top])| \leq \delta_0/4$. 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 Appendix C.2 for a proof.

Claim 3. *$m{=}O\left(\frac{M^4}{\delta_0^2}\log\frac{n}{\delta}\right)$ is sufficient to get $\left\|\frac{1}{m}DD^\top{-}\mathbb{E}\left[\frac{1}{m}DD^\top\right]\right\|_2 \leq \frac{\delta_0}{4}$, 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 $\Delta = \text{CH}(\{\mathbf{a}_1, \ldots, \mathbf{a}_k\})$. 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 Appendix D.1, 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 $i \in [k]$, $\Pr_{\mathbf{x}\sim D}\left[\hat{P}\mathbf{x} \in B_{\epsilon'/4}(\mathbf{a}_i)\right] \geq p_0\gamma$.*

On the other hand, any projected point that is far from the convex hull of $\mathbf{a}_1, \ldots, \mathbf{a}_k$ 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 $\text{dist}(\mathbf{z}, \Delta) \geq \epsilon'$, we have $\Pr_{\mathbf{x}\sim D}\left[\hat{P}\mathbf{x} \in B_{\epsilon'/2}(\mathbf{z})\right] \leq \frac{p_0\gamma}{4}$.*

Proof. We first show that $B_{\epsilon'/2}(\mathbf{z})$ does not include any non-noisy points. Take any non-noisy sample $\mathbf{x}$. Note that $P\mathbf{x} = \sum_{i=1}^k w_i\mathbf{a}_i$, where w_i are the mixture weights corresponding to point $\mathbf{x}$. We have,

$$
\begin{aligned}
\left\|\mathbf{z} - \hat{P}\mathbf{x}\right\| &= \left\|\mathbf{z} - \sum_{i=1}^k w_i\mathbf{a}_i + (P - \hat{P})\mathbf{x}\right\| \\
&\geq \left\|\mathbf{z} - \sum_{i=1}^k w_i\mathbf{a}_i\right\| - \|P - \hat{P}\|\|\mathbf{x}\| \geq \epsilon'/2
\end{aligned}
$$

Therefore, $B_{\epsilon'/2}(\mathbf{z})$ only contains noisy points. Since noisy points are perturbed by a spherical Gaussian, the projection of these points on any k-dimensional subspace can be thought of points generated from a k-dimensional Gaussian distributions with variance σ^2 and potentially different centers. One can show that the densest ball of any radius is at the center of a Gaussian. Here, we prove a slightly weaker claim. Consider one such Gaussian distribution, $\mathcal{N}(\mathbf{0}, \sigma^2 I_k)$.

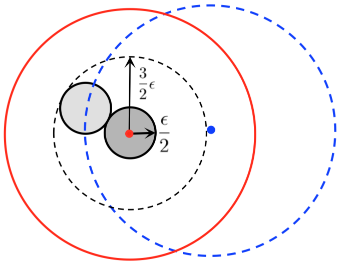

Figure 2: Density is maximized when blue and red gaussians coincide and the ball is at their center.

Note that the pdf of the Gaussian distribution decreases as we get farther from its center. By a coupling between the density of the points, $B_{\epsilon'/2}(\mathbf{0})$ has higher density than any $B_{\epsilon'/2}(\mathbf{c})$ with $\|\mathbf{c}\|_2 > \epsilon'$. Therefore,

$$\sup_{\mathbf{c}} \Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\sigma^2 I_k)}[\mathbf{x} \in B_{\epsilon'/2}(\mathbf{c})] \leq \Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\sigma^2 I_k)}[\mathbf{x} \in B_{3\epsilon'/2}(\mathbf{0})].$$

So, over D this value will be maximized if the Gaussians had the same center (see Figure 2). Moreover, in $\mathcal{N}(\mathbf{0}, \sigma^2 I_k)$, $\Pr[\|\mathbf{x}\|_2 \leq \sigma\sqrt{k(1-t)}] \leq \exp(-kt^2/16)$. Since $3\epsilon'/2 \leq \sigma\sqrt{k}/2 \leq \sigma\sqrt{k(1 - \sqrt{\frac{16}{k}\ln\frac{4}{p_0\gamma}}})$ we have

$$\Pr_{\hat{\mathbf{x}}\sim D}[\mathbf{x} \in B_{\epsilon'/2}(\mathbf{c})] \leq \Pr_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},\sigma^2 I_k)}[\|\mathbf{x}\|_2 \leq 3\epsilon'/2] \leq \frac{p_0\gamma}{4}.$$

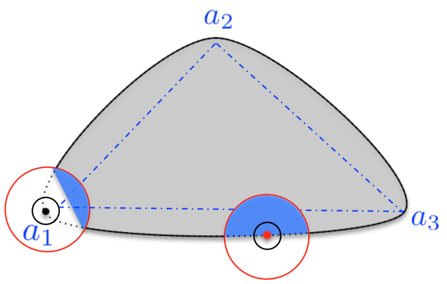

Figure 3: Demonstrating the distinction between points close to $\mathbf{a}_i$ and far from $\mathbf{a}_i$. The convex hull of $CH(\hat{S}_{||} \setminus B_{r_2}(\hat{\mathbf{x}}))$, which is a subset of the blue and gray region, intersects $B_{r_1}(\hat{\mathbf{x}})$ only for $\hat{\mathbf{x}}$ that is sufficiently far from $\mathbf{a}_i$'s.

□

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 $\Pr_{\mathbf{x} \sim D}[\mathbf{x} \in B] \geq 2\gamma$, $|\{\mathbf{x}_i \mid \mathbf{x}_i \in B\}| > \gamma m$ and for any ball $B \subseteq \mathbb{R}^k$ such that $\Pr_{\mathbf{x} \sim D}[\mathbf{x} \in B] \leq \gamma/2$, $|\{\mathbf{x}_i \mid \mathbf{x}_i \in B\}| < \gamma m$.*

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 3). 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 $B_{r_1}(\mathbf{x})$.

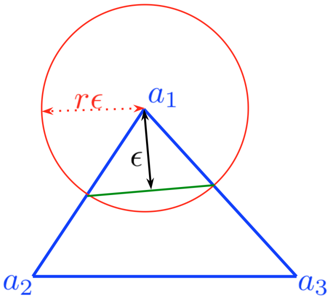

Figure 4: Parameter r is determined by the geometry of Δ.

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. The next claim, whose proof appears in Appendix E.1, demonstrates this concept.

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 Appendix F.

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, Appendix F 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 $f : \mathbb{R}^+ \to [0, 1]$ such that $f(0) = 0$. We describe how variations of the techniques discussed up to now can extend to this more general setting. See Appendix F for more information.

References

[1] Anima Anandkumar, Yi-kai Liu, Daniel J Hsu, Dean P Foster, and Sham M Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925. 2012.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[3] Sanjeev Arora, Rong Ge, Yonatan Halpern, David M Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 280–288, 2013.

[4] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization–provably. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC)*, pages 145–162. ACM, 2012.

[5] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models – going beyond svd. In *Proceedings of the 53rd Symposium on Foundations of Computer Science (FOCS)*, pages 1–10, 2012.

[6] Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):19, 2010.

[7] Maria-Florina Balcan, Avrim Blum, and Ke Yang. Co-training and expansion: Towards bridging theory and practice. In *Advances in neural information processing systems*, pages 89–96, 2004.

[8] Trapit Bansal, Chiranjib Bhattacharyya, and Ravindran Kannan. A provable svd-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems*, pages 1997–2005, 2014.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Conference on Computational Learning Theory (COLT)*, pages 92–100. ACM, 1998.

[11] Olivier Chapelle, Bernhard Schlkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 1st edition, 2010.

[12] Sanjoy Dasgupta, Michael L Littman, and David McAllester. Pac generalization bounds for co-training. *Advances in neural information processing systems*, 1:375–382, 2002.

[13] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Computing*, 7(1):1–46, 1970.

[14] Moritz Hardt and Ankur Moitra. Algorithms and hardness for robust subspace recovery. In *Proceedings of the 26th Conference on Computational Learning Theory (COLT)*, pages 354–375, 2013.

[15] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc., 1999.

[16] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Disentangling gaussians. *Communications of the ACM*, 55(2):113–120, 2012.

[17] Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

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

[19] G. W. Stewart. On the perturbation of pseudo-inverses, projections and linear least squares problems. *SIAM Review*, 19(4):634–662, 1977.

[20] GW Stewart and Ji-Guang Sun. Matrix perturbation theory (computer science and scientific computing), 1990.

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

[22] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

[23] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[24] Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972.

A Omitted Proof from Section 3 — No Noise

A.1 Proof of Lemma 3.1

For all $j \leq n-k$, let $Z_j = \{(\mathbf{x}_i^1 - \mathbf{x}_i^2) \mid i \leq \frac{j}{\zeta} \ln \frac{n}{\delta}\}$. We prove by induction that for all j, $\text{rank}(Z_j) < j$ with probability at most $j\frac{\delta}{n}$.

For $j = 0$, the claim trivially holds. Now assume that the induction hypothesis holds for some j. Furthermore, assume that $\text{rank}(Z_j) \geq j$. Then, $\text{rank}(Z_{j+1}) < j+1$ only if the additional $\frac{1}{\zeta} \ln \frac{n}{\delta}$ samples in Z_{j+1} all belong to $\text{span}(Z_j)$. Since, the space of such samples has rank $< n-k$, this happens with probability at most $(1-\zeta)^{\frac{1}{\zeta}\ln\frac{n}{\delta}} \leq \frac{\delta}{n}$. Together with the induction hypothesis that $\text{rank}(Z_j) \geq j$ with probability at most $j\frac{\delta}{n}$, we have that $\text{rank}(Z_{j+1}) < j+1$ with probability at most $\frac{(j+1)\delta}{n}$. Therefore $\text{rank}(Z) = \text{rank}(Z_{n-k}) = n-k$ with probability at least $1-\delta$.

A.2 Proof of Lemma 3.2

First note that V is a the pseudo-inverse of A, so their span is equal. Hence, $\sum_{i\in[k]} (\mathbf{v}_i \cdot \mathbf{x})\mathbf{a}_i \in \text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$. It remains to show that $\left(\mathbf{x} - \sum_{i\in[k]}(\mathbf{v}_i \cdot \mathbf{x})\mathbf{a}_i\right) \in \text{null}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$. We do so by showing that this vector is orthogonal to $\mathbf{v}_j$ for all j. We have

$$\left(\mathbf{x} - \sum_{i=1}^{k}(\mathbf{v}_i \cdot \mathbf{x})\mathbf{a}_i\right) \cdot \mathbf{v}_j = \mathbf{x} \cdot \mathbf{v}_j - \sum_{i=1}^{k}(\mathbf{v}_i \cdot \mathbf{x})(\mathbf{a}_i \cdot \mathbf{v}_j)$$
$$= \mathbf{x} \cdot \mathbf{v}_j - \sum_{i \neq j}(\mathbf{v}_i \cdot \mathbf{x})(\mathbf{a}_i \cdot \mathbf{v}_j) - (\mathbf{v}_j \cdot \mathbf{x})(\mathbf{a}_j \cdot \mathbf{v}_j)$$
$$= \mathbf{x} \cdot \mathbf{v}_j - \mathbf{x} \cdot \mathbf{v}_j = 0.$$

Where, the second equality follows from the fact when $A = V^+$, for all i, $\mathbf{a}_i \cdot \mathbf{v}_i = 1$ and $\mathbf{a}_j \cdot \mathbf{v}_i =$ for $j \neq i$. Therefore, $\sum_{i\in[k]}(\mathbf{v}_i \cdot \mathbf{x})\mathbf{a}_i$ is the projection of $\mathbf{x}$ on $\text{span}\{\mathbf{v}_1, \dots, \mathbf{v}_k\}$.

A.3 Proof of Lemma 3.3

Assume that S included samples that are purely of type i, for all $i \in [k]$. That is, for all $i \in [k]$ there is $j \leq m$, such that $\mathbf{v}_i \cdot \mathbf{x}_j^1 = \mathbf{v}_i \cdot \mathbf{x}_j^2 = 1$ and $\mathbf{v}_{i'} \cdot \mathbf{x}_j^1 = \mathbf{v}_{i'} \cdot \mathbf{x}_j^2 = 0$ for $i' \neq i$. By Lemma 3.2, the set of projected vectors form the set $\{\sum_{i=1}^k (\mathbf{v}_i \cdot \mathbf{x}_j)\mathbf{a}_i \mid j \in [m]\}$. Note that $\sum_{i=1}^k (\mathbf{v}_i \cdot \mathbf{x}_j)\mathbf{a}_i$ is in the simplex with vertices $\mathbf{a}_1, \dots, \mathbf{a}_k$. Moreover, for each i, there exists a pure sample in S of type i. Therefore, $\text{CH}\{\sum_{i=1}^k (\mathbf{v}_i \cdot \mathbf{x}_j)\mathbf{a}_i \mid j \in [m]\}$ is the simplex on linearly independent vertices $\mathbf{a}_1, \dots, \mathbf{a}_k$. As a result, $\mathbf{a}_1, \dots, \mathbf{a}_k$ are the extreme points of it.

It remains to prove that with probability $1-\delta$, the sample set has a document of purely type j, for all $j \in [k]$. By the assumption on the probability distribution $\mathcal{P}$, with probability at most $(1-\xi)^m$, there is no document of type purely j. Using the union bound, we get the final result.

B Technical Spectral Lemmas

Proposition B.1 (sin θ theorem [13])**.** *. Let $B, \hat{B} \in \mathbb{R}^{p\times p}$ be symmetric, with eigen values $\lambda_1 \geq \cdots \geq \lambda_p$ and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$, respectively. Fix $1 \leq r \leq s \leq p$ and let $V = (\mathbf{v}_r, \dots, \mathbf{v}_s)$ and $\hat{V} = (\hat{\mathbf{v}}_r, \dots, \hat{\mathbf{v}}_s)$ be the orthonormal eigenvectors corresponding to $\lambda_r, \dots, \lambda_s$ and $\hat{\lambda}_r, \dots, \hat{\lambda}_s$. Let $\delta = \inf\{|\hat{\lambda} - \lambda| : \lambda \in [\lambda_s, \lambda_r], \hat{\lambda} \in (-\infty, \hat{\lambda}_{s-1}] \cup [\hat{\lambda}_{r+1}, \infty)\} > 0$. Then ,*

$$\|\sin\Theta(V, \hat{V})\|_2 \leq \frac{\|\hat{B} - B\|_2}{\delta}.$$

where $\sin\Theta(V, \hat{V}) = P_V - P_{\hat{V}}$, where P_V and $P_{\hat{V}}$ are the projection matrices for V and $\hat{V}$.

Proposition B.2 (Corollary 5.50 of [23])**.** *Consider a Gaussian distribution in $\mathbb{R}^n$ with co-variance matrix Σ. Let $A \in \mathbb{R}^{n\times m}$ be a matrix whose rows are drawn i.i.d from this distribution, and let $\Sigma_m = \frac{1}{m}AA^\top$. For every $\epsilon \in (0,1)$, and t, if $m \geq cn(t/\epsilon)^2$ for some constant c, then with probability at least $1 - 2\exp(-t^2 n)$, $\|\Sigma_m - \Sigma\|_2 \leq \epsilon\|\Sigma\|_2$*

Proposition B.3 (Matrix Bernstein [22])**.** *Let $S_1, \dots, S_n$ be independent, centered random matrices with common dimension $d_1 \times d_2$, and assume that each one is uniformly bounded. That is, $\mathbb{E}S_i = 0$ and $\|S_i\|_2 \leq L$ for all $i \in [n]$. Let $Z = \sum_{i=1}^n S_i$, and let $v(Z)$ denote the matrix variance:*

$$v(Z) = \max\left\{\left\|\sum_{i=1}^n \mathbb{E}[S_i S_i^\top]\right\|, \left\|\sum_{i=1}^n \mathbb{E}[S_i^\top S_i]\right\|\right\}.$$

Then,

$$\mathcal{P}[\|Z\| \geq t] \leq (d_1 + d_2)\exp\left(\frac{-t^2/2}{v(Z) + Lt/3}\right).$$

Proposition B.4 (Theorem 4.10 of [20])**.** *Let $\hat{A} = A + E$ and let $\lambda_1, \dots, \lambda_n$ and $\lambda'_1, \dots, \lambda'_n$ be the eigen values of A and $A + E$. Then, $\max\{|\lambda'_i - \lambda_i|\} \leq \|E\|_2$.*

Proposition B.5 (Theorem 3.3 of [19])**.** *For any A and $B = A + E$,*

$$\|B^+ - A^+\| \leq \max 3\left\{\|A^+\|^2, \|B^+\|^2\right\}\|E\|,$$

where $\|\cdot\|$ is an arbitrary norm.

C Omitted Proof from Section 4.1 — Phase 1

C.1 Proof of Claim 2

Let $\mathbf{e}_i$ and $\mathbf{d}_i$ be the i^{th} row of E and D. Then $ED^\top = \sum_{i=1}^m \mathbf{e}_i\mathbf{d}_i^\top$ and $DE^\top = \sum_{i=1}^m \mathbf{d}_i\mathbf{e}_i^\top$. Let $S_i = \frac{1}{m}\begin{bmatrix} 0 & \mathbf{e}_i\mathbf{d}_i^\top \\ \mathbf{d}_i\mathbf{e}_i^\top & 0 \end{bmatrix}$. Then, $\|\frac{1}{m}DE^\top + \frac{1}{m}ED^\top\|_2 \leq 2\|\sum_{i=1}^m S_i\|_2$. We will use matrix Bernstein to show that $\sum_{i\in[m]} S_i$ is small with high probability.

First note that the distribution of $\mathbf{e}_i$ is a Gaussian centered at 0, therefore, $\mathbb{E}[S_i] = 0$. Furthermore, for each i, with probability $1-\delta$, $\|\mathbf{e}_i\|_2 \leq \sigma\sqrt{n}\log\frac{1}{\delta}$. So, with probability $1-\delta$, for all samples $i \in [m]$, $\|\mathbf{e}_i\|_2 \leq \sigma\sqrt{n}\log\frac{m}{\delta}$. Moreover, by assumption $\|\mathbf{d}_i\| = \|\mathbf{x}_i^1 - \mathbf{x}_i^2\| \leq 2M$. Therefore, with probability $1-\delta$,

$$L = \max_i \|S_i\|_2 = \frac{1}{m}\max_i \|\mathbf{e}_i\|\|\mathbf{d}_i\| \leq \frac{2}{m}\sigma\sqrt{n}M\,\text{polylog}\frac{n}{\epsilon\delta}.$$

Note that, $\left\|\mathbb{E}[S_i S_i^\top]\right\| = \frac{1}{m^2}\left\|\mathbb{E}[(\mathbf{e}_i \mathbf{d}_i^\top)^2]\right\| \leq L^2$. Since S_i is Hermitian, the matrix covariance defined by Matrix Bernstein inequality is

$$v(Z) = \max\left\{\left\|\sum_{i=1}^m \mathbb{E}[S_i S_i^\top]\right\|, \left\|\sum_{i=1}^m \mathbb{E}[S_i^\top S_i]\right\|\right\}$$
$$= \left\|\sum_{i=1}^m \mathbb{E}[S_i S_i^\top]\right\| \leq mL^2.$$

If $\epsilon \leq v(Z)/L$ and $m \in \Omega(\frac{n\sigma^2 M^2}{\epsilon^2} \text{polylog} \frac{n}{\epsilon\delta})$ or $\epsilon \geq v(Z)/L$ and $m \in \Omega(\frac{\sqrt{n}\sigma M}{\epsilon} \text{polylog} \frac{n}{\epsilon\delta})$, using Matrix Bernstein inequality (Proposition B.3), we have

$$\Pr\left[\left\|\frac{1}{m} DE^\top + \frac{1}{m} ED^\top\right\| \geq \epsilon\right] = \Pr\left[\left\|\sum_{i=1}^m S_i\right\| \geq \frac{\epsilon}{2}\right] \leq \delta.$$

C.2 Proof of Claim 3

Let $\mathbf{d}_i$ be the i^{th} row D. Then $DD^\top = \sum_{i=1}^m \mathbf{d}_i \mathbf{d}_i^\top$. Let $S_i = \frac{1}{m}\mathbf{d}_i \mathbf{d}_i^\top - \frac{1}{m}\mathbb{E}[\mathbf{d}_i \mathbf{d}_i^\top]$. Then, $\|\frac{1}{m} DD^\top - \mathbb{E}\left[\frac{1}{m} DD^\top\right]\|_2 = \|\sum_{i=1}^m S_i\|_2$. Since, $\mathbf{d}_i = \mathbf{x}_i^1 - \mathbf{x}_i^2$ and $\|\mathbf{x}_i^j\| \leq M$, we have that for any i, $\|\mathbf{d}_i \mathbf{d}_i^\top - \mathbb{E}[\mathbf{d}_i \mathbf{d}_i^\top]\| \leq 4M^2$. Then,

$$L = \max_i \|S_i\|_2 = \frac{1}{m} \max_i \|\mathbf{d}_i \mathbf{d}_i^\top - \mathbb{E}[\mathbf{d}_i \mathbf{d}_i^\top]\|_2 \leq \frac{4}{m} M^2,$$

and $\|\mathbb{E}[S_i S_i^\top] \leq L^2$. Note that S_i is Hermitian, so, the matrix covariance is

$$v(Z) = \max\left\{\left\|\sum_{i=1}^m \mathbb{E}[S_i S_i^\top]\right\|, \left\|\sum_{i=1}^m \mathbb{E}[S_i^\top S_i]\right\|\right\}$$
$$= \left\|\sum_{i=1}^m \mathbb{E}[S_i S_i^\top]\right\| \leq mL^2.$$

If $\delta_0 \leq 4M^2$ and $m \in \Omega(\frac{M^4}{\delta_0^2} \log \frac{n}{\delta})$ or $\delta_0 \geq 4M^2$ and $m \in \Omega(\frac{M^2}{\delta_0} \log \frac{n}{\delta})$, then by Matrix Bernstein inequality (Proposition B.3), we have

$$\Pr\left[\left\|\sum_{i=1}^m S_i\right\| \geq \frac{\delta_0}{2}\right] \leq \delta.$$

D Omitted Proof from Section 4.2 — Denoising

D.1 Proof of Claim 4

Recall that for any $i \in [k]$, with probability $\gamma = g(\epsilon'/(8k\alpha))$ a nearly pure weight vector $\mathbf{w}$ is generated from $\mathcal{P}$, such that $\|\mathbf{w} - \mathbf{e}_i\| \leq \epsilon'/(8k\alpha)$. And independently, with probability p_0 the point is not noisy. Therefore, there is $p_0\gamma$ density on non-noisy points that are almost purely of class i. Note that for such points, $\mathbf{x}$,

$$\|P\mathbf{x} - \mathbf{a}_i\| = \left\|\sum_{j=1}^k w_j \mathbf{a}_j - \mathbf{a}_i\right\| \leq k(\epsilon'/(8k\alpha))(\alpha) \leq \frac{\epsilon'}{8}.$$

Since $\|P - \hat{P}\| \leq \epsilon'/8M$, we have

$$\|\mathbf{a}_i - \hat{P}\mathbf{x}\| = \|\mathbf{a}_i - P\mathbf{x}\| + \|P\mathbf{x} - \hat{P}\mathbf{x}\| \leq \frac{\epsilon'}{8} + \frac{\epsilon'}{8} \leq \frac{\epsilon'}{4}$$

The claim follows immediately.

E Omitted Proof from Section 4.3 — Phase 2

E.1 Proof from Claim 7

Recall that by Lemma 4.3, for any $\hat{\mathbf{x}} \in \hat{S}_{\|}$ there exists $\mathbf{x} \in \Delta$ such that $\|\hat{\mathbf{x}} - \mathbf{x}\| \leq \epsilon'$ and for all $i \in [k]$, there exists $\hat{\mathbf{a}}_i \in \hat{S}_{\|}$ such that $\|\hat{\mathbf{a}}_i - \mathbf{a}_i\| \leq \epsilon'$. For the first part, let $\mathbf{x} = \sum_i \alpha_i \mathbf{a}_i \in \Delta$ be the corresponding point to $\hat{\mathbf{x}}$, where α_i's are the coefficients of the convex combination. Furthermore, let $\mathbf{x}' = \sum_i \alpha_i \hat{\mathbf{a}}_i$. We have,

$$\|\mathbf{x}' - \hat{\mathbf{x}}\| \leq \left\|\sum_{i=1}^k \alpha_i \hat{\mathbf{a}}_i - \sum_{i=1}^k \alpha_i \mathbf{a}_i + \mathbf{x} - \hat{\mathbf{x}}\right\|$$
$$\leq \left\|\max_{i \in [k]} (\hat{\mathbf{a}}_i - \mathbf{a}_i)\right\| + \|\mathbf{x} - \hat{\mathbf{x}}\| \leq 2\epsilon'.$$

The first claim follows from the fact that $\|\hat{\mathbf{x}} - \mathbf{a}_i\| > 8r\epsilon'$ and as a result $\mathbf{x}' \in \text{CH}(\hat{S}_{\|} \setminus B_{6r\epsilon'}(\hat{\mathbf{x}}))$. Next, note that $B_{4r\epsilon'}(\mathbf{a}_i) \subseteq B_{5r\epsilon'}(\hat{\mathbf{a}}_i)$. So, by the fact that $\|\mathbf{a}_i - \hat{\mathbf{a}}_i\| \leq \epsilon'$,

$$\text{dist}\left(\hat{\mathbf{a}}_i, \text{CH}(\Delta \setminus B_{5r\epsilon'}(\hat{\mathbf{a}}_i))\right) \geq \text{dist}\left(\mathbf{a}_i, \text{CH}(\Delta \setminus B_{4r\epsilon'}(\mathbf{a}_i))\right) - \epsilon'$$
$$\geq 3\epsilon'.$$

Next, we argue that if there is $\hat{\mathbf{x}} \in \text{CH}(\hat{S}_{\|} \setminus B_{5r\epsilon'}(\hat{\mathbf{a}}_i))$ then there exists $\mathbf{x} \in \text{CH}(\Delta \setminus B_{4r\epsilon'}(\hat{\mathbf{a}}_i))$, such that $\|\mathbf{x} - \hat{\mathbf{x}}\| \leq \epsilon'$.

To see this, let $\mathbf{x} = \sum_i \alpha_i \hat{\mathbf{z}}_i$ be the convex combination of $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_\ell \in \hat{S}_{\|} \setminus B_{d+\epsilon'}(\hat{\mathbf{a}}_i)$. By Claim 4.3, there are $\mathbf{z}_1, \ldots, \mathbf{z}_\ell \in \Delta$, such that $\|\mathbf{z}_i - \hat{\mathbf{z}}_i\| \leq \epsilon'$ for all $i \in [k]$. Furthermore, by the proximity of $\mathbf{z}_i$ to $\hat{\mathbf{z}}_i$ we have that $\mathbf{z}_i \notin B_d(\hat{\mathbf{a}}_i)$. Therefore, $\mathbf{z}_1, \ldots, \mathbf{z}_\ell \in \Delta \setminus B_d(\hat{\mathbf{a}}_i)$. Then, $\mathbf{x} = \sum_i \alpha_i \mathbf{z}_i$ is also within distance ϵ'.

Using this claim, we have $\text{dist}\left(\hat{\mathbf{a}}_i, \text{CH}(\hat{S}_{\|} \setminus B_{6r\epsilon'}(\hat{\mathbf{a}}_i))\right) \geq 2\epsilon'$.

F Additional Results, Extensions, and Open Problems

F.1 Sample Complexity Lower bound

As we observed the number of samples required by our method is $poly(n)$. However, as the number of classes can be much smaller than the number of features, one might hope to recover $\mathbf{v}_1, \ldots, \mathbf{v}_k$, 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 $\mathbf{v}_1, \ldots, \mathbf{v}_k$ regardless of the value of k.

For ease of exposition, let $k = 1$ and note that in this case every sample should be purely of one type. Assume that the class vector, $\mathbf{v}$, is promised to be in the set $C = \{\mathbf{v}^j \mid v_\ell^j = 1/\sqrt{2}, \text{ if } \ell = 2j-1 \text{ or } 2j, \text{ else } v_\ell^j = 0\}$. Consider instances $(\mathbf{x}_j^1, \mathbf{x}_j^2)$ such that the ℓ^{th} coordinate of $\mathbf{x}_j^1$ is $x_{j\ell}^1 = -1/\sqrt{2}$ if $\ell = 2j - 1$ and $1/\sqrt{2}$ otherwise, and $x_{j\ell}^2 = -1/\sqrt{2}$ if

$\ell = 2j$ and $1/\sqrt{2}$ otherwise. For a given $(\mathbf{x}_j^1, \mathbf{x}_j^2)$, we have that $\mathbf{v}^j \cdot \mathbf{x}_j^1 = \mathbf{v}^j \cdot \mathbf{x}_j^2 = 0$. On the other hand, for all $\ell \neq j$, $\mathbf{v}^\ell \cdot \mathbf{x}_j^1 = \mathbf{v}^\ell \cdot \mathbf{x}_j^2 = 1$. Therefore, sample $(\mathbf{x}_j^1, \mathbf{x}_j^2)$ is consistent with $\mathbf{v} = \mathbf{v}^\ell$ for any $\ell \neq j$, but not with $\mathbf{v} = \mathbf{v}^j$. That is, each instance $(\mathbf{x}_j^1, \mathbf{x}_j^2)$ renders only one candidate of C invalid. Even after observing at most $\frac{n}{2} - 2$ samples of this types, at least 2 possible choices for $\mathbf{v}$ remain. So, $\Omega(n)$ samples are indeed needed to find the appropriate $\mathbf{v}$. The next theorem, whose proof appears in Appendix G generalizes this construction and result to the case of any k.

Theorem F.1. *For any $k \leq n$, any algorithm that for all $i \in [k]$ learns $\mathbf{v}_i'$ such that $\|\mathbf{v}_i - \mathbf{v}_i'\|_2 \leq 1/\sqrt{2}$, requires $\Omega(n)$ samples.*

Note that in the above construction samples have large components in the irrelevant features. It would be interesting to see if this lower bound can be circumvented using additional natural assumptions in this model, such as assuming that the samples have length $\mathrm{poly}(k)$.

F.2 Alternative Noise Models

Consider the problem of recovering $\mathbf{v}_1, \ldots, \mathbf{v}_k$ 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. Furthermore, assume that the distribution over the instance space is rich enough such that any subspace other than $\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ is inconsistent with a set of instances of non-negligible density.[5] Since the VC dimension of the set of k dimensional subspaces in $\mathbb{R}^n$ is $\min\{k, n-k\}$, from the information theoretic point of view, one can recover $\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ as it is the only subspace that is inconsistent with less than $O(\epsilon)$ fraction of $\tilde{O}(\frac{k}{\epsilon^2})$ samples. Furthermore, we can detect and remove any noisy sample, for which the two views of the sample are not consistent with $\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. And finally, we can recover $\mathbf{a}_1, \ldots, \mathbf{a}_k$ using phase 2 of Algorithm 1.

In the above discussion, it is clear that once we have recovered $\mathrm{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, denoising and finding the extreme points of the projections can be done in polynomial time. For the problem of recovering a k-dimensional nullspace, [14] introduced an efficient algorithm that tolerates agnostic noise up to $\epsilon = O(k/n)$. Furthermore, they provide an evidence that this result might be tight. It would be interesting to see whether additional structure present in our model, such as the fact that samples are convex combination of classes, can allow us to efficiently recover the nullspace in presence of more noise.

Another interesting open problem is whether it is possible to handle the case of $p_0 = 0$. That is, when *every document* is affected by Gaussian noise $\mathcal{N}(0, \sigma^2 I_n)$, for $\sigma \gg \epsilon$. A simpler form of this problem is as follows. Consider a distribution induced by first drawing $\mathbf{x} \sim D$, where D is an arbitrary and unknown distribution over $\Delta = \mathrm{CH}(\{\mathbf{a}_1, \ldots, \mathbf{a}_k\})$, and taking $\hat{\mathbf{x}} = \mathbf{x} + \mathcal{N}(0, \sigma^2 I_n)$. *Can we learn $\mathbf{a}_i$'s within error of ϵ using polynomially many samples?* Note that when D is only supported on the corners of Δ, this problem reduces to learning mixture of Gaussians, for which there is a wealth of literature on estimating Gaussian means and mixture weights [12, 16, 17]. It would be interesting to see under what regimes $\mathbf{a}_i$ (and not necessarily the mixture weights) can be learned when D is an arbitrary distribution over Δ.

F.3 General function $f(\cdot)$

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 $f : \mathbb{R}^+ \to [0, 1]$ such that $f(0) = 0$. We describe how variations of the techniques discussed up to now can extend to this more general setting.

For ease of exposition, consider the non-noisy case. Since f is a strictly increasing function, $f(\mathbf{v}_i \cdot \mathbf{x}^1) = f(\mathbf{v}_i \cdot \mathbf{x}^2)$ if and only if $\mathbf{v}_i \cdot \mathbf{x}^1 = \mathbf{v}_i \cdot \mathbf{x}^2$. Therefore, we can recover $\mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ by the same approach as in Phase 1 of Algorithm 1. Although, by definition of pseudoinverse matrices, the projection of $\mathbf{x}$ is still represented by $\mathbf{x}_\parallel = \sum_i (\mathbf{v}_i \cdot \mathbf{x}) \mathbf{a}_i$, this is not necessarily a convex combination of $\mathbf{a}_i$'s anymore. This is due to the fact that $\mathbf{v}_i \cdot \mathbf{x}$ can add up to values larger than 1 depending on $\mathbf{x}$. However, $\mathbf{x}_\parallel$ is still a *non-negative combination* of $\mathbf{a}_i$'s. Moreover, $\mathbf{a}_i$'s are linearly independent, so $\mathbf{a}_i$ cannot be expressed by a nontrivial non-negative combination of other samples. Therefore, for all i, $\mathbf{a}_i/\|\mathbf{a}_i\|$ can be recovered by taking *the extreme rays of the convex cone* of the projected samples. So, we can recover $\mathbf{v}_1, \ldots, \mathbf{v}_k$, by taking the psuedoinverse of $\mathbf{a}_i/\|\mathbf{a}_i\|$ and re-normalizing the outcome such that $\|\mathbf{v}_i\|_2 = 1$. When samples are perturbed by noise, a similar argument that also takes into account the smoothness of f proves similar results.

It would be interesting to see whether a more general class of similarity functions, such as kernels, can be also learned in this context.

G Proof of Theorem F.1 — Lower Bound

For ease of exposition assume that n is a multiple of k. Furthermore, in this proof we adopt the notion $(\mathbf{x}_i, \mathbf{x}_i')$ to represent the two views of the i^{th} sample. For any vector $\mathbf{u} \in \mathbb{R}^n$ and $i \in [k]$, we use $(\mathbf{u})_i$ to denote the i^{th} $\frac{n}{k}$-dimensional block of $\mathbf{u}$, i.e., coordinates $u_{(i-1)\frac{n}{k}+1}, \ldots, u_{i\frac{n}{k}}$.

Consider the $\frac{n}{k}$-dimensional vector $\mathbf{u}_j$, such that $u_{j\ell} = 1$ if $\ell = 2j - 1$ or $2j$, and $u_{j\ell} = 0$, otherwise. And consider $\frac{n}{k}$-dimensional vectors $\mathbf{z}_j$ and $\mathbf{z}_j'$, such that $z_{j\ell} = -1$ if $\ell = 2j - 1$ and $z_{j\ell} = 1$ otherwise, and $z_{j\ell}' = -1$ if $\ell = 2j$ and $z_{j\ell}' = 1$ otherwise. Consider a setting where $\mathbf{v}_i$ is restricted to the set of candidate $C_i = \{\mathbf{v}_i^j \mid (\mathbf{v}_i^j)_i = \mathbf{u}_j/\sqrt{2}$ and $(\mathbf{v}_i^j)_{i'} = \mathbf{0}$ for $i' \neq i\}$. In other words, the ℓ^{th} coordinate of $\mathbf{v}_i^j$ is $1/\sqrt{2}$ if $\ell = (i-1)\frac{n}{k} + 2j - 1$ or $(i-1)\frac{n}{k} + 2j$, else 0. Furthermore, consider instances $(\mathbf{x}_i^j, \mathbf{x}_i'^j)$ such that $(\mathbf{x}_i^j)_i = \mathbf{z}_j/\sqrt{2}$ and $(\mathbf{x}_i'^j)_i = \mathbf{z}_j'/\sqrt{2}$ and for all

[5] This assumption is similar to the richness assumption made in the standard case, where we assume that there is enough "entropy" between the two views of the samples such that even in the non-noisy case the subspace can be uniquely determined by taking the nullspace of $X_1 - X_2$.

$i' \neq i$, $(\mathbf{x}_i^j)_{i'} = (\mathbf{x}_i'^j)_{i'} = \mathbf{0}$. In other words,

$$\mathbf{x}_i^j = \frac{1}{\sqrt{2}}\,(0,\ldots,0,\;\; 1,\ldots,1,\;\; \overbrace{1,-1}^{(i-1)\frac{n}{k}+2j-1,(i-1)\frac{n}{k}+2j}\;,1,\ldots,1,\;\; 0,\ldots,0),$$
$$\mathbf{x}_i'^j = \frac{1}{\sqrt{2}}\,(0,\ldots,0,\;\; 1,\ldots,1,-1,\;\; 1\;,1,\ldots,1,\;\; 0,\ldots,0),$$
$$\mathbf{v}_i^j = \frac{1}{\sqrt{2}}\,(0,\ldots,0,\;\; \underbrace{0,\ldots,0,\;\; 1,\;\; 1\;,0,\ldots,0,}_{i^{th}\;block}\;\; 0,\ldots,0).$$

First note that, for any $i, i' \in [k]$ and any $j, j' \in [\frac{n}{2k}]$, $\mathbf{v}_i^j \cdot \mathbf{x}_{i'}^{j'} = \mathbf{v}_i^j \cdot \mathbf{x}_{i'}'^{j'}$. That is, the two views of all instances are consistent with each other with respect to all candidate vectors. Furthermore, for any i and i' such that $i \neq i'$, for all j, j', $\mathbf{v}_i^j \cdot \mathbf{x}_{i'}^{j'} = 0$. Therefore, for any observed sample $(\mathbf{x}_i^j, \mathbf{x}_i'^j)$, the sample should be purely of type i.

For a given i, consider all the samples $(\mathbf{x}_i^j, \mathbf{x}_i'^j)$ that are observed by the algorithm. Note that $\mathbf{v}_i^j \cdot \mathbf{x}_i^j = \mathbf{v}_i^j \cdot \mathbf{x}_i'^j = 0$. And for all $j' \neq j$, $\mathbf{v}_i^{j'} \cdot \mathbf{x}_i^j = \mathbf{v}_i^{j'} \cdot \mathbf{x}_i'^j = 1$. Therefore, observing $(\mathbf{x}_i^j, \mathbf{x}_i'^j)$ only rules out $\mathbf{v}_i^j$ as a candidate, while this sample is consistent with candidates $\mathbf{v}_i^{j'}$ for $j' \neq j$. Therefore, even after observing $\leq \frac{n}{2k} - 2$ samples of this types, at least 2 possible choices for $\mathbf{v}_i$ remain valid. Moreover, the distance between any two $\mathbf{v}_i^j, \mathbf{v}_i^{j'} \in C_i$ is $\sqrt{2}$. Therefore, $\frac{n}{2k} - 1$ samples are needed to learn $\mathbf{v}_i$ to an accuracy better than $\sqrt{2}/2$.

Note that consistency of the data with $\mathbf{v}_{i'}$ is not affected by the samples of type $\mathbf{x}_i^j$ that are observed by the algorithms when $i' \neq i$. So, $\Omega(k\frac{n}{k}) = \Omega(n)$ samples are required to approximate all $\mathbf{v}_i$'s to an accuracy better than $\sqrt{2}/2$.