Near-optimal sample complexity bounds for learning Latent $k$-polytopes and applications to Ad-Mixtures

Chiranjib Bhattacharyya $^*$ $^1$ Ravindran Kannan $^*$ $^2$

Abstract

Deriving Optimal bounds on Sample Complexity of Latent Variable models is an active area of research. Recently such bounds were obtained for Mixture of Gaussians (Ashtiani et al., 2018), no such results are known for Ad-mixtures, a generalization of Mixture distributions. In this paper we show that $O^*(dk/m)$ samples are sufficient to learn each of $k$-topic vectors of LDA, a popular Ad-mixture model, with vocabulary size $d$ and $m \in \Omega(1)$ words per document, to any constant error in $L_1$ norm. The result is a corollary of the major contribution of this paper: the first sample complexity upper bound for the problem (introduced in (Bhattacharyya & Kannan, 2020)) of learning the vertices of a Latent $k$-Polytope in $\mathbb{R}^d$, given perturbed points from it. The bound, $O^*(dk/\beta)$, is optimal and linear in number of parameters. It applies to many stochastic models including a broad class Ad-mixtures. To demonstrate the generality of the approach we specialize the setting to Mixed Membership Stochastic Block Models(MMSB) and show for the first time that if an MMSB has $k$ blocks, the sample complexity is $O^*(k^2)$ under usual assumptions.

1. Introduction

Recently, in a seminal paper, building on long line of research, (Ashtiani et al., 2018) showed that a mixture of $k$ gaussians in $\mathbb{R}^d$ can be recovered using $O$(number of parameters) iid samples drawn from the mixture.

Our aim in this paper is to prove similar results for Ad-mixtures, a problem not tackled before. Ad-mixtures are generalizations of mixture distributions and have been studied extensively in various applications including Text Modelling, Population genetics, Community detection and many other fields (for a survey see (Airoldi et al., 2014)). Whereas in a mixture distribution, there is a fixed convex combination of component distributions from which all samples are generated, in Ad-mixtures, each sample is drawn from a different randomly picked convex combination of component distributions. This results in substantial overlap of distributions from which samples are picked and makes the learning problem harder.

Ad-mixtures themselves are special cases of a Geometric model called the Latent $k$-polytope (LKP) recently introduced in (Bhattacharyya & Kannan, 2020). In LKP, there is a latent $k$-polytope $K \subset \mathbb{R}^d$ from which $n$ iid observations in $\mathbb{R}^d$ are drawn, each obtained by stochastically perturbing an unobserved(or latent ) point in $K$. The latent points are chosen from a known prior distribution and each data point is chosen according to some known distribution. The LKP problem consists of approximating the vertices from iid observations, and (Bhattacharyya & Kannan, 2020) developed a polynomial time approximation algorithm for the LKP problem under certain assumptions.

The sample complexity question for even special cases, let alone the general LKP, has not been addressed. The central problem addressed here is to find the sample complexity of LKP problem where, we wish to learn each vertex within $L_1$ error $\varepsilon$Diameter$(K)$ and establish a near-optimal bound; we will show that $O^*(number of parameters)$ suffice under realistic assumptions.

Contributions Our results are summarized as follows.

Near-Optimal Sample Complexity for LKP: We prove an upper bound of $O^*(dk/\beta)$ on the sample complexity of LKP for $\varepsilon \in \Omega(1)$ where, $\beta \in \Omega(1)$ is a model-specific constant (Corollary (4.2)). The bound is proved by showing that the vertices of LKP can be well approximated by $k$ points residing in a polytope completely characterized by observed data, more specifically it is described by the convex hull of sample averages of certain sized subsets of observations.
Vertex Set Certificate from a candidate set: Central to our results is a novel method of independent interest which provides a certificate of when a set of $k$ points approximates the set of vertices of a latent polytope. See Theorem 4.6 for a precise statement.

Sufficient Conditions for learnability of certain models: As a consequence of the main result (Corollary (4.2)), any latent variable model which can be posed as LKP and satisfies general conditions on the prior and the probability distribution of data generation in the main result is learnable. This opens the doors for deriving sample complexity estimates for widely used models including admixtures.

Near-Optimal Upper bound on Sample Complexity for LDA: A widely used instance of a $k$-admixture is the LDA model(Blei et al., 2003). Sample complexity of LDA is an open problem. We provide an upper-bound (Theorem 5.1) of $O^*(dk/m)$ on the number of samples for a $k$ topic LDA model, with vocabulary size $d$ and with $m$ words per document, thus proving number of tokens in $\Omega^*(\text{Number of parameters})$ suffices. This is a consequence of the upper bound on the sample complexity of LKP and the fact that $\beta$ is $m$ (Lemma 5.4).

Near-Optimal Lower bounds on LDA: Using a combinatorial code-design, we show a matching lower bound of $\Omega^*(dk/m)$ for the sample complexity of LDA (Lemma 6.1). Since LDA is special case of Ad-Mixtures and Admixtures are a special case of LKP, the lower bound applies to them too (Lemma 6.2). To the best of our knowledge, this is the first near optimal sample complexity estimate for LDA.

Near-Optimal Upper Bound for MMSB: For a $k$-block mixed membership stochastic block model, another instance of an Ad-mixture, we argue that $n \in O^*(k^2)$ entries suffice to learn the $k \times k$ connection-probability matrix under mild conditions.

2. Related work

The problem of learning a distribution from iid samples have a rich history, for general background we refer the reader to (Devroye & Lugosi, 2001; Silverman, 1986). Sample Complexity was first studied in computational learning theory community in (Kearns et al., 1994). Since then there has been significant interest in the community to derive sample complexity estimates for different distributions. In particular there has been significant efforts in learning mixture of Gaussian distributions, see (Kalai et al., 2012) for a survey. The optimal sample complexity, optimal in the sense of Information theoretic limit of Gaussian distribution was only derived recently in (Atrashani et al., 2018). They showed that a mixture of $k$ Gaussians in $d$ dimensions can be approximated to an additive $\epsilon$ from $O\left(\frac{k^2 \log(d)}{\epsilon^2}\right)$ i.i.d samples. It is optimal in the sense that $kd^2$ is also the number of parameters.

In this paper we wish to derive optimal sample complexity estimates for Ad-mixture models. Topic models, in particular LDA(Blei et al., 2003), is an instance of Ad-mixture which has proven to very useful in practice(Blei, 2012). Many widely used algorithms for parameter estimation (e.g. (Griffiths & Steyvers, 2004)) do not provide sample complexity estimates. Recently there has been significant efforts in developing algorithms which have finite sample complexity((Arora et al., 2013; 2012; Anandkumar et al., 2012; Bansal et al., 2014). However the sample complexity estimates involve high polynomials in the parameters. Mixed membership models, including Topic models, are broad class of Ad-mixtures which have seen wide-spread applicability(see (Airoldi et al., 2014) for a survey). These models have been also extended to model overlapping communities (see e.g. (Airoldi et al., 2008)). The main technical challenge in Ad-mixtures which distinguishes it from Mixtures is the samples are drawn from distributions with high overlap. This renders known techniques not applicable in this context.

As argued before, studying the sample complexity estimate of LKP problem(Bhattacharyya & Kannan, 2020) would have immediate applications to Ad-mixtures. The geometric nature of the problem suggests connections to Convex set estimation problems where the aim is to discover the convex set from uniformly drawn, possibly noisy,1 samples, from it (see (Brunel, 2018) for a survey). In these problems the aim is to construct an approximation to the desired convex set in terms of the volume or distance of the convex set. In the LKP problem the requirement is more stringent; the extreme points of a polytope needs to be discovered. Direct application of the results (Brunel, 2018) to LKP does not yield anything interesting as in applications such as Topic models, it is neither realistic nor sufficient to have uniformly drawn samples.

3. Preliminaries

3.1. Latent $k$-polytope (LKP) problem

We begin by recalling the definition of LKP and the data generation process.

Definition 3.1 (Latent $k$-polytope) LKP is a polytope $K$ with $k$ extreme points described by the columns of a $d \times k$ matrix $M$. Let $\mathcal{D}_k$ be a probability density over the $k - 1$ dimensional simplex, $\Delta_{k-1} = \{w \in \mathbb{R}^k : w_t \geq 0, \sum_{t=1}^k w_t = 1\}$. The observed data drawn from LKP is described by a $d \times n$ matrix $A$ where each column $A_{:,j}$ is independently drawn
Near-optimal sample complexity via Latent \( k \)-Polytopes

as follows.

\[
W_{.,j} \sim \mathcal{D}_k, \quad P_{.,j} = MW_{j}, \quad A_{.,j} | P_{.,j} \sim \mathcal{P}_d, \quad (\text{LKP})
\]

where \( \mathcal{P}_d \) is a probability density or probability mass function with the property that \( \mathbb{E}_{\mathcal{P}_d} [p_j(A_{.,j})] = P_{.,j} \).

The definition implies in matrix notation,

\[
P = MW, \quad \mathbb{E}_{\mathcal{P}_d}[p(A)] = P.
\]

**Definition 3.2** (The LKP problem) The LKP problem is to estimate \( M \) from \( A \).

**Ad-Mixtures** Ad-Mixtures are special cases of LKP, where the probability distribution (density or pmf) \( F(A_{.,j} | P_{.,j}) \) is given by

\[
F(A_{.,j} | P_{.,j}) = \sum_{\ell=1}^{k} W_{\varepsilon,j} F(A_{.,j} | \mathcal{M}_{.,\ell}).
\]

In LDA (Blei et al., 2003), a popular instance of Ad-mixture, the columns of \( W \) are picked according to Dirichlet distribution. LDA is one of our main examples; in analyzing LDA, we will use the symmetric Dirichlet probability distribution \( \mathcal{D}_k = \text{Dir}(k, \alpha) \) whose density on the unit simplex, \( \Delta_{k-1} \) with concentration parameter \( \alpha \in (0,1) \) has p.d.f.:

\[
\frac{1}{\Gamma(\alpha)} \prod_{\ell=1}^{k} x_{\ell}^{\alpha-1}, \quad \text{where, } g(\alpha) \text{ is the normalizing constant.}
\]

The distribution of \( A_{.,j} | P_{.,j} \) is a multinomial.

**3.2. Sample Complexity**

We will use \( \| \cdot \| \) to denote the \( \ell_1 \) norm. \( \varepsilon \) will determine the error we allow in parameter estimation. The problem is to find approximations to the columns of \( M \) to within \( \| \cdot \| \) distance \( \varepsilon \nu \), where

\[
\nu = \text{Diameter}_{\ell_1}(K).
\]

\( \mathcal{M}(d, k) \) will denote a subset of \( d \times k \) matrices; in LDA, it is a subset of matrices with non-negative entries with column sums equal to 1 and in MMBM, it is a subset of matrices with each entry in \([0, 1]\).

We define a learner and its sample complexity now. A learner is an Algorithm which is given data \( A \) generated from an LKP model parametrized by an unknown \( M \) and returns \( k \) vectors which are approximations to the columns of \( M \), (which are also the extreme points of the polytope \( K \).) Note that the learner can only learn the columns up to a permutation. To allow this, we use the Hausdorff distance defined below.

**Definition 3.3** (Learner) A learner for \( \mathcal{M}(d, k) \) is a finite time deterministic algorithm which given data \( A \) drawn from LKP parametrized by an unknown \( M \in \mathcal{M}(d, k) \), outputs \( k \) vectors \( v_1, \ldots, v_k \in \mathbb{R}^d \).

As defined, the learner can output any set of \( k \) vectors. The learner is successful if the set \( \{v_1, v_2, \ldots, v_k\} \) is close to the set of columns of \( M \). Closeness will be measured by Hausdorff distance: The Hausdorff distance \( D(A, B) \) between two sets \( A, B \) of points in \( \mathbb{R}^d \) is defined by:

\[
D(A, B) = \max \{\text{Dist}(A, B), \text{Dist}(B, A)\}, \quad \text{where, for sets } X, Y, \text{ Dist}(X, Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|
\]

Next we define the notion of Sample complexity.

**Definition 3.4** (Sample Complexity) The \( \varepsilon \)-Sample Complexity of a learner for \( \mathcal{M}(d, k) \) is the minimum number \( f(\varepsilon, d, k) \), such that for every \( M \in \mathcal{M}(d, k) \) and \( n \geq f(\varepsilon, d, k) \) the \( \{v_1, v_2, \ldots, v_k\} \) returned by the learner when presented with \( \varepsilon \) and \( n \) iid draws from a LKP parametrized by \( M \) satisfies (2) with probability more than 0.99.

\[
D(\{v_1, v_2, \ldots, v_k\}, \{M_{.,1}, M_{.,2}, \ldots, M_{.,k}\}) \leq \varepsilon \nu. \quad (2)
\]

The optimal \( \varepsilon \)-sample complexity of \( \mathcal{M}(d, k) \) is the minimum sample complexity of a learner for \( \mathcal{M}(d, k) \).

For ease of exposition we have fixed the probability to be 0.99. It is easy to adapt it to a general setting where the statement holds with probability \( 1 - \delta \), where \( \delta \) is the failure probability.

In the following our first major result will be a finite time algorithm whose sample complexity is \( O^*(\text{number of parameters}) \), giving an upper bound on the Optimal Sample Complexity. Optimality is shown by exhibiting a carefully chosen LKP for which any successful learner will require at least samples linear in the parameters.

**4. Sample complexity of LKP using Subset Averages**

**Notations:** For any \( B \subset \mathbb{R}^d \), \( C(\gamma)(B) \) will denote the convex hull of \( B \). For any subset \( R \) of data points, described by the columns of \( A \), we let \( \bar{A}_{.,R} \) denote the average of the data points in \( R \).

In this section we derive an upper-bound on the Sample Complexity of LKP.

**4.1. An intuitive description of the learner**

In this subsection we give an intuitive description of the learner. The learner is based on several assumptions to be described later. Under the assumptions, for a \( \gamma \in (0, 1) \), we show that

---

1 A learner is successful if (2) is satisfied. The probability is on the iid draws; the learner itself is deterministic.
Near-optimal sample complexity via Latent $k$--Polytopes

1. For every $R \subseteq [n], |R| = \gamma n$, $A_{\cdot R}$ is close to $K$. A precise statement is presented in Lemma (4.4, (ii)). This observation leads us to define a Subset Smoothed Data Polytope (SSDP) as:

$$SSDP_\gamma(A) = CH(A_{\cdot R}, |R| = \gamma n) \quad (3)$$

2. There are $k$ subsets $S_1, S_2, \ldots, S_k$ of data, each of cardinality $\gamma n$, such that for each $M_{\cdot \ell}$, there is some $S_{\ell'}$ with $A_{\cdot S_{\ell'}}$ close to $M_{\cdot \ell}$ (see Lemma (4.4, (ii))). In other words, it says that there exists $k$ points in $SSDP_\gamma(A)$, each of which is close to one vertex of $K$.

3. From the above two statements, it can be deduced that the convex sets $K$, $SSDP_\gamma(A)$ and $CH(A_{\cdot S_1}, A_{\cdot S_2}, \ldots, A_{\cdot S_k})$ are close to each other. Note that while $K$ is unknown, $SSDP_\gamma(A)$, is completely determined by data, albeit, in exponential time.

4. Further, for any $k$ subsets of data $T_1, T_2, \ldots, T_k$, each of cardinality $\gamma n$, if every $A_{\cdot R}, |R| = \gamma n$ is approximately contained in $CH(A_{\cdot T_1}, A_{\cdot T_2}, \ldots, A_{\cdot T_k})$, then, for each $M_{\cdot \ell}$, there is some $\ell'$ with $A_{\cdot T_{\ell'}}$ close to $M_{\cdot \ell}$ and so $\{A_{\cdot \ell}, \ell = 1, 2, \ldots, k\}$ is a solution to LKP. (See the Vertex Certificate Theorem (4.6)).

Of these facts, the last one is by far the hardest to prove. Now, the procedure is clear: Enumerate all the $n \choose \gamma n$ collections of $k$ subsets $T_1, T_2, \ldots, T_k$ of cardinality $\gamma n$ of data and for each collection, solve a convex program to check if each $A_{\cdot R}, |R| = \gamma n$ is close to the convex hull of $CH(A_{\cdot T_1}, A_{\cdot T_2}, \ldots, A_{\cdot T_k})$. The first time we get a Yes answer (and we are guaranteed to get Yes at some point), we have a solution to LKP.

### 4.2. Assumptions on LKP

We formulate three model assumptions, one each on $M, D_k, P_d$.

The first assumption requires the vertices of $K$ are separated.

**Definition 4.1 (Well Separated Assumption on $M$)**

$M(d, k)$ satisfies the Well Separated Assumption with parameter $\varepsilon$ if for all $M \in M(d, k)$,

$$\forall \ell \in [k], \text{Dist}(M_{\cdot \ell}, CH(M_{\cdot \ell}, \ell' \neq \ell)) > 2\varepsilon n. \quad (4)$$

This assumption is akin to, but, stronger than, assuming that means of mixture components in Mixture of Gaussians are well separated, a common assumption often made in the literature (see (Kalai et al., 2012)).

The next assumption requires that $D_k$ puts sufficient mass close to the vertices of $K$.

**Definition 4.2 (Vertex Proximate Assumption on $D_k$)**

$D_k$ satisfies the Vertex Proximate Assumption with parameters $\gamma \in [0, 1/2k], \varepsilon \in (0, 1)$ if:

$$\forall \ell \in [k], \text{Prob}(W_{\ell, j} > 1 - \varepsilon^2/12) \geq 2\gamma. \quad (5)$$

It stipulates that there is sufficient probability of picking $W_{\ell, j}$ close to 1, which leads to $P_{j, \cdot}$ being near vertex $M_{\cdot \ell}$ of $K$. In Lemma 4.4 we will show that w.h.p it implies that a $\gamma$ fraction of $n$ draws, for $n$ large enough, have latent points close to each vertex. This is in a sense necessary: If the $P_{j, \cdot}$ are all concentrated in a proper subset of $K$, it is information theoretically impossible to find $K$. Later in Lemma 5.3 we will demonstrate that the assumption is obeyed in LDA with the prior $Dir(k, \frac{1}{2})$.

The third assumption is on $P_d$ which governs the distribution of $A_{\cdot j} | P_{\cdot j}$. To accommodate a wide variety of situations it is important that individual perturbations, $u_j = A_{\cdot j} - P_{\cdot j}$, be allowed as much freedom as possible yet allowing for tractable discovery of the vertices w.h.p. We pursue this objective by imposing a sub-gaussian assumption with parameter $\beta$ on the sum of the perturbation vectors over each subset $R$ of $\gamma n$ data points, i.e., on $X_R = \frac{1}{|R|} \sum_{j \in R} u_j$. Though we do not use this fact here, the maximum value of $\beta$, turns out to be the reciprocal of the maximum over all $R$, $|R| = \gamma n$, of the sub-Gaussian norm (with $\|\cdot\|$ as the underlying norm instead of the usual Euclidean norm) of the vector-random variable $X_R$ (see for example ((Vershynin, 2010), Definition 5.22)).

For $R \subset [n], |R| = \gamma n, \text{ let } A_{\cdot R} = \sum_{j \in R} A_{\cdot j} / |R|, \text{ P}_{\cdot R} = \sum_{j \in R} P_{\cdot j} / |R|$. We use $\mu$ to denote the mean of $X_R$ and $\beta$ the subgaussian norm of $X_R$.

**Definition 4.3 (Aggregate Sub-gaussian Perturbation on $P_d$)**

$P_d$ satisfies the Aggregate Sub-gaussian Assumption with parameters $\beta > 0, \gamma \in [0, 1]$ if $(\nu = \text{Dia}(k))$:

$$\forall R \subseteq \{1, 2, \ldots, n\}, |R| = \gamma n, \forall \lambda \geq 0, \forall v \in \{-1, +1\}^d, \text{ Prob}(\nu \cdot (A_{\cdot R} - P_{\cdot R}) \geq \lambda \nu) \leq 6 \exp(-\beta \lambda^2 |R|). \quad (6)$$

In the two applications- LDA, MMBM, we will show (6) holds with $\beta \in \Omega^*(1)$.

**Theorem 4.1 Suppose $\gamma, \beta, \varepsilon$ are such that**

(i) $D_k$ satisfies Vertex Proximate Assumption with parameters $\gamma, \varepsilon$.

(ii) $P_d$ satisfies Aggregate Sub-gaussian Assumption with parameters $\beta, \gamma$.
and (iii) \( M(d, k) \) satisfies Well Separated Assumption with parameter \( \varepsilon \).

\[
\text{If } n \geq \text{Max } \left( \frac{c \ln(ck)}{\gamma}, \frac{cd}{\beta \gamma \varepsilon^4} \right) \quad (7)
\]

\[
\beta \geq c \ln(1/\gamma)/\varepsilon^4 \quad (8)
\]

then there is a learner for \( M(d, k) \) which given \( n \) iid samples drawn according to \( M \in M(d, k) \), solves \((\gamma_n/k)\) convex programs each of \( \text{poly}(n, d) \) size and returns \( v_1, v_2, \ldots, v_k \) which with probability greater than 0.99 satisfy
\[
D(\{v_1, v_2, \ldots, v_k\}, \{M, 1, M, 2, \ldots, M, k\}) \leq \varepsilon \text{Dia}_{L_1}(K).
\]

The theorem applies to all parameter regimes, but we would investigate it in the interesting regime of \( \gamma = \text{poly}(\varepsilon)/k \), which applies to many widely-studied models. For such a choice of \( \gamma \) we obtain the following corollary.

**Corollary 4.2 (LKP Upper Bound)** If \( \gamma = \text{poly}(\varepsilon)/k \), \( d > \beta \ln(10k) \) and \( \beta > \text{poly}(k/\varepsilon)^4 \) and (i), (ii) and (iii) of Theorem (4.1) hold, then the \( \varepsilon \)-sample complexity of LKP is
\[
O^*(\frac{d k}{\beta \text{poly}(\varepsilon)}).
\]

The Aggregate Sub-gaussian Assumption is weaker than assuming a sub-gaussian bound on each individual perturbation. Hence, we can prove the following corollary.

**Corollary 4.3** Suppose \( \gamma, \beta, \varepsilon \) are such that hypotheses (i) and (iii) of Theorem (4.1) hold and in addition, we have: For every \( P_{j, j} \), every \( \lambda > 0 \), and every \( v \in \{-1, +1\}^d \), \( P_{\ell, \ell} \cdot (A_{j, j} - P_{j, j}) > \lambda \varepsilon \) \( \ln(\beta \lambda^2) \),
\[
\text{If } n \geq \text{Max } \left( \frac{c \ln(ck)}{\gamma}, \frac{cd}{\beta \gamma \varepsilon^4} \right) \quad (9)
\]

\[
\beta \geq c \ln(1/\gamma)/\varepsilon^4 \quad (10)
\]

then we can find a \( \varepsilon \)-learner for \( M(d, k) \) by solving \((n/k)\) convex programs each of \( \text{poly}(n, d) \).

The proof is in the Supplementary material. Though sub-gaussian assumption on individual perturbations does not hold for LDA or MMSB, it does apply to specific instances of (Yurochkin et al., 2019)- a detailed investigation will be presented elsewhere.

Some Lemmas are needed for Theorem 4.1.

**Definition 4.4 Candidate set** Let \( \varepsilon > 0 \) and \( Q \) be a \( k \)-polytope in \( \mathbb{R}^d \) with each of its vertices at \( L_1 \) distance at least \( \varepsilon \nu \) (where, \( \nu = \text{Diameter}_{L_1}(Q) \)) from the convex hull of other vertices. A set \( U \) of points in \( \mathbb{R}^d \) is said to be a \( \varepsilon \)-candidate set for \( Q \) if
\[
\text{Dist}(U, Q) < \varepsilon^2 \nu/12 \quad (11)
\]
\[
\text{Dist(set of vertices of } Q, U) < \varepsilon^2 \nu/6 \quad (12)
\]

**Lemma 4.4** (i) Suppose the Vertex Proximate Assumption (5) holds and \( n > c \ln(ck)/\gamma \). Then with probability at least .999, the following event holds.

**Proximate latent points Event**
\[
\forall \ell \in [k], \exists S_\ell \subseteq \{1, 2, \ldots, n\} : |S_\ell| = \gamma n : \forall j \in S_\ell : ||P_{j,j} - M, \ell|| < \varepsilon^2 \nu/12 ; ||P_{,\ell} - M, \ell|| < \varepsilon^2 \nu/12. \quad (13)
\]

(ii) Suppose Aggregate Sub-gaussian Assumption holds, and (8) and (7) are satisfied. Then, with probability at least .999, the following event holds.

**Subset-perturbation Event**
\[
\forall R \subseteq \{1, 2, \ldots, n\} : |R| = \gamma n, \quad ||A, R - P, R|| \leq \varepsilon^2 \nu/12 \quad (14)
\]

**Proof**: (i) uses a quantitative version of the coupon-collector problem. Consider the \( k \) events, one for each \( \ell \), namely
\[
\mathcal{E}_\ell : \exists S_\ell \subseteq [n] : |S_\ell| = \gamma n,
\]
\[
||P_{j,j} - M, \ell|| < \varepsilon^2 \nu/12 \forall j \in S_\ell
\]

The event \( \neg \mathcal{E}_\ell \) is \( \{j : ||M, j - P, j|| < \varepsilon^2 \nu/12\} < \gamma n \). But, we have
\[
||M, j - P, j|| = ||(1 - W_{\ell, j} M, \ell - \sum_{\ell' \neq \ell} W_{\ell', j} M, \ell'||
\]
\[
\leq (1 - W_{\ell, j}) \nu,
\]

since \( ||M, \ell'|| \leq \nu \) for all \( \ell' \) and \( \sum_{j} W_{\ell', j} = 1 \). This implies: \( \{j : ||P_{j,j} - M, \ell|| < \varepsilon^2 \nu/12\} \geq \{j : W_{\ell, j} \geq 1 - \varepsilon^2/12\} \). Now, \( \{j : W_{\ell, j} \geq 1 - \varepsilon^2/12\} \) is the sum of Bernoulli random variables and we have by (5), \( \mathbb{E}[\{j : W_{\ell, j} \geq 1 - \varepsilon^2/12\}] \geq 2 \nu n \). So by Höffding-Chernoff inequality, \( \Pr(\neg \mathcal{E}_\ell) \leq 4 \exp(-\gamma n/4) \leq k^8 \) by \( n > c \ln(ck)/\gamma \). Taking the union bound over the \( \ell \in [k] \), (i) follows.

(ii) From (6) with union bound over all \( R \) and all \( v \in \{-1, +1\}^d \), we get that (14) happens with probability at least \( 1 - c(n) \exp(-c \beta \varepsilon^4 \gamma n + d) \). Now, by Stirling,
\[
\binom{n}{\gamma n} \leq \exp(3 \gamma n \ln(3/\gamma)) \leq \exp(\beta \varepsilon^4 \gamma n),
\]

the last by (8). Also, \( \beta \varepsilon^4 \gamma n > d \) by (7). Thus (14) holds whp.
Lemma 4.5 If events (14) and (13) happen and (4) holds, then, the set \( U = \{ A, R \}_i | R = \gamma n \} \) forms a candidate set (defined in Definition (4.4) for \( K \).

**Proof:** Since \( P, R \in K \) for all \( R, (14) \) directly implies \( \text{Dist}(U, K) < e^2 \nu / 12 \). Now, from (13), the \( S_k \) there have \( || P_{S_k} - M_{\ell} || < e^2 \nu / 12 \) by convexity of \( || \cdot \| \| \). Also \( || P_{S_k} - A_{S_k} || < e^2 \nu / 12 \) by (14). Adding and using the triangle inequality, we get \( \text{Dist}(\{ M_{1}, M_{2}, \ldots, M_{k} \}, U) \leq e^2 \nu / 6 \) as required.

**Theorem 4.6 (Vertex Set Certificate Theorem)** Let \( U \) be a candidate set for \( Q \) as in Definition (4.4). Then, for any \( w_1, w_2, \ldots, w_k \in U \) satisfying

\[
\text{Dist}(U, CH(w_1, w_2, \ldots, w_k)) < e^2 \nu / 4, \tag{15}
\]

the following is true

\[
\text{Dist}(\text{Vertex set of } Q, \{ w_1, w_2, \ldots, w_k \}) < e \nu. \tag{16}
\]

**Proof Sketch:** The proof can be found in the Supplementary material. Here we only give a brief outline. We prove it for any polytope \( Q \) satisfying the geometric characteristics of \( K \), facilitating its subsequent applicability to the problem at hand. A piece of terminology will be useful: for a vector \( v \in \mathbb{R}^d \) and two points \( x, y \in \mathbb{R}^d \), the \( v \)-distance between \( x, y \) is \( |v \cdot (x - y)| \).

First, using (4), for each \( \ell \), we apply the Separating Hyperplane Theorem from Convex Geometry to produce a unit vector \( v^{(\ell)} \) such that the \( v^{(\ell)} \) distance of \( M_{\ell} \) from any other \( M_{\ell'} \) is at least \( 2e^{2} \nu \). Then, we define a region \( Q_{\ell} \) as the set of points within \( v^{(\ell)} \) distance \( e^{2} \nu / 12 \) of \( M_{\ell} \) and close to \( K \). The proof shows that if a set \( \{ w_1, w_2, \ldots, w_k \} \in U \) satisfies the hypotheses of Theorem (4.6), then we must have that each \( Q_{\ell} \) contains one of \( w_1, w_2, \ldots, w_k \). Renumber the \( w_k \) so that \( w_k \in Q/v^{(\ell)} \), \( w_k \) is close to some point, say, \( w_k' \in K \). \( w_k' \) is a convex combination of \( M_1, M_2, \ldots, M_k \), and if the convex combination did not attach weight almost 1 to \( M_{\ell} \), then, since the \( M_{\ell}, \ell' \neq \ell \) are at \( v^{(\ell)} \) distance at least \( 2e^{2} \nu \) from \( M_{\ell} \); we are able to show that \( w_k \) would end up being too far away from \( M_{\ell} \) in the \( v^{(\ell)} \) direction to be in \( Q_{\ell} \) producing a contradiction. So, the weight \( w_k' \) attaches to \( M_{\ell} \) must be nearly 1 and this will imply that \( w_k' \) and therefore \( w_k \) is close to \( M_{\ell} \) to finish the proof.

**Proof Sketch (of Theorem 4.1):** The proof has the following three steps:

**Step 1:** From Stochastic Assumptions to Events From Lemma (4.4) it follows if the stochastic assumptions \( (5) \) and \( (6) \), and the lower bounds on \( \beta \) (see (8)) and \( n \) (see (7)) hold, then Subset-perturbation Event \( (14) \) and Proximate latent points Event \( (13) \) happen with high probability.

**Step 2:** From Events to Candidate set If events \( (14) \) and \( (13) \) happen and \( (4) \) holds, Lemma (4.5) shows that the set \( U \) of \( \binom{n}{\gamma n} \) \( A, R \}_i | R | = \gamma n \} \) forms a candidate set (see Definition (4.4) below) for \( K \).

The Vertex Certificate theorem (Theorem (4.6) ) shows that \( U \) being a candidate set for \( K \) implies that if we find any \( k \) points \( w_1, w_2, \ldots, w_k \in U \) with \( \text{Dist}(U, CH(w_1, w_2, \ldots, w_k)) < \epsilon e^2 \nu / 4 \), then, \( w_1, w_2, \ldots, w_k \) is a solution to \( \text{LKP} \), namely,

\[
\text{Dist}(\{ M_1, M_2, \ldots, M_k \}, \{ w_1, w_2, \ldots, w_k \}) < \epsilon \nu.
\]

Further, the existence of \( w_1, w_2, \ldots, w_k \) is also guaranteed, as shown in the proof.

**Step 3:** From Candidate Set to Learner Based on the above discussion, to complete the proof we now need to prove that one can finds \( w_1, w_2, \ldots, w_k \) in finite time. More precisely, one needs to find \( R_1, R_2, \ldots, R_k \subset U, |R_i| = \gamma n, i \in [k] \) such that

\[
h(R_1, R_2, \ldots, R_k) \leq \frac{1}{4} e^2 \nu
\]

where \( h(R_1, R_2, \ldots, R_k) \) is defined as

\[
\max_{R \subseteq U : |R| = \gamma n} \text{Dist}(A, R, CH(A, R_1, A, R_2, \ldots, A, R_k))
\]

This problem is solved through enumerating all possible \( k + 1 \) subsets of size \( \gamma n \), and each enumeration requires solving a convex program. Since there are at most \( \binom{n}{\gamma n} \) \( k+1 \) convex programs to solve, the running time is finite.

**Remark 4.1** Intuitively, the proof builds on the fact that (i) each \( A, R \) is close to \( K \) and further, (ii) there is a collection of \( k \) subsets \( R_1, R_2, \ldots, R_k \) such that

\[
D(\{ A, R_1, A, R_2, \ldots, A, R_k \} , \{ M_1, M_2, \ldots, M_k \}) \tag{17}
\]

is small. [Definition (4.4) and (4.6) will state these precisely.] It seems like at this point, we are done - just enumerate all collections of \( k \) subsets \( R_1, R_2, \ldots, R_k \) and check the condition (17) for all. But we do not know \( K \) and so cannot check (17). This brings us to the technically crucial piece of the paper which is a data-based finitely checkable sufficient condition for when we have found approximations to the vertices of \( K \), i.e., when a proposed collection of \( k \) subsets \( R_1, R_2, \ldots, R_k \) in fact solves \( \text{LKP} \). This piece is answered by the Vertex Certificate theorem (Theorem (4.6) )

In a later section we will provide a matching lower bound thus proving the optimality of the bound.

**Sufficient conditions of Learnability:** The upper bound on Sample Complexity in Theorem 4.1 can be applied to
any model which can be posed as a special case of LKP. To apply the bound one needs to check if prior Dk on W satisfies Vertex Proximate Assumption(5) and if Aggregate Sub-gaussian Assumption holds with high enough β (defined in (6)). When expressed in words the sufficient conditions for learnability translates to (i) that the prior put sufficient weight near the corners of the standard simplex, (ii) that probability of data given latent point be sub-Gaussian with high enough β and (iii) and that the means are well-separated.

5. Ad-Mixtures: Upper bound

Ad-mixture is a special case of LKP. In this section we will discuss LDA, a special case of Ad-mixtures and applicability of the Theorem 4.1 presented in the earlier section.

5.1. LDA: Upper bound

In this section, we prove an upper bound on the sample complexity of Latent Dirichlet Allocation (LDA), the most widely used model for topic modeling. We let n, d, m be the number of documents, number of words in the dictionary and the number of words in each document respectively. The main theorem (Theorem 5.1) of this section proves that under mild assumptions (explained shortly), as long as the total number of words nm in all documents is at least $\Omega^*(dk)$ (where * hides logarithmic factors and factors in $\varepsilon$), there is a successful learner and so $O^*(dk)$ tokens (words) suffice. Since there are $dk$ parameters and the input has $nm$ free variables, this is near-optimal. Now we describe the mild assumptions.

The first assumption is on the number of documents, $n \in \Omega^*(k \ln k)$ (see (20)). It is necessary from the Coupon Collector problem, since, otherwise, with high probability, there will be a topic for which, we do not see any document. We also assume that each document has some minimum number of words, more precisely $m \geq c\ln(k/\varepsilon)/\varepsilon^4$(see 20).

We also assume (see (19)) each topic $\ell$ has an associated set of words $T_\ell$ on which it puts more total weight than other topics. This is similar in spirit to the assumptions made in the literature, and has been refereed to as Catchwords in (Bansal et al., 2014). But it is weaker in that the $T_\ell$ are not required to be disjoint, neither do we require as big a difference on the total weight on $T_\ell$ in topic $\ell$ as compared to other topics, nor is there a requirement that individual words have high frequency as assumed in earlier models in the literature.

Remark 5.1 The important paper of (Tang et al., 2014) deals with posterior contraction rather than sample complexity. But their main theorem essentially says in the limit, the Hausdorff distance in Euclidean norm between the true $M$ and the approximation computed by Gibbs sampling goes down as $O(\sqrt{\ln m}/m)$. So to make $L_1$ error at most $\varepsilon$ by a direct application of their result will need $m > \omega^*(d)$. This is strictly speaking incomparable to our result (indeed they have no dependence on $k$, which as they say is surprising), but often $m << n$ and so $m > d$ is perhaps a strong requirement. It is possible, however, that their techniques may be applied directly for $L_1$ norm to get better results than this implication of their current theorem.

We assume $P_{\cdot,j} \sim Dir(k, \frac{1}{k})$, an often used setting in Topic Modelling. It is known that for $Dir(k, \alpha)$ as $\alpha$ becomes $< 1$, the Dirichlet distribution puts substantial mass near the corners of the simplex (see (Telgarsky, 2013) for a proof). This favors Vertex Proximate Assumption and indeed we will show that with $\alpha = 1/k$, (see Lemma (5.3)) holds. This allows us to prove an upper bound stated below.

Theorem 5.1 (LDA Upper bound) Let $\mathcal{M}(d, k)$ be the collection of $d \times k$ matrices with non-negative entries and column sums equal to 1. Let $\varepsilon \in (0, 1)$. Suppose $W_{j, \cdot} = 1, 2, \ldots, n$ are iid, each distributed according to $Dir(k, 1/k)$. Let

$$\gamma = \frac{ce^4}{k}. \quad (18)$$

Assume there are subsets $T_1, T_2, \ldots, T_k$ of $[d]$ with

$$\forall \ell \neq \ell', \sum_{i \in \ell} M_{i, \ell} > \sum_{i \in \ell'} M_{i, \ell'} + 2\varepsilon. \quad (19)$$

Assume $d, n, m$ satisfy the conditions

$$n \geq c\frac{k \ln(1000k)}{\varepsilon^4}, \quad m \geq c\frac{\ln(k/\varepsilon)}{\varepsilon^4}, \quad nm \geq \frac{cdk}{\varepsilon^8} \quad (20)$$

and $d > 5$. Then, there is a learner which given $n$ iid samples drawn according to $\mathbf{M} \in \mathcal{M}(d, k)$, solves $(\frac{\alpha}{\gamma n})^k$ convex programs each of poly$(n, d)$ size and returns $v_1, v_2, \ldots, v_k$ which with probability greater than 0.99 satisfy

$$D(\{v_1, v_2, \ldots, v_k\}, \{M_{1,1}, M_{1,2}, \ldots, M_{1,k}\}) \leq \varepsilon.$$ 

So, the $\varepsilon$-sample complexity of $\mathcal{M}(d, k)$ is $O(dk/m\varepsilon^8)$ provided $d \in \Theta^*(m)$ and $m \in \Omega^{\ast}(1)$.

Proof: We will appeal to Theorem (4.1) to prove Theorem (5.1). To this end, we state three Lemmas after this theorem (indeed they have no dependence on $k$, which as they say is surprising), but often $m << n$ and so $m > d$ is perhaps a strong requirement. It is possible, however, that their techniques may be applied directly for $L_1$ norm to get better results than this implication of their current theorem.
then (4) holds, namely,
\[ \text{Dist}(M_{\ell}, C H(M_{\ell'}, \ell' \neq \ell)) \geq 2\varepsilon. \]

The existence of the sets \( T_i, i \in [k] \) have been already noted in Topic Modelling literature and is known as Catchwords (Bansal et al., 2014).

**Lemma 5.3 (Vertex Proximate Assumption)** Suppose \( W_{ij}, j = 1, 2, \ldots, n \) are i.i.d., each distributed according to \( \text{Dir}(k, 1/k) \). Let \( \gamma = \frac{c\varepsilon^4}{k} \). Then, (6) holds.

**Lemma 5.4 (Aggregate Sub-gaussian Assumption)** Let \( \gamma = \frac{c\varepsilon^4}{k} \) and \( d > 5 \). (6) holds with \( \beta = m \).

### 6. Lower bounds

We derive a lower bound matching the upper bound (within \( O^*(1) \)) using a code-design similar to the one used in (Suresh et al., 2014). For the lower bound, we assume \( \gamma = 1/k \) and most importantly we also assume that each document is on a single topic. Thus Vertices Proximate Assumption (5) is automatically satisfied. Since we are proving lower bounds, these restrictions only make the bound stronger.

We generate documents as in LDA - each word is independently generated according to the multinomial with the number of tokens is \( o(m) \), the total number of parameters.

The proof is by a counting argument; we show by a simple code-design that there is a large a collection of \( \mathbf{M} \) which are

(i) pairwise far enough apart that a learner must output different answers on any two different \( \mathbf{M} \) to be successful and

(ii) each \( \mathbf{M} \) in the collection satisfies the Separation Assumption (4).

(iii) After constructing this collection of \( \mathbf{M} \), we show that the number of possible sets of \( n \) documents with \( m \) words each is much smaller than the number of \( \mathbf{M} \) in the collection, and since the identifier is deterministic, and we arrive at a contradiction. The proof applies even if one assumes no stochastic model of data generation.

Of these it turns out that (ii) is harder to ensure.

**Theorem 6.1 (LDA Lower Bound)** Suppose \( \gamma = 1/k \) and each document is purely on a single topic (hence Proximate Data Assumption (13) is satisfied) and suppose Separation Assumption (4) is satisfied. Let \( \varepsilon = .001 \). If
\[ nm < \frac{(0.09d - \ln k)k}{\ln d}, \]

then, no learner achieving error less than \( \varepsilon \) on all \( \mathbf{M} \) exists.

The proof is in the Supplemenatry. The lower bound on the number of tokens required for LDA also implies a lower bound on the number of samples. Since LDA is a special case of LKP, it also implies a lower bound on the sample complexity. More precisely, the following corollary of the LDA lower bound theorem establishes the lower bound on the sample complexity of LKP.

**Corollary 6.2 (LKP Lower bound)** In the LKP problem, if \( n \in o(dk/\beta \ln d) \), there is no .001-learner.

**Proof:** The proof follows directly from Theorem (6.1) using the fact that \( \beta \) was shown to be equal to \( m \), the number of words per document in Lemma (5.4).

The matching upper bounds and lower bounds suggests that the assumptions of the Main theorem are indeed sufficient conditions for learning and furthermore the generality of the result suggests immediate applicability to many other models.

### 7. Mixed Membership Stochastic Blocks (MMSB)

MMSB models are generative models of overlapping communities widely used in Social Network Analysis (Airoldi et al., 2008). In this section, we prove the first upper bound on the sample complexity of \( O^*(k^2) \) for MMBM, which is also near-optimal.

**The Model:** There are \( n \) people and \( k \) communities. As in LKP, there is a prior \( D_k \) on the simplex \( \Delta_{k-1} \). For \( j = 1, 2, \ldots, n \), person \( j \) chooses \( (V_{ij}, j = 1, 2, \ldots, n) \) according to \( D_k \) independently of \( j \). \( V_{ij} \) is the membership-weight of person \( j \) in community \( \ell \). There is a \( k \times k \) “meeting-probability” matrix \( B \). The probability \( P_{j,j'} \) that person \( j \) meets person \( j' \) is defined as follows
\[ P_{j,j'} = \sum_{\ell_1, \ell_2} V_{i_1,j} B_{\ell_1, \ell_2} V_{i_2,j'}, \quad \mathbf{P} = \mathbf{V}^T \mathbf{B} \mathbf{V}. \]

The data consists of 0-1 matrix \( \mathbf{A} \), where, \( A_{jj'} \) are independent Bernoulli random variables with \( \Pr(A_{jj'} = 1) = P_{j,j'} \).

**MMSB as a special case of LKP:** MMSB can be formulated as an LKP by considering the following setup. Partition the set of people into two sets \( S_1, S_2 \) where with \( |S_1| = d \) and \( |S_2| = n \); the partition is chosen randomly subject to their cardinalities. \( d \) will be specified later. We then consider only the bipartite graph with edges between \( S_1 \) and \( S_2 \). For ease of notation, we henceforth write \( |S_1| = d; |S_2| = n \) by resetting the value of \( n \).
Let $U$ be the $k \times d$ submatrix of $V$ containing the columns of $V$ corresponding to $S_1$ and $W$ be the $k \times n$ submatrix containing columns in $S_2$. $U$, respectively, $W$ has the community membership weights of members of $S_1$, respectively $S_2$ and so $U$ is statistically independent of $W$. [The partitioning trick is used precisely for this “de-conditioning” of $U$ and $W$.] Now the new $P$, $A$ are $d \times n$ matrices with entry $(i, j)$ for $i \in S_1$, $j \in S_2$ and we have with $M = U^T B$:

$$P = U^T B W : P = MW.$$

Now, we do have a LKP, where we are given $A$ and need to estimate $M$. This does not automatically give us $B$ requiring further assumption. Recall $\nu$ is the diameter of $K$ and is equal to $\nu = \max_{i,j} \sum_{i=1}^d M_{i,j}$.

**Core group Assumption** Let $d \in \Omega(c k \ln k / \text{poly}(\varepsilon))$ and $S_1 = \{1, 2, \ldots, d\}$ denote a random subset of people. Let $U$ be the $k \times d$ submatrix of $V$ containing the $[d]$ columns of $V$. Our assumption is that given an $M$ satisfying $D(\{M_{1,1}, M_{2,2}, \ldots, M_{k,k}\}, \{M_{1,1}, M_{2,2}, \ldots, M_{k,k}\}) \leq \varepsilon^2 / 4 \ln k$, we can identify the columns of $B$ to within $L_1$ error $\varepsilon^2$, where $\nu$ is the maximum column sum of $B$.

Under the Core Group Assumption, we can now reformulate the problem as one of finding $U^T B$ which is a proxy for $B$. The columns of $M = U^T B$ are now the vertices of the latent polytope. To avoid any conditioning, we assume the people in $S_1$ have already chosen their $V_{j,k}$’s.

**Theorem 7.1 (MMSB Upper Bound)** Let $\varepsilon \in (0, 1)$. Consider an MMSB model parametrized by a $k \times k$ meeting probability matrix $B$ with $\nu' = \max_{j \in [k]} \sum_{i=1}^k B_{i,j}$. If the model satisfies

a.) each $V_{j,k}$ is picked according to $\text{Dir}(k, 1/k)$.

b.) there are subsets $T_1, T_2, \ldots, T_k$ of $[d]$ with

$$\forall \ell \neq \ell', \sum_{i \in T_\ell} M_{i,\ell} > \sum_{i \in T_{\ell'}} M_{i,\ell'} + 2\varepsilon \nu. \quad (21)$$

c.) Core Group Assumption holds,

d.) $n \nu > \frac{c^2 (\ln k)^2}{\text{poly}(\varepsilon)}$; $\nu > \frac{c \ln^2 (k / \varepsilon')}{\varepsilon^2}$.

then the following holds

i.) The three assumptions, namely Well Separated Assumption(4), Vertex Proximate Assumption(5) and Aggregate Sub-gaussian Assumption(6) are satisfied, with $\gamma = \frac{c^2}{k}, \beta = \nu$.

ii.) there is a finite time procedure which given $n$ iid samples, returns a $k \times k$ matrix $B$ such that with probability at least 0.99, we have $D(\{B_{1,1}, B_{2,2}, \ldots, B_{k,k}\}, \{B_{1,1}, B_{2,2}, \ldots, B_{k,k}\}) \leq \varepsilon^2 n \nu$.

**Remark 7.1** The inequality (21) is akin to the Catchwords assumption for LDA. In this setting, it is intuitively validated by the following reasoning: As we argued above, whp, there are $O^*(1)$ people in $S_1$, each with weight almost 1 on community $\ell$. Let $T_\ell$ be the set of these people. So, for $i \in T_\ell$, $M_{i,\ell}$ is the intra-community meeting probability, whereas, for any $\ell' \neq \ell$, $M_{i,\ell'}$ is an inter-community meeting probability. Under usual assumptions that intra-community meeting probabilities are substantially greater than inter-community probabilities, (21) holds.

The sample complexity being $O^*(k^2 / \nu)$ means it suffices to have $n \nu$ (which is the expected number of edges in the graph) be at least $O^* (k^2)$ (which is the number of parameters of the model, namely, the number of entries of $B$ we are trying to learn). Hence the claim of near-optimality.

The full proof of the Theorem is given in the Supplementary. We note that of the three conclusions, Aggregate Sub-gaussian Assumption(6) and Well Separated Assumption(4) are proved exactly as in LDA. The crucial difference is in the proof of Vertex Proximate Assumption(5) which uses now a different probability inequality, since, the model is different. The (ii) follows by application of Theorem 4.1 and Core group assumption.

**Theorem 7.2 MMBM Lower Bound** Suppose $\gamma = 1/k$ and each column of $P$ is equal to some column of $M$ (hence Extreme Data Assumption (13) is satisfied) and suppose Separation Assumption (4) is satisfied. Let $\varepsilon = .001$. If

$$n \nu < \frac{(0.09 d - \ln k)k}{\ln (d/2 \nu)},$$

then, no identifier exists.

8. Conclusion

The Optimal Sample complexity of LKP demonstrates the linear dependence on the number of parameters. The generality of LKP along with the tools developed suggests a straightforward procedure for deriving Sample complexity estimates for Ad-mixture models. Similar results can be derived for other models, such as Dirichlet Simplex Nest (Yurochkin et al., 2019), as well. The techniques developed are also novel and should also be of interest, specially for set-estimation problems (Brunel, 2018).
Near-optimal sample complexity via Latent \( k \)-Polytopes

References


