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

C. Bhattacharyya [1] **R. Kannan** [2]

In this supplement, apart from the proofs, we also provide intuitions behind our techniques. As in the main paper, $\|\cdot\|$ stands for ℓ_1 norm and B for the unit ball of ℓ_1 norm, namely, $B = \{x \in \mathbf{R}^d : ||x|| \leq 1\}$.

A. Section 4: Intuition behind techniques

At the outset, the challenge is that data points are (highly) perturbed versions of latent points and most/all data points lie outside the latent polytope K. But, in principle, this can be overcome by averaging large subsets of data. Indeed, a standard statistical technique for the toy case of $k = 1$ is that the single vertex of K is well estimated by the average of all data points. An extension of this is used in traditional Clustering: once the data points have been partitioned into clusters, the mean is estimated by the average of the points in the cluster. [Of course the algorithmically harder part is to find the partition.] A starting observation for finding the Latent $k-$ polytope based on taking averages of subsets of data points is: Let U be the set of $\binom{n}{\gamma n}$ averages of all $\gamma n-$sized subsets of data. By Proximate Latent points assumption (intuitively), there are approximations to the k vertices of K among elements of U. Under the Sub-Gaussian assumption, (intuitively), since all ($\gamma n-$ sized) subset averages of data are close to corresponding subset averages of latent points (which are all in$K$), we have the following two properties:

(a) Each element of U is close to K and

(b) There are sets $S_1, S_2, \ldots, S_k$, $|S_\ell| = \gamma n$, such that average of data in subset S_ℓ is close to the ℓ th vertex of K for $\ell = 1, 2, \ldots, k$.

Property (a) implies that the convex hull of any k points of U is approximately contained in K. Property (b) implies that there are k points in U whose convex hull approximately contains K. If these two statements had been true exactly instead of approximately, the extreme points of U would suffice. Namely, k points of U with the property that all points of U are in the convex hull of these k points would be the vertices of the latent polytope and we would be done. Since (a) and (b) only hold approximately, the central problem we tackle is an approximate analog of the above reasoning. We formulate specific conditions (11) and (12) of the Definition 4.4. of a candidate set which are the quantitative versions of (a) and (b)

Before we describe the use of candidate set, a note on our stochastic model, especially, the Sub-Gaussian assumption (4) is in order. A simpler model of perturbations (of data points from latent points) would have been an upper bound on individual perturbations (rather than subset average perturbations). But we do not know of a hypothesis on individual perturbations which is realistic and still guarantees that for EVERY $\gamma n-$subset of data, the average is close to the average of the corresponding subset of latent points. For example, for LDA, the best absolute bound we can assume on $||A_{\cdot,j} - P_{\cdot,j}||$ is that $||A_{\cdot,j} - P_{\cdot,j}|| \leq c\forall j$. This only implies that for each $R, |R| = \gamma n$ and each $v \in \{-1, 1\}^d$,

$$\text{Prob}(v \cdot (A_{\cdot,R} - P_{\cdot,R}) \geq \lambda) \leq \exp(-c\lambda^2\gamma n).$$

But, we have to union this over the $\binom{n}{\gamma n} \approx \exp(\gamma n \ln(1/\gamma))$ subsets and we do not get any non-trivial bound unless $\lambda > \sqrt{\ln(1/\gamma)}$, which is too large for our use. We circumvent this by requiring $\beta \geq c\ln(1/\gamma)/\varepsilon^4$ (condition (8) of Theorem 4.1). We prove that (8) holds in applications.

Also, another word of explanation for the remarks immediately preceding Definition 4.3 is in order. The vector valued random variable

$$X_R = \frac{\sqrt{|R|}}{\nu}(A_{\cdot,R} - P_{\cdot,R}) = \frac{1}{\nu\sqrt{|R|}}\sum_{j\in R}(A_{\cdot,j} - P_{\cdot,j})$$

is the sum of $|R|$ independent random variables: $(A_{\cdot,j} - P_{\cdot,j})/\nu$ normalized (as in Central Limit Theorem) by $\sqrt{|R|}$. For real-valued random variables, CLT gives us sub-Gaussian tail bounds. By looking at all 1-d marginals,

[1]Department of Computer Science and Automation, Indian Institute of Science, Bengaluru-560012, India [2]Microsoft Research Labs India, Bengaluru-56001, India. Correspondence to: C. Bhattacharyya <chiru@iisc.ac.in>, R. Kannan <kannan@microsoft.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

namely, $\frac{1}{\nu\sqrt{|R|}}\sum_{j\in R} v\cdot(A_{\cdot,j}-P_{\cdot,j})$, we can first deduce sub-Gaussian behavior of the normalized sum for a fixed v and then take the union bound over all $v\in\{-1,1\}^d$ if we use (as we do here) ℓ_1 norm or over an ε-net for other norms which puts an additive term $+d$ in the exponent. CLT does not give us finite sample bounds, but, by definition of Sub-Gaussian norm of a vector valued random variable (see Definition 5.22 of (Vershynin, 2010)), we get (4) with β as the sub-Gaussian norm.

Now, with (a) and (b) in place, since, we are not worried here about our algorithm taking exponential time (sample complexity is our focus), we could:

- enumerate all collections $R_1, R_2, \ldots, R_k$ of k subsets of data of cardinality γn each and
- if only we could check for each enumerated collection $R_1, R_2, \ldots, R_k$, whether, $\text{Dist}(\{M_{\cdot,1}, M_{\cdot,2}, \ldots, M_{\cdot,k}\}, \{A_{\cdot,R_1}, A_{\cdot,R_2}, \ldots, A_{\cdot,R_k}\})$ was small enough, we would find the answer at some point, when we enumerate $S_1, S_2, \ldots, S_k$, if not earlier. But, we do not know $M_{\cdot,\ell}$, and we know of no such easy check.

So, the question is: Is there a (purely) data-determined test characterizing $R_1, R_2, \ldots, R_k$ which satisfy: $\text{Dist}(\{M_{\cdot,1}, M_{\cdot,2}, \ldots, M_{\cdot,k}\}, \{A_{\cdot,R_1}, A_{\cdot,R_2}, \ldots, A_{\cdot,R_k}\})$ being small ?

We do not know of a characterizing test. But we note that we do not need a characterization. A (sufficient) condition on $R_1, R_2, \ldots, R_k$ which implies $\text{Dist}(\{M_{\cdot,1}, M_{\cdot,2}, \ldots, M_{\cdot,k}\}, \{A_{\cdot,R_1}, A_{\cdot,R_2}, \ldots, A_{\cdot,R_k}\})$ is small and which has properties (c) and (d) below will do:

(c) If the condition is satisfied by $R_1, R_2, \ldots, R_k$, then $\text{Dist}(\{M_{\cdot,1}, M_{\cdot,2}, \ldots, M_{\cdot,k}\}, \{A_{\cdot,R_1}, A_{\cdot,R_2}, \ldots, A_{\cdot,R_k}\})$ is small.

(d) There exists some collection of k sets $R_1, R_2, \ldots, R_k$ which satisfy the condition.

We can first ask what seems to be a simpler question: Given one set $R, |R| = \gamma n$, is there a data-determined sufficient condition such that if R satisfies the condition, then it is certified (not just with high probability, but deterministically) that $A_{\cdot,R}$ is close to some vertex of K ? Actually, no such condition is known.

What we do prove here is Theorem 4.6 (Vertex Set Certificate Theorem): If $R_1, R_2, \ldots, R_k$ satisfy the condition that $\text{Dist}(\{A_{\cdot,R} : |R| = \gamma n\}\ ,\ \{A_{\cdot,R_1}, A_{\cdot,R_2}, \ldots, A_{\cdot,R_k}\})$ is small and Separation Condition (4) and (11) and (12) are satisfied, that is sufficient for $\text{Dist}(\{M_{\cdot,1}, M_{\cdot,2}, \ldots, M_{\cdot,k}\}, \{A_{\cdot,R_1}, A_{\cdot,R_2}, \ldots, A_{\cdot,R_k}\})$ to be small.

The reader may wonder if there is a simpler test for when a given set $\{w_1, w_2, \ldots, w_k\}$ is within small Hausdorff distance of the set of vertices of K. We do not know for sure if it is possible, but here we discuss why all of our conditions (4), (11) and (12) seem necessary by giving examples to show that if one of the conditions fails, then the conclusion also fails.

First, Theorem 4.6 and Lemma A.1. together imply that if the separation assumption (4), conditions (11) and (12) of Definition 4.4 and condition (15) of Theorem 4.6 hold, then, condition (16) of Theorem 4.6 gives us a test of when we have approximation to the set of vertices of K.

Examples when (4) is violated are obvious. We give an example when (11) alone is violated and the conclusion fails: Take $d = k$ and say K is a proper subset of the simplex $\Delta_k = \{x : x_\ell \geq 0, \sum_\ell x_\ell = 1\}$ satisfying (4). Take as U an $c\varepsilon^2-$net of the convex hull of the k unit vectors and also add to U the k unit vectors. (12) is satisfied by the ε^2-net property. (15) is satisfied if we take $w_\ell =$ to be the ℓ th unit vector. But clearly, the conclusion (16) is not satisfied if K is substantially smaller than Δ_k.

An example where (12) alone is violated is the following: Take $k = 3, d = 2$ and K to be the triangle with vertices $(-1,0),(1,0),(0,5\varepsilon)$ and $U = \{(-1,0),(1,0),(-0.9,0.5\varepsilon)\} = \{w_1, w_2, w_3\}$, for ε a small enough positive real. It is easy to see that (4) (11), (15) are satisfied, but (12) is not and the conclusion (16) is not.

A slightly tighter example is essentially the same as the one above, but with $U = \{(-1,0),(1,0),(-0.25,3.75\varepsilon)\} = \{w_1, w_2, w_3\}$. In this case (12) is violated, but a weaker condition: Dist(Set of vertices of $K, CH(U)) \leq \varepsilon\nu$ (note: $\nu = 2$) is satisfied. This shows that we need that each vertex of K be well-approximated by a point of U, rather than a point just in $CH(U)$.

We formulate a sufficient condition (stated in (15)) for a collection of k sets $R_1, R_2, \ldots, R_k$ to have the property that the set $\{A_{\cdot,R_\ell}, \ell = 1, 2, \ldots, k\}$ is close in Hausdorff distance to the set of the k vertices of K.

The main technical workhorse of the paper is Theorem 4.6 which proves (c) for this condition. The sufficient condition in the notation of this discussion reads:

$$\text{Dist}(U, CH(A_{\cdot,R_1}, A_{\cdot,R_2}, \ldots, A_{\cdot,R_k}) \text{ is small }. \quad \text{(A.1)}$$

Note that this condition is data-determined (though in exponential time). We also prove (d) for this condition in Lemma (A.1) stated below.

A.1. Proofs for Section 4

Proof: (Of Theorem 4.1) Theorem 4.1 follows directly from Lemma 4.4, Lemma 4.5 and Theorem 4.6. Lemma 4.4 proves that Vertex Proximate Assumption (5) which is a hypothesis of Theorem 4.1 implies the event (13) and the Aggregate Subgaussian Assumption (6) (another hypothesis of Theorem 4.1) implies the event (14). Under these conditions, Lemma 4.5 proves that the collection U of $\binom{n}{\gamma n}$ averages of all γn sized subsets of data forms a "candidate set" as defined in Definition 4.4. Then, Theorem 4.6 (the Vertex Certificate Theorem) proves sufficient condition (15) which certifies a set of k points in U to be close to the set of vertices of K. It is easy to check that condition (15) can be checked by trying out all $\binom{\binom{n}{\gamma n}}{k}$ collections of k points of U and for each solving convex programs to check if each $A_{\cdot,R}, |R| = \gamma n$ is close enough to the convex hull of the set of k points of U. ∎

Proofs of Lemma 4.4 and Lemma 4.5 are in the main paper and we do not discuss them here.

Lemma A.1 *For the $S_1, S_2, \ldots, S_k$ in (13), we have*

$$\begin{aligned} &\textit{Dist}(\{A_{\cdot,R} : |R| = \gamma n\}, CH(A_{\cdot,S_1}, A_{\cdot,S_2}, \ldots, A_{\cdot,S_k})) \\ &< \varepsilon^2\nu/4. \end{aligned} \tag{A.2}$$

Intuitive Description of the proof $P_{\cdot,R}$ is in K, so is a convex combination of the $M_{\cdot,\ell}$'s. Using (13), we can show that each $P_{\cdot,S_\ell}$ is close to $M_{\cdot,\ell}$ and this will imply that the same convex combination of the $P_{\cdot,S_\ell}$ will be close to $P_{\cdot,R}$. By (14), each $A_{\cdot,S_\ell}$ is close to $P_{\cdot,S_\ell}$ and also $A_{\cdot,R}$ is close to $P_{\cdot,R}$. So the same convex combination of the $A_{\cdot,S_\ell}$. we prove, is close to $A_{\cdot,R}$ using triangle inequality on norms several times.

Proof: For any R, $P_{\cdot,R} \in K$ and can be expressed as a convex combination of $M_{\cdot,\ell}$. Suppose $P_{\cdot,R} = \sum_{\ell=1}^{k} \alpha_\ell M_{\cdot,\ell}$ is the convex combination. We will prove that $\sum_\ell \alpha_\ell A_{\cdot,S_\ell} \approx A_{\cdot,R}$ to establish (A.2). For all $j \in S_\ell$, by (13), $||P_{\cdot,j} - M_{\cdot,\ell}|| < \varepsilon^2\nu/12$ which implies by convexity of $||\cdot||$ that

$$||P_{\cdot,S_\ell} - M_{\cdot,\ell}|| \leq \frac{1}{|S_\ell|}\sum_{j\in S_\ell} ||P_{\cdot,j} - M_{\cdot,\ell}|| < \varepsilon^2\nu/12. \tag{A.3}$$

Now, using (14) and (A.3),

$$\begin{aligned} &||\sum_\ell \alpha_\ell A_{\cdot,S_\ell} - P_{\cdot,R}|| = ||\sum_\ell \alpha_\ell (A_{\cdot,S_\ell} - M_{\cdot,\ell})|| \\ &\leq \sum_\ell \alpha_\ell ||A_{\cdot,S_\ell} - M_{\cdot,\ell}|| \\ &\leq \sum_\ell \alpha_\ell ||A_{\cdot,S_\ell} - P_{\cdot,S_\ell}|| + \sum_\ell \alpha_\ell ||P_{\cdot,S_\ell} - M_{\cdot,\ell}|| \\ &< \varepsilon^2\nu/6, \end{aligned}$$

where, the first inequality is by convexity of $||\cdot||$ and the second by triangle inequality. This implies by (14):

$$\begin{aligned} ||\sum_\ell \alpha_\ell A_{\cdot,S_\ell} - A_{\cdot,R}|| &\leq ||\sum_\ell \alpha_\ell A_{\cdot,S_\ell} - P_{\cdot,R}|| \\ &\quad + ||P_{\cdot,R} - A_{\cdot,R}|| \\ &< \varepsilon^2\nu/6 + ||P_{\cdot,R} - A_{\cdot,R}|| \leq \varepsilon^2\nu/4, \end{aligned}$$

completing the proof of (A.2). ∎

Next, we prove Theorem 4.6. This theorem is the technical heart of the paper. A brief explanation of its role was given in Remark 4.1. We will provide intuition (so marked) before each stage of the proof.

Proof: (Of Theorem 4.6) The Theorem holds for any polytope Q. But to avoid extra notation, we prove it (without loss of generality) for the polytope K.

Intuition The first part of the proof (up to just before Claim (A.1)) uses just the hypothesis that each vertex of Q is at L_1 distance at least $\varepsilon \text{Dia}_{L_1}(Q)$ from the convex hull of the other vertices which is satisfied by K from assumption (4) to construct a separating hyperplane and then a region Q_ℓ near vertex ℓ of K (which we may think of as a "region of attraction" for that vertex). Formally:

Proof Assumption (4) says that the following two convex sets are disjoint:

$$[M_{\cdot,\ell} + 2\varepsilon\nu B] \cap CH(M_{\cdot,\ell'} : \ell' \neq \ell) = \emptyset.$$

Thus, by the Separating Hyperplane Theorem from Convex Geometry, there is a vector $v^{(\ell)}$ such that

$$\forall \ell' \neq \ell\ ,\ \forall x \in 2\varepsilon\nu B\ ,\ v^{(\ell)} \cdot M_{\cdot,\ell} + v^{(\ell)} x > v^{(\ell)} \cdot M_{\cdot,\ell'}. \tag{A.4}$$

After scaling by $||v^{(\ell)}||_\infty$, we may assume that $||v^{(\ell)}||_\infty = 1$ and (A.4) is still satisfied.

There is a y, with $v \cdot y = -2\varepsilon\nu$ and $y \in 2\varepsilon\nu B$. [There is an i with $v_i^{(\ell)} = \pm 1$. Define $y \in \mathbf{R}^d$ by $y_i = -2\varepsilon\nu v_i^{(\ell)}$ and $y_{i'} = 0 \forall i' \neq i$.]

Now, we get from (A.4):

$$v^{(\ell)} \cdot M_{\cdot,\ell} > v^{(\ell)} \cdot M_{\cdot,\ell'} + 2\varepsilon\nu \forall \ell' \neq \ell. \tag{A.5}$$

For $\ell = 1, 2, \ldots, k$, define a set Q_ℓ as follows (cf: Paragraph 3 of Outline):

$$\begin{aligned} Q_\ell = &\left(CH(\mathbf{M}) + (\varepsilon^2\nu/12)B\right) \\ &\cap \{x : v^{(\ell)} \cdot x > v^{(\ell)} \cdot M_{\cdot,\ell} - 5\varepsilon^2\nu/12\}. \end{aligned}$$

Claim A.1 *Suppose $w_1, w_2, \ldots, w_k \in U$ satisfy*

$$\textit{Dist}(U\ ,\ CH(w_1, w_2, \ldots, w_k) < \varepsilon^2\nu/4.$$

Then we must have

$$\forall \ell \in [k], \exists \ell' \in [k] : w_{\ell'} \in Q_\ell.$$

Intuition The claim is crucial; it argues that indeed if any set of k points $w_1, w_2, \ldots, w_k \in U$ satisfies the (quantitative version of the) sufficient condition (A.1) stated formally with $w_\ell = A_{\cdot,R_\ell}$, then, one of the k points must lie in the "region of attraction" of each vertex of K. After the claim is proved, we will prove in Lemma (A.2) that every point in Q_ℓ is close to $M_{\cdot,\ell}$ completing the proof of the Theorem.

Proof: Suppose the hypothesis of the claim is satisfied, but there is some Q_ℓ, which we assume wlg is Q_k such that no w_ℓ is in Q_k. But, $w_1, w_2, \ldots, w_k$ all belong to $CH(\mathbf{M}) + \varepsilon^2\nu/(12)B$ (by (9)). Since $w_1, w_2, \ldots, w_k \notin Q_k$, we must have (from the definition of Q_k):

$$w_\ell \cdot v^{(k)} \leq v^{(k)} \cdot M_{\cdot,k} - 5\varepsilon^2\nu/12 \text{ for } \ell = 1, 2, \ldots, k.$$

Consequently,

$$\begin{aligned} &\forall y \in CH(w_1, w_2, \ldots, w_k) \\ &v^{(k)} \cdot y \leq v^{(k)} \cdot M_{\cdot,k} - 5\varepsilon^2\nu/12. \end{aligned} \tag{A.6}$$

By (10), there is a point $u_k \in CH(U)$ such that

$$||M_{\cdot,k} - u_k|| < \varepsilon^2\nu/6. \tag{A.7}$$

By hypothesis of the current Claim that Dist$(U, CH(w_1, w_2, \ldots, w_k) < \varepsilon^2\nu/4$, we have using the convexity of the Dist function, there is a point $y \in CH(w_1, w_2, \ldots, w_k)$, with

$$||y - u_k|| < \varepsilon^2\nu/4.$$

So, using (A.6),

$$v^{(k)} \cdot u_k < v^{(k)} \cdot y + \varepsilon^2\nu/4 \leq v^{(k)} \cdot M_{\cdot,k} - \varepsilon^2\nu/6. \tag{A.8}$$

But, $||u_k - M_{\cdot,k}|| < \varepsilon^2\nu/6$, which implies $v^{(k)} \cdot u_k > v^{(k)} \cdot M_{\cdot,k} - \varepsilon^2\nu/6$ contradicting (A.8) This proves the claim. ∎

Intuition Now, we go back to the proof of Theorem 4.6. For the $w_1, w_2, \ldots, w_k$ defined in (15), the hypothesis of Claim (A.1) holds. So for each ℓ, there is some $\ell' \in [k]$ with $w_{\ell'} \in Q_\ell$. Renumber and assume $w_\ell \in Q_\ell$.

Lemma (A.2) proves that any $w_\ell \in Q_\ell$, is close to $M_{\cdot,\ell}$. This is done as follows: By definition of Q_ℓ, $w_\ell \in Q_\ell$ implies that w_ℓ is close to some point, say, $w'_\ell \in K$. w'_ℓ is a convex combination of $M_{\cdot,1}, M_{\cdot,2}, \ldots, M_{\cdot,k}$. If the convex combination did not attach weight almost 1 to $M_{\cdot,\ell}$, then, it attaches non-trivial total weight to the other $M_{\cdot,\ell'}$. Then, since the $M_{\cdot,\ell'}, \ell' \neq \ell$ are well-separated from $M_{\cdot,\ell}$ in the $v^{(\ell)}$ direction, (cf. (A.5)), namely, $M_{\cdot,\ell'} \notin \{x : v^{(\ell)} \cdot x \geq v^{(\ell)} \cdot M_{\cdot,\ell} - 2\varepsilon\nu\}$, we will have that w'_ℓ and hence also w_ℓ is far from $M_{\cdot,\ell}$ in the $v^{(\ell)}$ direction, contradicting the definition of Q_ℓ.

Lemma A.2 *$w_\ell \in Q_\ell \cap U$ implies*

$$||M_{\cdot,\ell} - w_\ell|| \leq \varepsilon\nu, \textit{ for } \ell = 1, 2, \ldots, k.$$

Proof: By (9), there is a $w'_\ell \in K$ such that $||w_\ell - w'_\ell|| \leq \varepsilon^2\nu/12$. Since $w'_\ell \in K$, it can be written as a convex combination of $M_{\cdot,1}, M_{\cdot,2}, \ldots M_{\cdot,k}$, say,

$$\begin{aligned} &w'_\ell = \alpha_\ell M_{\cdot,\ell} + (1 - \alpha_\ell)x, \\ &\text{where, } x \in CH(M_{\cdot,\ell'}, \ell' \neq \ell)\ ;\ \alpha_\ell \in [0, 1]. \end{aligned} \tag{A.9}$$

$$v^{(\ell)} \cdot x < v^{(\ell)} \cdot M_{\cdot,\ell} - 2\varepsilon\nu \text{ by (A.5)}\ .$$

So,

$$v^{(\ell)} \cdot w'_\ell < v^{(\ell)} \cdot M_{\cdot,\ell} - (1 - \alpha_\ell)2\varepsilon\nu.$$

Since $||v^{(\ell)}||_\infty = 1$,

$$v^{(\ell)} \cdot w_\ell \leq v^{(\ell)} \cdot M_{\cdot,\ell} - (1 - \alpha_\ell)2\varepsilon\nu + \varepsilon^2\nu/12.$$

But since $w_\ell \in Q_\ell$, we have by the definition of Q_ℓ

$$v^{(\ell)} \cdot w_\ell > v^{(\ell)} \cdot M_{\cdot,\ell} - 5\varepsilon^2\nu/12.$$

Thus by the last two inequalities,

$$(1 - \alpha_\ell) < \frac{\varepsilon}{4}.$$

So, since $\nu = \max_{\ell,\ell'} ||M_{\cdot,\ell} - M_{\cdot,\ell'}||$, by (A.9),

$$\begin{aligned} &||w'_\ell - M_{\cdot,\ell}|| = ||(1 - \alpha_\ell)(x - M_{\cdot,\ell})|| \\ &\leq (1 - \alpha_\ell)\text{Max}_{\ell,\ell'}||M_{\cdot,\ell} - M_{\cdot,\ell'}|| \leq \varepsilon\nu/4. \end{aligned}$$

$$||w_\ell - M_{\cdot,\ell}|| < ||w'_\ell - M_{\cdot,\ell}|| + \varepsilon^2\nu/4 \leq \varepsilon\nu$$

proving Lemma (A.2) as well as Theorem 4.6. ∎

∎

Proof: (Of Corollary 4.3) By hypothesis of the Corollary and Lemma 4.4, we have for all $R, |R| = \gamma n$ and all $v \in \{-1.1\}^d$, and for all $\lambda > 0$:

$$\Pr(v \cdot (A_{\cdot,R} - P_{\cdot,R}) \geq \lambda) \leq c\exp(-c\beta\lambda^2),$$

from which (6) follows. The other hypotheses of Theorem 4.1 follow directly from the hypotheses of Corollary 4.3. So Theorem 4.1 implies the corollary. ∎

B. Proofs for Section 5

Proof: (Of Lemma 5.2): If $x \in CH(M_{\cdot,\ell'} : \ell' \neq \ell)$, then, $\sum_{i \in T_\ell} x_i < \sum_{i \in T_\ell} M_{i,\ell} - 2\varepsilon$ which implies, that $||x - M_{\cdot,\ell}|| \geq 2\varepsilon$ as required. ∎

The proof above is technically straightforward. But note that we did not require the T_ℓ to be disjoint. This allows for example the case when $k > d$ which is ruled out if one assumes at the outset that the T_ℓ are disjoint (as is done in the literature, for example, (Bansal et al., 2014) and (Arora et al., 2018)). Also, we do not need another assumption prevalent in the literature, namely, that there be one (or more) individually high frequency words for each topic.

Intuition Next, we prove Lemma 5.3. It asserts that under Dirichlet distribution, with small concentration parameter (namely, $\alpha = 1/k$, which is a standard value), there is substantial prior mass near the corners. I.e., there is substantial probability that a document be nearly purely on a single topic. This type of fact is well-known, for example, in (Telgarsky, 2013). We supply the short proof of the exact result we need here for completeness. Note that in contrast, if we had a uniform prior on the simplex Δ_k, then, the mass near the vertices can be (exponentially in k) small.

Proof: (Of Lemma 5.3): For a random variable x distributed according to $\mathrm{Dir}(k, 1/k)$, the marginal density of x_1 is given by

$$q(x_1) = \frac{1}{\Gamma(1/k)\Gamma(2-(1/k))} x_1^{(1/k)-1}(1-x_1)^{1-(1/k)}.$$

We have

$$\Pr(x_1 \geq 1 - (\varepsilon^2/12)) = \int_{x_1=1-\varepsilon^2/12}^{1} q(x_1)dx_1$$
$$\geq \frac{c}{k}\int_{1-\varepsilon^2/12}^{1}(1-x_1)^{1-(1/k)} \geq \frac{c\varepsilon^4}{k}.$$

■

Next, we prove Lemma 5.4 which asserts that LDA satisfies the Sub-Gaussian assumption (6). For this, we will use the fact that for any $R, |R| = \gamma n$, and any $\{-1, 1\}$ vector $v \in \mathbf{R}^d$, $\sum_{j\in R}(v\cdot(A_{\cdot,j} - P_{\cdot,j}))$ is a function of γnm independent random variables (namely, the word choices for all documents in R). We remark that we do not know a proof of this type of concentration based only on viewing each document as a vector-valued random variable with the appropriate moment bounds, since, we can only upper bound up to the m th moment of individual $A_{\cdot,j} - P_{\cdot,j}$.

Proof: (Of Lemma 5.4): Let

$$f(R, v) = \sum_i \sum_{j\in R} v_i(A_{ij} - P_{ij}). \tag{B.10}$$

$f(R, v)$ is a function of γnm independent random variables, namely, the words in the γn documents in R.We note that changing any one word changes $f(R, v)$ by $1/m$ at most. So from the bounded difference inequality (McDiarmid & Reed, 2006), we get that

$$\Pr(f(R, v) \geq \varepsilon\gamma n) \leq 2\exp(-2\varepsilon^2\gamma nm).$$

(6) follows with $\beta = m$ as claimed. ■

C. Proofs for Section 6

Now, we prove Theorem 6.1.

Lemma C.1 *For any positive integer $d \geq 6$, there is a family $\mathcal{L}$ of subsets of $[d]$ so that*

$$\forall L \in \mathcal{L}, |L| = \frac{d}{2}$$
$$\forall L \neq L' \in \mathcal{L}, |L \cap L'| \leq \frac{3d}{8}$$
$$|\mathcal{L}| \geq (1.1)^d.$$

Proof: We choose sets in $\mathcal{L}$ one by one. Each set we decide to put into $\mathcal{L}$ rules out all other $d/2$-sets with intersection greater than $3d/8$ with it, namely, each set rules out $\binom{d/2}{3d/8}\binom{d/2}{d/8}$ sets. It is as simple caculation using Striling inequalities for the factorial function to see that

$$(1.1)^d\binom{d/2}{3d/8}\binom{d/2}{d/8} < \binom{d}{d/2},$$

which proves the Lemma. ■

For each $L \in \mathcal{L}$, we define a vector $v(L) \in \mathbf{R}^d$ with $||v(L)||_1 = 1$ as follows:

$$v(L)_i = \begin{cases} 0 & \text{if } i \notin L \\ \frac{2}{d} & \text{if } i \in L. \end{cases}$$

Lemma C.2 *For each $L \in \mathcal{L}$, we have*

$$\mathit{Dist}_1\big(v(L)\ ,\ CH\big(v(L') : L' \neq L, L' \in \mathcal{L}\big)\big) \ \geq\ \frac{1}{4}.$$

Proof: By Lemma (C.1) and the construction of $v(L)$, we have for every $L' \neq L$,

$$\sum_{i\notin L} v(L') \geq \frac{d}{8}\frac{2}{d} = \frac{1}{4}.$$

So, $\sum_{i\notin L} x_i = 1/4$ is a separating hyperplane between the convex sets: $CH\big(v(L') : L' \neq L, L' \in \mathcal{L}\big)$ and $v(L) + \frac{1}{4}B_1$ proving the Lemma. ■

Proof: (Of Theorem 6.1): Let $\mathcal{M}$ be a family of $d \times k$ matrices, where each $\mathbf{M} \in \mathcal{M}$ is obtained by choosing a $k-$subset of $\{v(L) : L \in \mathcal{L}\}$. We note that

$$|\mathcal{M}| \geq \frac{1}{k!}(1.1)^{dk},$$

the $1/k!$ accounting for the over-counting the $k!$ permutations of the columns of $\mathbf{M}$. Also, by Lemma (C.2), each $\mathbf{M} \in \mathcal{M}$ satisfies Assumption 2.

Since each document has m words, there are at most d^{mn} possible sets of n documents. Under the hypothesis of the Theorem, $d^{mn} < 0.5|\mathcal{M}|$. Since the learner is deterministic, each input generates a unique output. Thus, there is at least one $\mathbf{M} \in \mathcal{M}$ so that no input leads to an output of any $\mathbf{M}'$ with $\text{Dist}(\mathbf{M}, \mathbf{M}') \leq \varepsilon$. ∎

Remark C.1 *The reader may be momentarily puzzled about such a class of specially structured* $\mathbf{M}$ *being intractable. Note that for documents which are purely on a topic with topic vector* $v(L), L \in \mathcal{L}$*, indeed, they will have no words at all from* $[d] \setminus L$*. But the reason we cannot still decipher* l *or* $v(L)$ *from the documents is the following: there are* n/k *documents with a total of* nm/k *words in them. Since Theorem 6.1 assumes that* $nm < 0.09dk$*, the total number of words in all these documents is less than* $0.09d$*. So many words* $i \in L$ *also never appear in these documents. This is intuitively the reason for the intractability.*

D. Proofs for Section 7

Proof: (Of Theorem 7.1) First we prove that (4) holds. For any $S \subseteq [n]$, $|S| = \gamma n$,

$$||A_{\cdot,S} - P_{\cdot,S}||_1 = \frac{1}{\gamma n}\text{Max}_{v\in\{-1,+1\}^d} \underbrace{\sum_{i=1}^{d}\sum_{j\in S} v_i(A_{ij} - P_{ij})}_{f(S,v)}.$$

We prove that with high probability each $f(S, v)$ cannot be too large and then take the union bound over the 2^d v 's. For a fixed v, we note that $f(S, v)$ is the sum of γnd independent random variables $v_i(A_{ij} - P_{ij})$, where each $|v_i(A_{ij} - P_{ij})| \leq 1$ and $\text{Var}(v_i(A_{ij} - P_{ij})) = P_{ij}(1 - P_{ij})$. We use Freedman's inequality to get that for any $\lambda > 0$,:

$$\begin{aligned}&\Pr\left(f(S, v) \geq \lambda\right)\\ &\leq \exp\left(-\lambda^2/(2(\lambda + \sum_i \sum_{j\in S} \text{Var}(v_i(A_{ij} - P_{ij}))))\right)\\ &\leq \exp\left(-\frac{\lambda^2}{2(\lambda + \gamma n\nu)}\right),\end{aligned}$$

since, for each j, $\sum_i P_{ij} = \sum_\ell W_{\ell,j} \sum_i M_{i,\ell} \leq \nu$.

$$\begin{aligned}&\Pr\left(\exists v : f(S, v) > \varepsilon\nu\gamma n\right)\\ &\leq \exp\left(d - \frac{\varepsilon^2\nu\gamma n}{6}\right),\end{aligned}$$

using the hypothesis of the Theorem. This proves 4 with $\beta = \nu$.

The proofs of (3) and (2) are identical to the case of LDA, so we do not repeat them here.

∎

Proof: (Of Theorem 7.2): As in proof of Theorem 6.1, we construct a set of $(1.1)^d$ vectors $v(L)$ with one difference: each non-zero component of $v(L)$ is set to $2\nu/d$ (rather than $2/d$). We still have that each $v(L)$ is at L_1 distance at least $\varepsilon\nu$ from the convex hull of the other $v(L')$ and so for any $k-$ subset of the $v(L)$, the Separation Assumption 2 is satisfied.

Now the argument about the number of possible data sets is different. The expected number of 1's in all columns of $\mathbf{A}$ is νn and with very high probability, the actual number is at most $2\nu n$. There are $2\nu n\binom{nd}{2\nu n}$ possible data sets with a total of at most $2\nu n$ 1's. ∎

References

Arora, S., Ge, R., Halpern, Y., Mimno, D. M., Moitra, A., Sontag, D., Wu, Y., and Zhu, M. Learning topic models - provably and efficiently. *Commun. ACM*, 61(4):85–93, 2018.

Bansal, T., Bhattacharyya, C., and Kannan, R. A provable SVD-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems 27*, pp. 1997–2005, 2014.

McDiarmid, C. and Reed, B. A. Concentration for self-bounding functions and an inequality of talagrand. *Random Struct. Algorithms*, 29(4):549–557, 2006.

Telgarsky, M. Dirichlet draws are sparse with high probability. *CoRR*, abs/1301.4917, 2013.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.