Fitting Topic-Rich models to a Billion Token corpus in a box
Supplementary Material

For $\zeta \in \{0, 1, 2, \ldots, m\}$, let $p_i(\zeta, l)$ be the probability that a random j belongs to T_l and $A_{ij} = \zeta/m$ and $q_i(\zeta, l)$ the corresponding "empirical probability":

$$p_i(\zeta, l) = \frac{1}{s} \sum_{j \in T_l} \binom{m}{\zeta} P_{ij}^{\zeta} (1 - P_{ij})^{m-\zeta}. \tag{1}$$

$$q_i(\zeta, l) = \frac{1}{s} \left| \{ j \in T_l : A_{ij} = \zeta/m \} \right|. \tag{2}$$

Note In the first step of the algorithm, we will pick (uniformly at) random subset of r documents, where, $r < s$. r will be large enough that we will assume that

(i) there are $w_l r$ documents in the subset with dominant topic l and (ii) p_i, q_i defined on just the subset (with r instead of s in the denominator) are the same as for the whole set of s documents.

The errors involved in these assumptions are small and can be ignored.

Note that $p_i(\zeta, l)$ is a real number, whereas, $q_i(\zeta, l)$ is a random variable with

$$E(q_i(\zeta, l) \mid \mathbf{P}) = p_i(\zeta, l).$$

For an interval $I \subset [0, m]$, we let

$$p_i(I, l) = \sum_{\zeta \in I} p_i(\zeta, l) \; ; \; q_i(I, l) = \sum_{\zeta \in I} q_i(\zeta, l). \tag{3}$$

We need a technical assumption on the $p_i(\zeta, l)$ (which is weaker than unimodality).

No-Local-Min Assumption We assume that $p_i(\zeta, l)$ does not have a local minimum, in the sense:

$$p_i(\zeta, l) \;>\; \text{Min}(p_i(\zeta-1, l), p_i(\zeta+1, l)) \;\forall\, \zeta \in \{1, 2, \ldots, m-1\}. \qquad (4)$$

The plot of $q_i(\zeta, l)$ versus ζ often has a Zipf's law behavior whence it is monotone decreasing. Or it could increase to a mode and fall (for catchwords). Both satisfy the assumption.

c refers to a generic constant independent of $m, s, 1/w_0, \varepsilon, \delta$; its value may be different in different contexts.

1 Proof of Correctness

We start by recalling the Höffding-Chernoff (H-C) inequality in the form we use it.

Lemma 1. Höffding-Chernoff *If X is the average of r independent random variables with values in $[0, 1]$ and $E(X) = \mu$, then, for any $t > 0$,*

$$Pr(X \geq \mu + t) \leq \exp\left(-\frac{t^2 r}{2(\mu + t)}\right) \;;\; Pr(X \leq \mu - t) \leq \exp\left(-\frac{t^2 r}{2\mu}\right).$$

1.1 General results

The first lemma is a consequence of the no-local-minimum assumption. We use that assumption solely through this Lemma.

Lemma 2. *Suppose a, b are integers with $0 \leq a \leq b \leq m$ and let $I = [a, b]$. We have*

$$p_i([a, b], l) \geq \frac{b - a + 1}{m + 1} Min\left(p_i([0, b], l), p_i([a, m], l)\right).$$

Proof. Abbreviate $p_i(\cdot, l)$ by $f(\cdot)$. It is easy to see that by the No-Local-Min property (4), for $\zeta_0 = \text{Argmax}_\zeta f(\zeta)$, we have

$$\begin{aligned} f(\zeta) &\geq f(\zeta - 1) && \text{for } \zeta = 1, 2, \ldots, \zeta_0 \\ f(\zeta) &\leq f(\zeta - 1) && \text{for } \zeta = \zeta_0 + 1, \zeta_0 + 2, \ldots, m. \end{aligned}$$

Now, let $f([a, \zeta_0) = x; f([\zeta_0 + 1, b]) = y; f([0, a - 1]) = u; f([b + 1, m]) = v$.

Case 1 $\zeta_0 \in [a, b]$: We have:

$$\begin{aligned}
x &\geq \frac{\zeta_0 - a + 1}{a} u \geq \frac{\zeta_0 - a + 1}{m - b + a} u \\
y &\geq \frac{b - \zeta_0}{m - b} v \geq \frac{b - \zeta_0}{m - b + a} v \\
x + y &\geq \frac{b - a + 1}{m - b + a} \min(u, v) \\
x + y &\geq \frac{1}{1 + \frac{m-b+a}{b-a+1}} \min(u + x + y, v + x + y),
\end{aligned}$$

from which we get the Lemma for this case. The other cases are easier and we omit the proofs. □

Next, we state a technical Lemma which is used repeatedly. It states that for every i, ζ, l, the empirical probability that $A_{ij} = \zeta/m$ is close to the true probability, even when conditioned on any value of $\mathbf{P}$. Unsurprisingly, we prove it using H-C. But we will state a consequence in the form we need in the sequel.

Lemma 3. *Let $I \subseteq [0, m]$ be an interval and $L \subseteq \{1, 2, \dots, k\}$. With probability at least $1 - 2\exp(-c\varepsilon w_0 s)$, we have*

$$0.9 \sum_{l \in L} p_i(I, l) - \frac{\varepsilon w_0}{4} \leq \sum_{l \in L} q_i(I, l) \leq 2 \sum_{l \in L} p_i(I, l) + \frac{\varepsilon w_0}{4}.$$

Proof. Note that

$$\sum_{l \in L} q_i(\zeta, l) = \frac{1}{s} \left| \{ j \in \cup_L T_l : A_{ij} = \zeta/m \} \right| = \frac{1}{s} \sum_{j=1}^{s} X_{ij},$$

where, X_{ij} is the indicator variable of $A_{ij} = \zeta/m \wedge j \in \cup_L T_l$. Now, (recalling the bound on the perturbation allowed in $\mathbf{P}$)

$$E(X_{ij}) = \frac{1}{s} \sum_{j \in T_l} (\mathbf{MW})_{ij} \quad \text{and} \quad |X_{ij} - (\mathbf{MW})_{ij}| \leq \frac{\varepsilon w_0}{8}.$$

We can apply H-C with $t = \mu + \frac{\varepsilon w_0}{4}$ and $\mu = \sum_L p_i(\zeta, l)$ to get

$$\begin{aligned}
&\Pr\Big(\sum_{l \in L} q_i(I, l) > 2 \sum_{l \in L} p_i(I, l) + \frac{\varepsilon w_0}{4}\Big) \\
&\leq \exp\Big(-\big(\mu + \frac{\varepsilon w_0}{4}\big)^2 s / 2\big(2\mu + \frac{\varepsilon w_0}{4}\big)\Big).
\end{aligned}$$

The last expression (viewed as a function of μ) is maximized when $\mu = 0$ and so we get an upper bound of $\exp(-\varepsilon w_0 s/8)$.

For the other side, H-C implies

$$\begin{aligned}
&\Pr(\sum_L q_i(I,l) < 0.9 \sum_L p_i(I,l) - \frac{\varepsilon w_0}{4}) \\
&\leq \exp(-(0.1 \sum_L p_i(I,l) + \frac{\varepsilon w_0}{4})^2 s \;/\; 2 \sum_L p_i(I,l)) \\
&\leq \exp(-0.05 \varepsilon w_0 s).
\end{aligned}$$

$\square$

1.1.1 Properties of Thresholding

Say that a threshold ζ_i "splits" $T_l^{(2)}$ if $T_l^{(2)}$ has a significant number of j with $A_{ij} > \zeta_i/m$ and also a significant number of j with $A_{ij} \leq \zeta_i/m$. Intuitively, it would be desirable if no threshold splits any T_l, so that, in $\mathbf{B}$, for each i, l, either most $j \in T_l^{(2)}$ have $B_{ij} = 0$ or most $j \in T_l^{(2)}$ have $B_{ij} = \sqrt{\zeta_i}$. We now prove that this is indeed the case with proper bounds. We henceforth refer to the conclusion of the Lemma below by the mnemonic "no threshold splits any T_l".

Lemma 4. *(***No Threshold Splits any** T_l*) For a fixed i, l, with probability at least $1 - m^2 \exp(-c\varepsilon w_0 r)$, the following holds:*

$$\mathit{Min}\ (p_i([0,\zeta_i],l)\ ,\ p_i([\zeta_i+1,m],l)) \quad \leq 4\varepsilon w_0/\varepsilon_0.$$

Proof. Note that ζ_i is a random variable which depends only on $A^{(1)}$. So, for $j \in T_l^{(2)}$, A_{ij} are independent of ζ_i. Now, suppose

$$p_i([0,\zeta_i],l) > \frac{4\varepsilon w_0}{\varepsilon_0} \text{ and } p_i([\zeta_i+1,m],l) > \frac{4\varepsilon w_0}{\varepsilon_0}.$$

Let

$$I = \left[\mathrm{Max}(0, \frac{\zeta_i}{m} - \varepsilon_0), \mathrm{Min}(m, \frac{\zeta_i}{m} + \varepsilon_0)\right].$$

Since $\varepsilon_0 m$ is an integer, we can write I as $[\frac{a}{m}, \frac{b}{m}]$ and apply Lemma (2) to get:

$$p_i(I,l) > 4\varepsilon w_0.$$

Pay a failure probability of $m^2 \exp(-c\varepsilon r w_0)$ and assume the conclusion of Lemma (3) holds for every interval $I \subseteq [0, m]$. [Note: The Lemma was for the case when the empirical probability $q_i(I, l)$ was for a sample of s documents, but is valid for any s. Here we apply it with r samples instead of s, since $\mathbf{A}^{(1)}$ has just r columns. - Recall we assumed these quantities are the same for the sub-sample of r documents as well - see note just after (2).] We have:

$$\frac{1}{r}\left|\{j \in T_l^{(1)} : A_{ij} \in I\}\right| = q_i(I,l) \geq 0.9 p_i(I,l) - \frac{\varepsilon w_0}{4} > 3\varepsilon w_0,$$

contradicting the definition of ζ_i in the algorithm. This completes the proof of the Lemma. □

Define k vectors $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(k)}$ by

$$\mu^{(l)} = E(B_{\cdot,j} \mid j \in T_l^{(2)}), l = 1, 2, \ldots, k,$$

where, expectation refers to uniform random sample j.

We also abuse notation slightly and let $\mu_{\cdot,j}$ denote $\mu^{(l)}$ for $j \in T_l^{(2)}$, so we also think of μ as a $d \times s$ matrix. The entries of the matrix μ are fixed (real numbers) once we have $\mathbf{A}^{(1)}$ (and the thresholds ζ_i are determined). But μ_{ij} are random variables before we fix $\mathbf{A}^{(1)}$. The following Lemma is a direct consequence of "no threshold splits any T_l".

Lemma 5. *Suppose $\zeta_i \geq 8\ln(20/\varepsilon w_0)$. With probability at least $1 - 2m^2 kd \exp(-c\varepsilon r w_0)$ (over the choice of $\mathbf{A}^{(1)}$):*

$$\forall l, \forall j \in T_l, \forall i : \mu_{ij} \leq \varepsilon_l \sqrt{\zeta_i} \text{ OR } \mu_{ij} \geq \sqrt{\zeta_i}(1 - \varepsilon_l)$$

where, $\varepsilon_l = 4\varepsilon w_0 / \varepsilon_0 w_l$.

Proof. After paying a failure probability of $2m^2 kd \exp(-c\varepsilon r w_0)$, assume no threshold splits any T_l. [The factors of k and d come in because we are taking the union bound over all words and all topics.] Then,

$$p_i([0, \zeta_i], l) \leq 4\varepsilon \frac{w_0}{\varepsilon_0} = .4\varepsilon_l w_l$$
$$\text{or } p_i([\zeta_i + 1, m], l) \leq 4\varepsilon \frac{w_0}{\varepsilon_0} = .4\varepsilon_l w_l.$$

Wlg, assume that the first inequality holds. Then, by Lemma (3),

$$\begin{aligned} q_i([0,\zeta_i],l) &\leq .8\varepsilon_l w_l + \varepsilon w_0/4 \leq \varepsilon_l w_l \sqrt{\zeta_i} \\ \frac{1}{w_l s}\sum_{j\in T_l} B_{ij} &\geq \frac{w_l s(1-\varepsilon_l)}{w_l s} = (1-\varepsilon_l)\sqrt{\zeta_i} \end{aligned} \tag{5}$$

which implies

$$\mu_{ij} \geq (1-\varepsilon_l)\sqrt{\zeta_i}.$$

This proves the lemma in this case. The other case is symmetric. □

So far, we have proved that for every i, the threshold does not split any T_l. But this is not sufficient in itself to be able to cluster (and hence identify the T_l), since, for example, this alone does not rule out the extreme cases that for most j in every T_l, $A^{(2)}_{ij}$ is above the threshold (whence $\mu_{ij} \geq (1-\varepsilon_l)\sqrt{\zeta'_l}$ for almost all j) or for most j in no T_l is $A^{(2)}_{ij}$ above the threshold, whence, $\mu_{ij} \leq \varepsilon_l\sqrt{\zeta'_i}$ for almost all j. Both these extreme cases would make us loose all the information about T_l due to thresholding; this scenario and milder versions of it have to be proven not to occur. We do this by considering how thresholds handle catchwords. Indeed we will show that for a catchword $i \in S_l$, each $j \in T_l$ has $A^{(2)}_{ij}$ above the threshold and each $j \notin T_l$ has $A^{(2)}_{ij}$ below the threshold. (Both statements will only hold with high probability, of course.) To do this, we first define a η_i (which is not random and only depends on parameters, not data) and show (Lemma (6)) that whp, for $j \in T_l$, $A_{ij} > \eta_i/m$ and for $j \notin T_l$,, $A_{ij} < \eta_i/m$. Then we show (Lemma (7)) that with high probability, $\zeta_i \geq \eta_i$. So it follows for $i \in S_l$, $j \notin T_l$, whp, $A_{ij} < \zeta_i/m$ and so $B_{ij} = 0$. Since ζ_i can be greater than η_i, it does not automatically follow that for $j \in T_l$, $A_{ij} > \zeta_i/m$. Since for $j \notin T_l$, $A_{ij} < \zeta_i/m$, and since by definition of ζ_i in the algorithm, we have at least $w_0 r/2$ documents j with $A_{ij} > \zeta_i/m$, most of these must be in T_l. Now, the "no threshold splits any T_l" lemma comes in handy to show that indeed, most of T_l lies above threshold. This is used in the proof of the Lemma that $\mu_{\cdot,j}$ and $\mu_{\cdot,j'}$ differ a lot for j, j' in different T_l.

Lemma 6. *For $i \in S_l$, and $l' \neq l$, we have with $\eta_i = \lfloor M^{(1)}_{il}(\alpha+\beta+\rho)m/2 \rfloor$,*

$$p_i([0,\eta_i+\varepsilon_0 m],l) \leq \varepsilon w_0 w_l/20,$$
$$p_i([\eta_i-\varepsilon_0 m,m],l') \leq \varepsilon w_0 w_l/20.$$

Proof. Recall that P_{ij} is the probability of word i in document j conditioned on $\mathbf{W}$. Fix an $i \in S_l$. From the dominant topic assumption,

$$\forall j \in T_l, (\mathbf{MW})_{ij} = \sum_{l_1} M^{(1)}_{il_1} W^{(1)}_{l_1,j} \geq M^{(1)}_{il} W^{(1)}_{lj} \geq M^{(1)}_{il}\alpha_l \implies P_{ij} \geq M^{(1)}_{il}\alpha - \varepsilon w_0/8. \tag{6}$$

Note that (6) holds with probability 1. From Catchword assumption we get that

$$M^{(1)}_{il}\alpha_l - (\eta_i/m) - (\varepsilon w_0/8) \geq M^{(1)}_{il}\alpha - M^{(1)}_{il}((\alpha+\beta+\rho)/2) - (\varepsilon w_0/8) \geq M^{(1)}_{il}\alpha\delta/2.$$

Now, we will apply H-C with $\mu - t = \varepsilon_0 + \eta_i/m$ and $\mu \geq M^{(1)}_{il}\alpha_l - (\varepsilon w_0/8)$ for the m independent words in a document. By Calculus, the probability bound from H-C of

$$\exp(-t^2 m/2\mu) = \exp(-(\mu - \varepsilon_0 - (\eta_i/m))^2 m/2\mu)$$

is highest subject to the constraints $\mu \geq M^{(1)}_{il}\alpha_l; \eta_i \leq mM^{(1)}_{il}(\alpha+\beta+\rho)/2$, when $\mu = M^{(1)}_{il}\alpha - (\varepsilon w_0/8)$ and $t = M^{(1)}_{il}\alpha_l - \frac{\eta_i}{m} - \varepsilon_0$, whence, we get

$$p_i([0, \eta_i + \varepsilon_0 m], l) \leq \exp(-M^{(1)}_{il}\alpha\delta^2 m/16) \leq \varepsilon w_0/20,$$

using (7). Now, we prove the second assertion of the Lemma.

$$\begin{aligned}
&\forall j \in T_{l'}, l' \neq l, \sum_{l_1} M^{(1)}_{il_1} W^{(1)}_{l_1,j} = M^{(1)}_{il} W^{(1)}_{lj} + \sum_{l_1 \neq l} M^{(1)}_{il_1} W^{(1)}_{l_1,j} \\
&\leq M^{(1)}_{il} W^{(1)}_{lj} + \left(\mathrm{Max}_{l_1 \neq l} M^{(1)}_{il_1}\right)(1 - W^{(1)}_{lj}) \\
&\leq M^{(1)}_{il}(\beta+\rho) \quad \implies \quad P_{ij} \leq M^{(1)}_{il}(\beta+\rho) + (\varepsilon w_0/8).
\end{aligned} \tag{7}$$

$$\frac{\eta_i}{m} - \varepsilon_0 - M^{(1)}_{il}(\beta+\rho) \geq \frac{M^{(1)}_{il}(\alpha+\beta+\rho)}{2} - M^{(1)}_{il}(\beta+\rho) - \frac{1}{m} - \varepsilon_0 \geq 0.4 M^{(1)}_{il}\alpha\delta,$$

using the bounds on α, β, ρ. Applying the first inequality of Lemma (1) with $\mu + t = \eta_i/m - \varepsilon_0$ and $\mu \leq M^{(1)}_{il}(\beta+\rho)$ and we get the second assertion of the Lemma. □

Lemma 7. *For $i \in S_l$, $Pr(\zeta_i < \eta_i) \leq 3km^2 e^{-c\varepsilon r w_0}$, with η_i as defined in Lemma 6.*

Proof. Let $I = [\frac{\eta_i}{m} - \varepsilon_0, \frac{\eta_i}{m} + \varepsilon_0]$. Fix attention on an $i \in S_l$. After paying the failure probability of $3m^2ke^{-c\varepsilon r w_0}$, assume the conclusions of Lemma (3) hold for all l and all intervals I. It suffices to show that

$$\left|\{j : A^{(1)}_{ij} > \eta_i/m\}\right| \geq \frac{w_0 r}{2} \quad , \quad \left|\{j : A^{(1)}_{ij} \in I\}\right| < 3w_0\varepsilon r,$$

since, η_i is an integer and ζ_i is the largest integer satisfying the inequalities. For the first statement, we have from Lemma 6 with $I' = [\eta_i+1, m]$: $p_i(I', l) \geq w_l(1 - (\varepsilon w_0/20w_l)) \geq 0.9w_l$. So,

$$|\{j : A^{(1)}_{ij} > \eta_i/m\}| \geq rq_i(I, l) \geq w_l r(.81 - (\varepsilon w_0/4)) \geq w_0 r/2.$$

The second statement is slightly more complicated. Using both the first and second assertions of Lemma 6, we get that for all l' (including $l' = l$), we have

$$p_i(I, l') \leq \varepsilon w_0 w_l/20 \implies \sum_{l'=1}^{k} p_i(I, l') \leq \varepsilon w_0/20.$$

Now, Lemma (3) implies

$$\left|\{j : A^{(1)}_{ij} \in I\}\right| = r\sum_{l'=1}^{k} q_i(I, l') \leq \left(\frac{\varepsilon w_0}{10} + \frac{\varepsilon w_0}{4}\right) r \leq \varepsilon w_0 r,$$

thus completing the proof. □

Lemma 8. *Define $I_l = \{i \in S_l : \zeta_i \geq \eta_i\}$. With probability at least $1 - 6m^2dk\exp(-c\varepsilon r)$, we have for all l,*

$$\sum_{i \in I_l} \zeta'_i \geq m\alpha p_0/4.$$

Proof. After paying the failure probability, we assume the conclusion of Lemma 3 holds for all i, ζ, l. Now, by Lemma 7, we have (with $\mathbf{1}$ denoting the indicator function)

$$E\left(\sum_{i \in S_l} M^{(1)}_{il} \mathbf{1}(\zeta_i < \eta_i)\right) \leq 3m^2k\exp(-\varepsilon r w_0/8)\sum_{i \in S_l} M^{(1)}_{il},$$

which using Markov inequality implies that with probability at least $1 - 6m^2k\exp(-c\varepsilon s w_0)$,

$$\sum_{i \in I_l} M^{(1)}_{il} \geq \frac{1}{2}\sum_{i \in S_l} M^{(1)}_{il} \geq p_l/2. \tag{8}$$

Note that no catchword has ζ_i' set to zero. So,

$$\sum_{i\in I_l}\zeta_i' = \sum_{i\in I_l}\zeta_i \geq \sum_{i\in I_l}\eta_i \geq \sum_{I_l} mM_{il}^{(1)}\alpha_l/2 \geq \alpha_l p_l m/4.$$

□

Lemma 9. *With probability at least* $1 - 8m^2dk\exp(-c\varepsilon w_0 r)$*, we have for* $l \neq l'$*,*

$$|\mu^{(l)} - \mu^{(l')}|^2 \geq \frac{2m}{9}\alpha p_0.$$

Proof. For this proof, i will denote an element of I_l. By Lemma 6,

$$\forall i \in I_l, l' \neq l, p_i([\zeta_i, m], l') \leq \frac{\varepsilon w_0 w_l}{20}. \tag{9}$$

This implies by Lemma 3,

$$\sum_{l'\neq l}\left|\{j \in T_{l'}^{(1)} : A_{ij}^{(1)} > \frac{\zeta_i}{m}\}\right| \leq \sum_{l'\neq l} r\frac{\varepsilon w_0}{10}w_{l'} + r\frac{\varepsilon w_0}{4} \leq \varepsilon w_0 r. \tag{10}$$

Now the definition of ζ_i in the algorithm implies that:

$$r\sum_{\zeta>\zeta_i} q_i(\zeta, l) = \left|\{j \in T_l^{(1)} : A_{ij} > \frac{\zeta_i}{m}\}\right| \geq \left(\frac{w_0}{2} - \varepsilon w_0\right) r \geq w_0 r/4.$$

So, by Lemma 3,

$$\begin{aligned} &p_i([\zeta_i + 1, m], l) \geq \frac{1}{2}q_i([\zeta_i + 1, m], l) - \frac{1}{4}\varepsilon w_0 \\ &\geq \frac{w_0}{8} - \frac{1}{4}\varepsilon w_0 \geq w_0/9, \end{aligned}$$

using (7). Next let $I = [\frac{\zeta_i}{m} - \varepsilon_0, \frac{\zeta_i}{m} + \varepsilon_0]$ and $\tilde{p} = p_i(I, l)$. Since $|\{j \in T_l^{(1)} : A_{ij} \in I\}| \leq 3\varepsilon w_0 r$, by the definition of ζ_i in the algorithm, we get from Lemma 3 again:

$$\tilde{p} \leq 2q_i(I, l) + \varepsilon w_0/4 \leq 7\varepsilon w_0. \tag{11}$$

Now, by Lemma 2, we have

$$\tilde{p} \geq \mathrm{Min}\left(\frac{2\varepsilon_0 w_0}{9}\ ,\ 2\varepsilon_0 p_i([0, \zeta_i], l)\right).$$

By (7), $7\varepsilon w_0 < 2\varepsilon_0 w_0/9$ and so $\tilde{p} < 2\varepsilon_0 w_0/9$ and we get:

$$p_i([0, \zeta_i], l) \leq 7\varepsilon w_0/2\varepsilon_0.$$

Noting that by (2,3,4), no catchword has ζ_i' set to zero, $\Pr(B_{ij} = 0 | j \in T_l^{(2)}) \leq 7\varepsilon w_0/2\varepsilon_0 w_l \leq 1/6$, by the bounds on ε. This implies

$$\mu_{ij} \geq \frac{5}{6}\sqrt{\zeta_i'}.$$

Now, by (9), we have for $j' \notin T_l$,

$$\mu_{ij'} \leq \sqrt{\zeta_i'}/6.$$

So, we have

$$\sum_{i \in I_l} (\mu_{ij} - \mu_{ij'})^2 \geq (4/9) \sum_{i \in I_l} \zeta_i'.$$

Similarly, we get $\sum_{i \in I_{l'}} (\mu_{ij} - \mu_{ij'})^2 \geq \frac{4}{9} \sum_{i \in I_{l'}} \zeta_i'$. Now Lemma (8) implies the current Lemma. $\square$

Lemma 10. *With probability at least* $1 - \exp(-c\varepsilon w_0 s)$*, we have*

$$||\mathbf{B}||_F^2 \geq \frac{sm\alpha p_0}{20}.$$

Proof. By Lemma (8),

$$E\left(|B_{\cdot,j}|^2 \mid j \in T_l\right) \geq \frac{1}{2} E(\sum_{i \in S_l} \zeta_i') \geq \frac{m\alpha p_0}{10}.$$

$$\text{So, } E(||\mathbf{B}||_F^2) \geq \frac{m\alpha p_0 s}{10}.$$

Now, $||\mathbf{B}||_F^2 = \sum_j |B_{\cdot,j}|^2$ is the sum of independent random variables $|B_{\cdot,j}|^2$ which are each at most $8km$ by Lemma (11) . So applying H-C to $|B_{\cdot,j}|^2/(8km)$, we get the current Lemma.

$\square$

Since with high probability, for all $i \in S_l$ and $j \in T_l$, $B_{ij} = \zeta_i'$ and also by the argument of Lemma (6), $\zeta_i' \geq mM_{il}^{(1)}\alpha/2$, we have whp for $j \in T_l$, $|B_{\cdot,j}|^2 \geq c\alpha p_0 m$ and so $\sum_{j \in T_l} |B_{\cdot,j}|^2| \geq csw_l p_0 \alpha m$.

Also, for any j, $\sum_{i: B_{ij} > 0} \zeta_i' \leq m$ and so $||\mathbf{B}||_F^2 \leq sm$.

Now also we have that for $i \in S_l, j \notin T_l$, $B_{ij} = 0$ whp. Further, for $i \in S_0$, $B_{ij}^2 \leq \lambda_i$ implies (recall the definition of p_0 from the Notation section) that whp $\sum_{i \in S_0} B_{ij}^2 \leq p_0 m..$ Thus, whp, $||\mathbf{B}||_F^2 \leq s(p_0 + \sum_{l'} w_{l'} p_{l'}.$

1.2 $k-$means find dominant topics

We need a piece of notation: For $t = 1, 2, \ldots, r$, if $B_{\cdot,j}, j \in T_l$ was picked to be the t th column of $\mathbf{C}$, we form a $d \times r$ matrix $\tilde{\mu}$ with $\tilde{\mu}_{\cdot,t} = \mu_{\cdot,j}$. We denote by $\tilde{T}_l$ the set of columns in T_l which were sampled and included in $\mathbf{C}$.

We first prove:

Theorem 1.1. *With probability at least* $1 - cm^2 dk \exp(-c\varepsilon w_0 r)$*, we have*

$$||\mathbf{C} - \tilde{\mu}||_F^2 \leq ck^3 \frac{\varepsilon w_0 m}{p_0 \alpha \varepsilon_0}\, r.$$

Proof.

$$\text{Let } \mathcal{E}_1 : \sum_{i=1}^{d} \zeta_i' \leq ckm \quad ; \quad \mathcal{E}_2 \qquad\qquad : ||\mathbf{B}||_F^2 \geq csm\delta_0. \tag{12}$$

After paying the failure probability of $m^2 dk \exp(-c\varepsilon w_0 r)$, we may assume from Lemmas (11) and (10), that $\mathcal{E}_1, \mathcal{E}_2$ hold.

Consider the random variable $X = ||\mathbf{C} - \tilde{\mu}||_F^2$. It is the sum of r independent i.i.d. random variables: $X_t = |C_{\cdot,t} - \tilde{\mu}_{\cdot,t}|^2$. Changing one $C_{\cdot,t}$ changes X by at most ckm since each $|B_{\cdot,j}|^2 \leq \sum_{i=1}^d \zeta_i'$ and under $\mathcal{E}_1$, $\sum_i \zeta_i' \leq ckm$. So we have by Bounded Difference Inequality that with high probability, $|X - EX|$ is small. So, now, it suffices to bound $E(X)$. Now,

$$E(X) = rE_{\text{length}^2}\left(|C_{\cdot,1} - \tilde{\mu}_{\cdot,1}|^2\right) = E\left(\sum_{j=1}^{s} \frac{|B_{\cdot,j}|^2}{||\mathbf{B}||_F^2} |B_{\cdot,j} - \mu_{\cdot,j}|^2\right). \tag{13}$$

$$E\left(\sum_{j=1}^{s} \frac{|B_{\cdot,j}|^2}{||\mathbf{B}||_F^2} |B_{\cdot,j} - \mu_{\cdot,j}|^2\right) \leq E\left(\sum_{j=1}^{s} \frac{|B_{\cdot,j}|^2}{||\mathbf{B}||_F^2} |B_{\cdot,j} - \mu_{\cdot,j}|^2 \mid \mathcal{E}_1, \mathcal{E}_2\right) + m^2 dk \exp(-c\varepsilon w_0 r) m^2,$$

where, for the second term, we have used $|B_{\cdot,j}|^2 \leq ||\mathbf{B}||_F^2$ and $|B_{\cdot,j}|^2, |\mu^{(l)}|^2 \leq m^2$. The second term is easily seen to be lower order, so we may ignore it and just bound the first term. Now since $|B_{\cdot,j}|^2 \leq \sum_{i=1}^d \zeta_i'$,

$$E\left(\sum_{j=1}^{s} \frac{|B_{\cdot,j}|^2}{||\mathbf{B}||_F^2}|B_{\cdot,j} - \mu_{\cdot,j}|^2 \mid \mathcal{E}_1, \mathcal{E}_2\right) \leq \frac{ckm}{sm\alpha p_0} E\left(\sum_{j=1}^{s} |B_{\cdot,j} - \mu_{\cdot,j}|^2\right)$$
$$\leq \frac{ck}{s\delta_0}\sum_{l=1}^{k} w_l s \; E\left(|B_{\cdot,j} - \mu^{(l)}|^2 \mid j \in T_l\right)$$
$$\leq \frac{ck^2\varepsilon w_0}{\varepsilon_0\delta_0} E(\sum_i \zeta_i') \leq \frac{ck^3\varepsilon w_0 m}{\varepsilon_0\delta_0},$$

where, we have used Lemma (5) and $\mathcal{E}_1$. Since $|B_{\cdot,j}|^2 \leq ckm$ under $\mathcal{E}_1$ and $\mathcal{E}_2$, we can put in Theorem (1.4) $\nu \leq c\sqrt{km}$. Also put $t = ck\sqrt{\frac{\varepsilon w_0}{\delta_0\varepsilon_0}}\sqrt{r}$. Then the current theorem follows. □

1.2.1 Proximity

We need a piece of notation: For $t = 1, 2, \ldots, r$, if $B_{\cdot,j}, j \in T_l$ was picked to be the t th column of $\mathbf{C}$, we form a $d \times r$ matrix $\tilde{\mu}$ with $\tilde{\mu}_{\cdot,t} = \mu_{\cdot,j}$. We denote by $\tilde{T}_l$ the set of columns in T_l which were sampled and included in $\mathbf{C}$.

We wish to show that clustering as in ℓ_2^2 identifies the dominant topics correctly for most documents, i.e., that $R_l \approx \tilde{T}_l$ for all l. For this, we will use a theorem from [2] [see also [1]] which in this context says:

Theorem 1.2. *If all but a f fraction of the the $C_{\cdot,t}$ satisfy the "proximity condition", then ℓ_2^2TSVD identifies the dominant topic in all but $c_1 f$ fraction of the documents correctly after polynomial number of iterations.*

To describe the proximity condition, first let σ be the maximum over all directions v of the square root of the mean-squared distance of $C_{\cdot,t}$ to $\tilde{\mu}_{\cdot,t}$, i.e.,

$$\sigma^2 = \text{Max}_{\|v\|=1}\frac{1}{r}|v^T(\mathbf{C} - \tilde{\mu})|^2 = \frac{1}{r}\|\mathbf{C} - \tilde{\mu}\|^2.$$

The parameter σ should remind the reader of standard deviation.

Recall: We showed that $|\tilde{T}_l|$ is at least $\Omega(w_l\alpha_l p_l r/k)$.

Definition: $C_{\cdot,t}, t \in \tilde{T}_l$ is said to satisfy the proximity condition with respect to μ, if for each $l' \neq l$, the projection of $C_{\cdot,t}$ onto the line joining $\mu^{(l)}$ and $\mu^{(l')}$

is closer to $\mu^{(l)}$ than it is to $\mu^{(l')}$ by at least at least

$$\Delta_{l.l'} = c_0 k \left(\frac{\sqrt{r}}{\sqrt{|\tilde{T}_l|}} + \frac{\sqrt{r}}{\sqrt{|\tilde{T}_{l'}}} \right) \sigma$$
$$\leq c_0 k^{3/2} \left(\frac{1}{\sqrt{w_l \alpha_l p_l}} + \frac{1}{\sqrt{w_{l'} \alpha_{l'} p_{l'}}} \right) \sigma.$$

If this fails for an l', we say that t is not proximate with respect to l'.

To prove proximity, we need to upper bound σ. This will be the task of the subsection 1.3 which relies heavily on Random Matrix Theory.

1.3 Bounding the Spectral norm

In this section, we prove:

Theorem 1.3. *With δ_0 as in (6), we have: With probability at least $1 - cm^2 dk \exp(-c\varepsilon w_0 r)$, we have*

$$||\mathbf{C} - \tilde{\mu}||^2 \leq ck^3 \frac{\varepsilon w_0 m}{\delta_0 \varepsilon_0}\, r.$$

Theorem 1.4. *[3, Theorem 5.44] Suppose R is a $d \times r$ matrix with columns $R_{\cdot,j}$ which are independent identically distributed vector-valued random variables. Let $U = E(R_{\cdot,j} R_{\cdot,j}^T)$ be the inertial matrix of $R_{\cdot,j}$. Suppose $|R_{\cdot,j}| \leq \nu$ always. Then, for any $t > 0$, with probability at least $1 - de^{-ct^2}$, we have*

$$||R|| \leq ||U||^{1/2} \sqrt{r} + t\nu.$$

We need the following Lemma first.

Lemma 11. *Let*

$$\zeta_i' = \begin{cases} \zeta_i & \text{if } \zeta_i \geq 8 \ln(20/\varepsilon w_0)) \\ 0 & \text{if } \zeta_i < 8 \ln(20/\varepsilon w_0) \end{cases}.$$

Let $\zeta_0 = Max_i \zeta_i'$. With probability at least $1 - \exp(-r\varepsilon w_0/3)$, we have

$$\zeta_0 \leq 4m\lambda \ ; \ \sum_i \zeta_i' \leq 4km \tag{14}$$

Proof. The probability of word i in document j, is given by: $(\mathbf{MW})_{ij} = \sum_l M^{(1)}_{il} W^{(1)}_{lj} \leq \lambda_i$ (where, $\lambda_i = \max_l M^{(1)}_{il}$). If $\lambda_i < \frac{1}{m}\ln(20/\varepsilon w_0)$, then, $\Pr(A_{ij} > (8/m)\ln(20/\varepsilon w_0)) \leq \varepsilon w_0$ by H-C (since A_{ij} is the average of m i.i.d. trials). Let X_j be the indicator function of $A_{ij} > (8/m)\ln(20/\varepsilon w_0)$. X_j are independent and so using H-C, we see that with probability at least $1 - \exp(-\varepsilon w_0 r/3)$, less than $w_0 s/2$ of the A_{ij} are greater $(8/m)\ln(20/\varepsilon w_0)$, whence, $\zeta'_i = 0$. So we have (using the union bound over all words):

$$\Pr\left(\sum_{i:\lambda_i<(1/m)\ln(20/\varepsilon w_0)} \zeta'_i > 0\right) \leq d\exp(-\varepsilon w_0 s/3).$$

If $\lambda_i \geq (1/m)\ln(20/\varepsilon w_0)$, then

$$\Pr(A_{ij} > 4\lambda_i) \leq e^{-\lambda_i m} \leq \varepsilon w_0/2,$$

which implies by the same X_j kind of argument that with probability at least $1 - \exp(-\varepsilon w_0 r/4)$, for a fixed i, $\zeta_i \leq 4\lambda_i m$. Using the union bound over all words and adding all i, we get that with probability at least $1 - 2d\exp(-\varepsilon w_0 s/4)$,

$$\sum_i \zeta'_i \leq 4m\sum_i \lambda_i \leq 4m\sum_{i,l} M^{(1)}_{il} \leq 4km.$$

Now we prove the bound on ζ_0. For each fixed i, j, we have $\Pr(A_{ij} \geq 4\lambda) \leq e^{-m\lambda} \leq \varepsilon w_0$. Now, let Y_j be the indicator variable of $A_{ij} \geq 4\lambda$. The $Y_j, j = 1, 2, \ldots, s$ are independent (for each fixed i). So, $\Pr(\zeta_i \geq 4m\lambda) \leq \Pr(\sum_j Y_j \geq w_0 s/2) \leq e^{-\varepsilon w_o r/3}$. Using an union bound over all words, we get that $\Pr(\zeta_0 > 4m\lambda) \leq de^{-\varepsilon w_0 r/3}$ by H-C. $\square$

Proof. (of Theorem 1.3)

Let $U = E\left((C_{\cdot,1} - \tilde{\mu}_{\cdot,1})\,(C_{\cdot,1} - \tilde{\mu}_{\cdot,1})^T\right)$ be the intertial matrix of $C_{\cdot,1} - \tilde{\mu}_{\cdot,1}$.

$$\begin{aligned} ||U|| &\leq \mathrm{Max}_{v:|v|=1} E_{\text{length}^2}\left((v^T(C_{\cdot,1} - \tilde{\mu}_{\cdot,1}))^2\right) \\ &\leq E_{\text{length}^2}\left(|C_{\cdot,1} - \tilde{\mu}_{\cdot,1}|^2\right) = E\left(\sum_{j=1}^{s} \frac{|B_{\cdot,j}|^2}{||\mathbf{B}||_F^2} |B_{\cdot,j} - \mu_{\cdot,j}|^2\right). \end{aligned} \tag{15}$$

$$\text{Let } \mathcal{E}_1 : \sum_{i=1}^{d} \zeta'_i \leq ckm \quad ; \quad \mathcal{E}_2 \qquad : ||\mathbf{B}||_F^2 \geq csm\delta_0. \tag{16}$$

After paying the failure probability of $m^2dk\exp(-c\varepsilon w_0 r)$, we may assume from Lemmas (11) and (10), that $\mathcal{E}_1, \mathcal{E}_2$ hold. We use this to bound the right hand side of (15). To this end,

$$E\left(\sum_{j=1}^{s}\frac{|B_{\cdot,j}|^2}{||\mathbf{B}||_F^2}|B_{\cdot,j}-\mu_{\cdot,j}|^2\right)\leq E\left(\sum_{j=1}^{s}\frac{|B_{\cdot,j}|^2}{||\mathbf{B}||_F^2}|B_{\cdot,j}-\mu_{\cdot,j}|^2\mid\mathcal{E}_1,\mathcal{E}_2\right)+m^2dk\exp(-c\varepsilon w_0 r)m^2,$$

where, for the second term, we have used $|B_{\cdot,j}|^2\leq||\mathbf{B}||_F^2$ and $|B_{\cdot,j}|^2, |\mu^{(l)}|^2\leq m^2$. The second term is easily seen to be lower order, so we may ignore it and just bound the first term. Now since $|B_{\cdot,j}|^2\leq\sum_{i=1}^d\zeta_i'$,

$$E\left(\sum_{j=1}^{s}\frac{|B_{\cdot,j}|^2}{||\mathbf{B}||_F^2}|B_{\cdot,j}-\mu_{\cdot,j}|^2\mid\mathcal{E}_1,\mathcal{E}_2\right)\leq\frac{ckm}{sm\delta_0}E\left(\sum_{j=1}^{s}|B_{\cdot,j}-\mu_{\cdot,j}|^2\right)$$
$$\leq\frac{ck}{s\delta_0}\sum_{l=1}^{k}w_l s\, E\left(|B_{\cdot,j}-\mu^{(l)}|^2\mid j\in T_l\right)$$
$$\leq\frac{ck^2\varepsilon w_0}{\varepsilon_0\delta_0}E(\sum_i\zeta_i')\leq\frac{ck^3\varepsilon w_0 m}{\varepsilon_0\delta_0},$$

where, we have used Lemma (5) and $\mathcal{E}_1$. Since $|B_{\cdot,j}|^2\leq ckm$ under $\mathcal{E}_1$ and $\mathcal{E}_2$, we can put in Theorem (1.4) $\nu\leq c\sqrt{km}$. Also put $t=ck\sqrt{\frac{\varepsilon w_0}{\delta_0\varepsilon_0}}\sqrt{r}$. Then the current theorem follows. □

1.4 Proving Proximity

From Theorem (1.3), the σ in definition 1.4 is $ck^{3/2}\sqrt{\varepsilon w_0 m}/\sqrt{\delta_0\varepsilon_0}$. So, the Δ in definition 1.4 is

$$\Delta_{l.l'}\leq ck^3\sqrt{\frac{\varepsilon w_0 m}{\delta_0\varepsilon_0}}\left(\frac{1}{\sqrt{w_l\alpha_l p_l}}+\frac{1}{\sqrt{w_{l'}\alpha_{l'}p_{l'}}}\right).$$

So it suffices to prove:

Lemma 12. *For $t\in\tilde{T}_l$ and $l'\neq l$, let $\hat{C}_{\cdot,t}$ be the projection of $C_{\cdot,t}$ onto the line joining $\mu^{(l)}$ and $\mu^{(l')}$. The probability that $|\hat{C}_{\cdot,t}-\mu^{(l')}|\leq|\hat{C}_{\cdot,t}-\mu^{(l)}|+\Delta_{l,l'}$ is at most $c\varepsilon w_0 k^{5/2}/\delta_0\varepsilon_0\min_l\sqrt{\alpha_l p_l}$. Hence, with probability at least $1-cm^2dk\exp(-cw_0\varepsilon r)$, the number of t for which $C_{\cdot,t}$ does not satisfy the proximity condition is at most $\min_l(w_l a_l\delta_l)r/(10c_1)$, where, c_1 is the constant in Theorem (1.2) .*

Proof. After paying the failure probability of $cm^2dk\exp(-cw_0r\varepsilon)$, of Lemmas (11) and (9), assume that $\zeta_0 \leq 4m\lambda$, $|\mu_{.,j} - \mu_{.,j'}|^2 \geq (\alpha_l p_l + \alpha_{l'}p_{l'})m/9$ and $\sum_i \zeta'_i \leq 4km$.

For $j \in T_l$, define $X_{j,l'} = (B_{.,j} - \mu_{.,j}) \cdot (\mu^{(l')} - \mu^{(l)})$. Since $\Pr(B_{ij} = \sqrt{\zeta'_i} | j \in T_l) = \mu_{ij}/\sqrt{\zeta'_i}$, we have:

$$\begin{aligned}
E(|X_{j,l'}| \mid j \in T_l) &\leq E\sum_i |B_{ij} - \mu_{ij}|\, |\mu_i^{(l')} - \mu_i^{(l)}| \\
&= \sum_i \left[(\sqrt{\zeta'_i} - \mu_{ij}) \frac{\mu_{ij}}{\sqrt{\zeta'_i}} + (1 - \frac{\mu_{ij}}{\sqrt{\zeta'_i}}) \mu_{ij} \right] |\mu_{ij} - \mu_{ij'}| \\
&\leq 2\varepsilon_l \sum_i \sqrt{\zeta'_i} |\mu_{ij} - \mu_{ij'}| \quad \text{by Lemma 5} \\
&\leq 2\varepsilon_l \left(\sum_i \zeta'_i \right)^{1/2} |\mu_{.,j} - \mu_{.,j'}| \leq 4\varepsilon_l \sqrt{km} |\mu_{.,j} - \mu_{.,j'}|.
\end{aligned}$$

We claim that: If $|X_{j,l'}| \leq |\mu_{.,j} - \mu_{.,j'}|^2/8$, then, $|\hat{B}_{.,j} - \mu_{.,j'}| \geq |\hat{B}_{.,j} - \mu_{.,j}| + 3|\mu_{.,j} - \mu_{.,j'}|/4 \geq |\hat{B}_{.,j} - \mu_{.,j}| + \Delta_{l,l'}$.

To prove the claim, it suffices to show that $|\mu^{(l)} - \mu^{(l')}|^2 \geq 4\Delta^2_{l,l'}$. There are two cases: **Case 1** $w_l\alpha_l p_l \leq w_{l'}\alpha_{l'}p_{l'}$: Then we have $\left(\frac{1}{w_l\alpha_l p_l} + \frac{1}{w_{l'}\alpha_{l'}p_{l'}}\right) \leq \frac{2}{w_l\alpha_l p_l}$ and so $4\Delta^2_{l,l'} \leq cm\alpha_l p_l$, using (7). **Case 2** $w_l\alpha_l p_l > w_{l'}\alpha_{l'}p_{l'}$. By a similar argument, $4\Delta^2_{l,l'} \leq cm\alpha_{l'}p_{l'}$. Since $|\mu^{(l)} - \mu^{(l')}|^2 \geq cm(\alpha_l p_l + \alpha_{l'}p_{l'})$, the claim follows.

Let $Y_{j,l'}$ be the indicator of non-proximity of j for l'.

$$\Pr(Y_{j,l'} \mid j \in T_l) \leq \Pr(X_{j,l'} \geq (1/8)|\mu_{.,j} - \mu^{(l')}|^2) \leq \frac{c\varepsilon_l\sqrt{k}}{\sqrt{\alpha_l p_l}}.$$

Let Y_j = indicator of non-proximity of j. Union over all $l' \neq l$. Now, under $\mathcal{E}_1$ and $\mathcal{E}_2$ of (16),

$$E\left(\frac{|B_{.,j}|^2}{||\mathbf{B}||_F^2} Y_j \mid j \in T_l\right) \leq \frac{ck}{s\delta_0} \frac{\varepsilon_l k^{3/2}}{\sqrt{\alpha_l p_l}}.$$

$$\Pr(C_{.,1} \text{ doesn't satisfy proximity }) = E_{\text{length}^2}\left[\sum_{j=1}^{s} Y_j\right] \leq \frac{ck^{5/2}\varepsilon w_0}{\delta_0 \text{Min}_l \sqrt{\alpha_l p_l}\varepsilon_0}.$$

Now using H-C on the r independent columns of $\mathbf{C}$, and (7), the second statement of Lemma follows. ☐

The last Lemma implies by Theorem (1.2):

Lemma 13. *With probability at least* $1 - \exp(-cw_0 \varepsilon r)$, l_2^2 *TSVD correctly identifies the dominant topic in all but at most* $\min_l(w_l a_l)\delta/10$ *fraction of documents in each* $\tilde{T}_l$.

1.5 Identifying Catchwords

Recall the definition of J_l from Step 5a of the algorithm. The two lemmas below are roughly converses of each other which prove roughly that J_l consists of those i for which $M_{il}^{(1)}$ is strictly higher than $M_{il'}^{(1)}$.

Lemma 14. *Let* J_l *be as in step 6b of the Algorithm. For* $i \in J_l$, *and* $l' \neq l$, $M_{il}^{(1)} \geq (1+4\delta)M_{il'}^{(1)}$ *and* $M_{il}^{(1)} \geq \frac{3}{m\delta^2}\ln(20/\varepsilon\delta \min_l(w_l a_l p_l \alpha_l))$.

Proof. By the definition of J_l in the algorithm, $g(i,l) \geq (6/m\delta^2)\ln(20/\varepsilon\delta\min_l(w_l a_l p_l \alpha_l))$. We claim that this implies:

$$\max_{l_1} M_{il_1}^{(1)} \geq \frac{3}{m\delta^2}\ln(20/\varepsilon\delta\min_l(w_l a_l p_l \alpha_l)). \tag{17}$$

Suppose not. Then $(\mathbf{MW})_{ij} < \frac{3}{m\delta^2}\ln(20/\varepsilon\delta\min_l(w_l a_l p_l \alpha_l))$, and we have

$$\begin{aligned}&\Pr\left(A_{ij} \geq (4/m\delta^2)\ln(20/\varepsilon\delta w_0 a p_0 \alpha)\right) \leq \\ &\exp(-\ln(20/\varepsilon\delta w_0 a p_0 \alpha)/8\delta^2) \leq \varepsilon\delta a w_0 p_0 \alpha/20,\end{aligned}$$

using (6,7). Thus, $\Pr(g(i,l) \geq (4/m\delta^2)\ln(20/\varepsilon\delta\min_l(w_l a_l p_l \alpha_l))) \leq c\exp(-\varepsilon w_0 r)$, which is a contradiction, proving (17).

Let $l' = \arg\max_{l_1 \neq l} M_{il_1}^{(1)}$ and assume for contradiction that $M_{il}^{(1)} \leq (1+4\delta)M_{il'}^{(1)}$. Now, by Lemma, there are at least $cw_{l'}a_{l'}p_{l'}\alpha_{l'}r/k$ documents in $\mathbf{C}$ which are $(1-\delta)$-pure for topic l' and by Lemma (13), at least $cw_{l'}a_{l'}p_{l'}\alpha_{l'}r/k$ of these are in $R_{l'}$ Further, (17) implies that for $(1-\delta)$-pure documents in $T_{l'}$, whp, $A_{ij} \geq M_{il'}^{(1)}(1-2\delta)$. Thus,

$$g(i,l') \geq M_{il'}^{(1)}(1-2\delta). \tag{18}$$

On the other hand, we have for all l_1, $M_{il_1}^{(1)} \leq \max(M_{il}^{(1)}, M_{il'}^{(1)}) \leq (1+4\delta)M_{il'}^{(1)}$ and so we have $g(i,l) \leq M_{il'}^{(1)}(1+5\delta)$ which together with (18) contradicts the fact that i is in J_l.

□

Lemma 15. *If* $M_{il}^{(1)} \geq$ *Max* $\left(\frac{5}{m\delta^2}\ln(20/\varepsilon\delta\min_l(a_l w_l p_l \alpha_l)), Max_{l'\neq l}(1+12\delta)\ M_{il'}^{(1)}\right)$, *then, with probability at least* $1-\exp(-c\varepsilon w_0 r)$*, we have that* $i \in J_l$*. So,* $S_l \subseteq J_l$*.*

Proof. Using the pure documents for topic l and proceeding as in Lemma (14), we get:

$$g(i,l) \geq M_{il}^{(1)}(1-1.5\delta). \tag{19}$$

On the other hand, for $j \in T_{l'}$ and for $l' \neq l$ and $i : M_{il}^{(1)} \geq (1+12\delta)M_{il'}^{(1)}$ (hypothesis of the Lemma),

$$(\mathbf{MW})_{ij} \leq M_{il}^{(1)}W_{lj}^{(1)} + \frac{1}{1+12\delta}M_{il}^{(1)}(1-W_{lj}^{(1)}) \leq M_{il}^{(1)}\left(\beta_l + \frac{1-\beta_l}{1+12\delta}\right) \leq M_{il}^{(1)}\frac{1+1.2\delta}{1+12\delta},$$

since $\beta \leq 0.1$. So whp,

$$g(i,l') \leq M_{il}^{(1)}\frac{1+2\delta}{1+12\delta}. \tag{20}$$

From (19) and (20) and hypothesis of the Lemma, it follows that

$$g(i,l) \geq \text{Max}\left(\frac{4}{m\delta^2}\ln(1/\varepsilon w_0), (1+8\delta)\ g(i,l')\right).$$

So, $i \in J_l$ as claimed. It only remains to check that i in S_l satisfies the hypothesis of the Lemma which is obvious. □

Proof. (of Lemma 4): let $\hat{\mathbf{A}}$ be defined by $\hat{A}_{lj} = \sum_{i\in J_l} A_{ij}$, for $l = 1, 2, \ldots, k$ and $\hat{A}_{ij} = A_{ij}$ for $i \notin \cup_l J_l$ (except the rows are rearrnaged so that all $i \notin \cup J_l$ are put in rows $k+1, k+2, \ldots$). Similarly define $\hat{\mathbf{M}}$ and $\hat{\mathbf{P}}$. Call j with $W_{lj}^{(1)} \geq 1-\delta$ "pure" for topic l. For j pure for topic l, we have:

$$\hat{(\mathbf{MW})}_{lj} = \sum_{l'} \hat{M}_{ll'}^{(1)}W_{l'j}^{(1)} \geq \hat{M}_{ll}^{(1)}W_{lj}^{(1)} \geq (1-\delta)\hat{M}_{ll}^{(1)}.$$

Also, since each $\hat{A}_{lj}$ is the average of m independent trials, and $\hat{(\mathbf{MW})}_{lj} \leq M_{ll}^{(1)}$, we have by Höffding-Chernoff, for j pure for topic l:

$$\Pr\left(\hat{A}_{lj} \leq \hat{(\mathbf{MW})}_{lj} - \delta\hat{M}_{ll}^{(1)}\right) \leq ce^{-mc\delta^2\hat{M}_{ll}^{(1)}} \leq \varepsilon\delta/10,$$

since, $\hat{M}_{ll}^{(1)} \in \Omega^*(1/m\delta^2)$. This implies whp:

$$\left|\{j : W_{lj}^{(1)} \geq 1-\delta\ ;\ \hat{A}_{lj} \geq (1-2\delta)\hat{M}_{ll}^{(1)}\}\right| \geq \frac{3\varepsilon n}{4} \tag{21}$$

Now, consider j with $W_{lj}^{(1)} \leq 1-10\delta$. For such j,

$$\hat{(}\mathbf{MW})_{lj} = \hat{M}_{ll}^{(1)} + \sum_{l'\neq l} \hat{M}_{ll'}^{(1)} W_{l'j}^{(1)} \leq (1-10\delta)\hat{M}_{ll}^{(1)} + 10\delta(\hat{M}_{ll}^{(1)}/2) \leq (1-5\delta)\hat{M}_{ll}^{(1)}.$$

So for these j, we have

$$\Pr\left(\hat{A}_{lj} \geq (1-4\delta)\hat{M}_{ll}^{(1)}\right) \leq \varepsilon\delta/10,$$

which implies

$$\left|\{j : W_{lj}^{(1)} \leq 1-10\delta\ ;\ \hat{A}_{lj} \geq (1-4\delta)\hat{M}_{ll}^{(1)}\}\right| \leq \varepsilon\delta n/10. \tag{22}$$

This implies:

$$\left|U_l \cap \{j : W_{lj}^{(1)} \leq 1-10\delta\}\right| \leq \varepsilon\delta/5. \tag{23}$$

This can imply that at least for all $i \in J_l$, we have the desired inequality:

$$\widetilde{M}_{il}^{(1)} \geq (1-c\delta)M_{il}^{(1)}.$$

But we need this inequality for all i and for this, we proceed as follows.

Now, we go back to the original $\mathbf{A}, \mathbf{M}, \mathbf{P}$. Let

$$\frac{2}{\varepsilon n}\sum_{j\in U_l} A_{ij} = N_{il}.$$

Lemma 16. *Whp,* $\forall i, l : N_{il} \geq (1-14\delta)M_{il}^{(1)} - \frac{10\ln(d/\varepsilon)}{\varepsilon nm}$.

Proof. from (23), it follows that

$$\frac{2}{\varepsilon n}\sum_{j\in U_l}(\mathbf{MW})_{ij} \geq (1-10\delta)(1-(\delta/2))M_{il}^{(1)} \geq (1-12\delta)M_{il}^{(1)}.$$

Now, $\frac{2}{\varepsilon n}\sum_{j\in U_l} A_{ij}$ is the average of εnm independent Bernoulli random variables (where each is a word of a document in U_l). So by H-C, we can show for a single i,

$$\Pr\left(\frac{2}{\varepsilon n}\sum_{j\in U_l} A_{ij} < (1-14\delta)M_{il}^{(1)} - \frac{10\ln(d/\varepsilon)}{\varepsilon nm\delta}\right) \leq \frac{\varepsilon}{10d},$$

which implies by union bound that whp,

$$\sum_{j \in U_l} A_{ij} \geq (1 - 14\delta) M_{il}^{(1)} - \frac{10 \ln(d/\varepsilon)}{\varepsilon nm} \forall i,$$

proving the Lemma. □

Now the total amount we add to all the N_{il} is at most $10 \ln(d/\varepsilon)\ d\ /\varepsilon\delta nm$, which is at most δ by assumption on n. So (4) follows.

We now sketch the proof of (5).

$$M_{il}^{(1)} (\mathbf{M^{(2)}W})_{lj} \geq M_{il}^{(1)} \alpha \forall \text{ non-pure } j \in T_l.$$

Now, we can show that $(\mathbf{MW})_{i,j_0}$ is strictly lower for $j_0 \notin T_l$. This implies (by a calculation) that $l_1(j)$ is correct for all j. For the proof that $l_2(j)$ is correct, similar calculations are used.

[Note that we can easily extend this to more than 2 dominant basic topics per document.]

The run time complexity estimate is obtained through the following observations. Step (1) is dominated by (c) which has a runtime complexity of $O(nd)$ as it requires Thresholding all n documents. Between Step (2) is $O(rd)$ while Step (3) requires truncated SVD with complexity $O(rdk_0^2)$. Step (4) requires $O(dk_0^2)$.The remaining steps are all $O(nk)$, where k is the number of edge topics. All this together yields the desired estimate.

□

References

[1] Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

[2] Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Foundations of Computer Science (FOCS)*, pages 299–308. IEEE, 2010.

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