B Online Appendix

B.1 Proof of Remark 4

Claim: Let $\|\cdot\|$ be an arbitrary matrix norm. For any column-stochastic matrix P of nonnegative rank K we have

$$\mathcal{C}_{\mathsf{K}}(\mathsf{P}) \equiv \mathcal{C}_{\mathsf{K}} \cap \left\{ \mathsf{C} \in \mathbb{R}^{\mathsf{V}\times\mathsf{V}} \mid \mathsf{C}\mathsf{P}^{\text{row}} = \mathsf{P}^{\text{row}} \right\} \neq \emptyset$$

if and only if

$$\min_{\mathsf{C}\in\mathcal{C}_{\mathsf{K}}} \|\mathsf{C}\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| = 0.$$

Proof. We first show the " $\Longrightarrow$ " direction. Since $\mathsf{C}_{\mathsf{K}}(\mathsf{P}) \neq \emptyset$, then there exists $\mathsf{C}^* \in \mathcal{C}_{\mathsf{K}}$ such that $\mathsf{C}^*\mathsf{P}^{\text{row}} = \mathsf{P}^{\text{row}}$. Since,

$$0 \leqslant \inf_{\mathsf{C}\in\mathcal{C}_{\mathsf{K}}} \|\mathsf{C}\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| \leqslant \|\mathsf{C}^*\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| = 0,$$

then

$$\inf_{\mathsf{C}\in\mathcal{C}_{\mathsf{K}}} \|\mathsf{C}\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| = \|\mathsf{C}^*\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| = 0.$$

Thus, the infimum is attained and

$$\min_{\mathsf{C}\in\mathcal{C}_{\mathsf{K}}} \|\mathsf{C}\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| = 0.$$

For the " $\Longleftarrow$ " we note that if

$$\min_{\mathsf{C}\in\mathcal{C}_{\mathsf{K}}} \|\mathsf{C}\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| = 0,$$

then, by definition, there exists $\mathsf{C}^* \in \mathcal{C}_{\mathsf{K}}$ such that

$$\|\mathsf{C}^*\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}}\| = 0.$$

But since $\|\cdot\|$ is a norm, this implies $\mathsf{C}^*\mathsf{P}^{\text{row}} - \mathsf{P}^{\text{row}} = 0$. □

B.2 Proof of Remark 5

Let P, Q be column-stochastic matrices of dimension $\mathsf{V} \times \mathsf{D}$. Define the total-variation distance between P and Q as

$$\|\mathsf{P} - \mathsf{Q}\|_{\text{TV}} = \frac{1}{2} \sum_{\nu=1}^{\mathsf{V}} \sum_{\mathrm{d}=1}^{\mathsf{D}} |\mathsf{p}_{\nu,\mathrm{d}} - \mathsf{q}_{\nu,\mathrm{d}}|.$$

This extends the typical definition of the total-variation distance for discrete distributions; see p. 48, Proposition 4.2 in Levin & Peres (2017).

Claim: Suppose that P is a column-stochastic matrix of nonnegative rank $K \leqslant \min\{V, D\}$ that a) does not admit an anchor-word factorization in the sense of Definition 2, and b) there exists some $\epsilon > 0$

$$p_\nu \equiv \sum_{d=1}^{D} p_{\nu,d} > \epsilon, \quad \forall\, \nu = 1, \ldots, V.$$

Then, there is no sequence of matrices $\{P_i\}_{i\in\mathbb{N}}$ for which $P_i = A_i W_i$, $(A_i, W_i) \in \Theta_0$ and $\|P - P_i\|_{TV} \to 0$.

Proof. We establish this result by contradiction. Suppose there is a sequence $\{P_i\}_{i\in\mathbb{N}}$ for which $P_i = A_i W_i$, $(A_i, W_i) \in \Theta_0$ and $\|P - P_i\|_{TV} \to 0$. Theorem 1 shows that for each $i \in \mathbb{N}$, there exists a matrix $C_i \in \mathcal{C}_K$ such that

$$C_i P_i^{\text{row}} = P_i^{\text{row}}.$$

Let $\|\cdot\|$ denote the Frobenius norm. For any C_i satisfyng $CP_i = P_i$ we have

$$\begin{aligned}
\|C_i P^{\text{row}} - P^{\text{row}}\| &= \|C_i P^{\text{row}} - C_i P_i^{\text{row}} + C_i P_i^{\text{row}} - P_i^{\text{row}} + P_i^{\text{row}} - P^{\text{row}}\|, \\
&\leqslant \|C_i(P^{\text{row}} - P_i^{\text{row}})\| + \|C_i P_i^{\text{row}} - P_i^{\text{row}}\| + \|P_i^{\text{row}} - P^{\text{row}}\|, \\
&= \|C_i(P^{\text{row}} - P_i^{\text{row}})\| + \|P_i^{\text{row}} - P^{\text{row}}\|, \\
&\leqslant \big(\|C_i\| + 1\big) \cdot \|P_i^{\text{row}} - P^{\text{row}}\|.
\end{aligned}$$

Consequently,

$$\inf_{C\in\mathcal{C}_K} \|CP^{\text{row}} - P^{\text{row}}\| \leqslant \big(\|C_i\| + 1\big) \cdot \|P_i^{\text{row}} - P^{\text{row}}\| \tag{62}$$

for every $i \in \mathbb{N}$. Because $\mathcal{C}_K$ is bounded (as the matrices $C \in \mathcal{C}_K$ have elements in $[0, 1]$), then the sequence $\{\|C_i\|\}_{i\in\mathbb{N}}$ is bounded. Moreover,

$$\begin{aligned}
\|P^{\text{row}} - P_i^{\text{row}}\| &= \sqrt{\sum_{d=1}^{D}\sum_{\nu=1}^{V}(p_{\nu,d}^{\text{row}} - p_{i,(\nu,d)}^{\text{row}})^2} \\
&\leqslant \sum_{d=1}^{D}\sum_{\nu=1}^{V}|p_{\nu,d}^{\text{row}} - p_{i,(\nu,d)}^{\text{row}})| \\
&= \sum_{d=1}^{D}\sum_{\nu=1}^{V}\Big|\frac{p_{\nu,d}}{p_\nu} - \frac{p_{i,(\nu,d)}}{p_{i\nu}}\Big|,
\end{aligned}$$

where p_ν and $p_{i\nu}$ represent the row sums of P and P_i, respectively. Since

$$\left|\frac{p_{\nu,d}}{p_\nu} - \frac{p_{i,(\nu,d)}}{p_{i\nu}}\right| = \left|\frac{p_{\nu,d}}{p_\nu} - \frac{p_{i,(\nu,d)}}{p_\nu} + \frac{p_{i,(\nu,d)}}{p_\nu} - \frac{p_{i,(\nu,d)}}{p_{i\nu}}\right|,$$

then

$$\begin{aligned}\|P^{row} - P_i^{row}\| &\leqslant \sum_{d=1}^{D}\sum_{\nu=1}^{V}\frac{1}{p_\nu}\cdot|p_{\nu,d} - p_{i,(\nu,d)}| \\ &+ \sum_{d=1}^{D}\sum_{\nu=1}^{V}\frac{p_{i,(\nu,d)}}{p_\nu \cdot p_{i\nu}}\cdot|p_{i\nu} - p_\nu|.\end{aligned}$$

Since $\|P_i - P\|_{TV} \to 0$ implies that $|p_{i,(\nu,d)} - p_{\nu,d}| \to 0$ for all $\nu = 1,\ldots,V$ and $d = 1,\ldots,D$ then

$$\|P^{row} - P_i^{row}\| \to 0,$$

and, because of (62)

$$\inf_{C\in\mathcal{C}_K}\|CP^{row} - P^{row}\| = 0.$$

This implies, by Theorem 1 that P admits an anchor-word factorization. A contradition.

□

B.3 Proof that $\inf_{C\in\mathcal{C}_K}\|C\widehat{P}^{row} - \widehat{P}^{row}\|$ is always attained.

Claim: Let $\|\cdot\|$ denote the Frobenius norm. For any column-stochastic, row normalized matrix P^{row},

$$\inf_{C\in\mathcal{C}_K}\|CP^{row} - P^{row}\| = \min_{C\in\mathcal{C}_K}\|CP^{row} - P^{row}\|.$$

Proof. We want to show the minimum of $\|CP^{row} - P^{row}\|$ is attainable in $\mathcal{C}_K$ when the norm is Frobenius. By the extreme value theorem—e.g., Munkres (2000) Theorem 27.4 on page 174—it is sufficient to show function $f_P(C) \equiv \|CP^{row} - P^{row}\|$ is continuous in C over $\mathcal{C}_K$ and that $\mathcal{C}_K$ is compact. For the rest of the proof, we work with the topology induced by the Euclidean metric in R^{V^2}, and the topology over $R^{V\times V}$ induced by the Frobenius norm.

First, we show that $f_P(C)$ is continuous. For any $\varepsilon > 0$, there exists $\delta = \varepsilon/\|P^{row}\|$ such that if $\|C - C_0\| < \delta$, then

$$|\,\|CP^{row} - P^{row}\| - \|C_0P^{row} - P^{row}\|\,| \leqslant \|CP^{row} - C_0P^{row}\| \leqslant \|C - C_0\|\cdot\|P^{row}\| < \varepsilon.$$

The first inequality holds due to the reverse triangle inequality and the second inequality comes from

the submultiplicativity of the Frobenius norm; see Horn & Johnson (2012) page 340.

Second, we show that the set $\mathcal{C}_K$ is compact. It is sufficient to show $\mathcal{C}_K$ is closed since it is a subset of a compact space $[0,1]^{K\times K}$; see Munkres (2000) Theorem 26.2 on page 165. For the compactness of the space $[0,1]^{K\times K}$, we rely on facts that the space $[0,1]^{K^2}$ is compact and the image of a compact space under a continuous map is compact—see, for example, Munkres (2000) Theorem 26.5 on page 166—where we depend on the continuous bijection $h_{ij}(\tilde{C}) = \tilde{C}_{V(i-1)+j}$ for any $\tilde{C} \in [0,1]^{K^2}$.

For a sequence $\{C_n \in \mathcal{C}_K\}_{n\in\mathcal{N}}$ that converges, we want to show its limit C is in $\mathcal{C}_K$. Notice the matrix converges in the Frobenius norm is equivalent to entry-wise convergences in absolute values. That is, if $\lim_{n\to\infty} C_n = C$, for any $\varepsilon > 0$, there exists N such that if $n > N$, $|C_{n,ij} - C_{i,j}| \leqslant \|C_n - C\| \leqslant \varepsilon$. Also, if $\lim_{n\to\infty} C_{n,ij} = C_{ij}$ for all i and j, for any $\varepsilon/V > 0$, there exists $\{N_{ij}\}$ such that if $n > \sup\{N_{ij}\}$, $\|C_n - C\| \leqslant \sqrt{V^2(\frac{\varepsilon}{V})^2} = \varepsilon$. The last inequality is from the definition of the Frobenius norm.

Finally, by the definition of the convergence, the diagonal elements are bounded by 0 and 1, and the off-diagonal elements also share the same bounds because if $C_{n,ij} \leqslant C_{jj}$, $\lim C_{n,ij} \leqslant C_{jj}$. Therefore, C is in $\mathcal{C}_K$ and $\mathcal{C}_K$ is closed.

□

B.4 An anchor word factorization always exists when $K = 2 \leqslant \min\{V, D\}$

B.4.1 Proof using condition (19) of Theorem 1

Let P be a nonnegative column-stochastic matrix of rank $K = 2 \leqslant \min\{V, D\}$. Thomas (1974) has shown that every rank two nonnegative matrix admits a nonnegative matrix factorization. Let (A, W) be the nonnegative matrices in $\mathbb{R}^{2\times V} \times \mathbb{R}^{2\times D}$ that factorize P; that is $P = AW$.

Without loss of generality we can assume that A and W are column stochastic (that is, their columns add up to one). Also, suppose that the first term in the vocabulary solves the problem $c_1 \equiv \min_{v\in V} a_{v2}/a_{v1}$. That is, we assume that the first term of the vocabulary receives the lowest possible probability under topic two, relative to the probability that the same term receives under topic one. Analogously, suppose that the second term in the vocabulary solves $c_2 \equiv \min_{v\in V} a_{v1}/a_{v2}$. Note that if A were not organized in such a way, we could always permute the rows of A to achieve this structure. Note also that the ratios involving a_{v1} and a_{v2} are always well defined because none of the rows of P equal zero.

We will make use of the 2×2 matrix

$$T \equiv \begin{pmatrix} \frac{1}{1-c_2} & -\frac{c_1}{1-c_1} \\ \frac{-c_2}{1-c_2} & \frac{1}{1-c_1} \end{pmatrix},$$

where c_1 and c_2 are defined in the previous paragraph. Because A has rank two, both $c_1, c_2 \in (0,1)$.

This implies that T is well defined; that its determinant is strictly positive, and that T^{-1} is a column-stochastic matrix.

In a slight abuse of notation, write A as the following block matrix

$$A = \begin{bmatrix} \underbrace{A^*}_{2\times 2} \\ \underbrace{\tilde{A}}_{V-2\times 2} \end{bmatrix}.$$

Consider then the $V \times V$ matrix given by

$$\mathsf{C} \equiv \begin{bmatrix} \mathbb{I}_2 & \mathbf{0}_{2\times V-2} \\ (\mathcal{R}_{\tilde{A}W})^{-1}\tilde{A}\mathsf{T}\mathcal{R}_{\mathsf{T}^{-1}W} & \mathbf{0}_{V-2\times V-2} \end{bmatrix}. \tag{63}$$

We will show that this matrix satisfies the necessary and sufficient condition for anchor word factorization in Theorem 1.

We first show that C is an element of the set C_2 defined in Equation (17). Note first that $\mathrm{Tr}(\mathsf{C}) = 2$ and that the diagonal elements of the matrix C are either 0 or 1. Thus, we only need to show that the elements of the matrix

$$(\mathcal{R}_{\tilde{A}W})^{-1}\tilde{A}\mathsf{T}\mathcal{R}_{\mathsf{T}^{-1}W} \tag{64}$$

are nonnegative and bounded above by one.

We first show that the elements of (64) are nonnegative. Note that $\tilde{A}W$ (which corresponds to the lower $V-2 \times D$ block of P) is a nonnegative matrix, which implies $\mathcal{R}_{\tilde{A}W}$ is nonnegative. Note also that because T^{-1} is column stochastic, then $\mathsf{T}^{-1}W$ is a column-stochastic matrix. Finally, since $\tilde{A}$ is column stochastic and $\mathsf{c}_1, \mathsf{c}_2 \in (0,1)$, it follows that $\tilde{A}\mathsf{T}$ is nonnegative.

We then show that the elements of (64) are bounded above by one. Since, by definition, $\mathcal{R}_M$ is the diagonal matrix that contains the row sums of a matrix M, algebra shows that

$$\mathcal{R}_{\tilde{A}W} = \mathcal{R}_{(\tilde{A}\mathsf{T})(\mathsf{T}^{-1}W)} = \mathcal{R}_{\tilde{A}\mathsf{T}\mathcal{R}_{\mathsf{T}^{-1}W}}.$$

Thus, the elements of the $V-2 \times 2$ matrix (64) are bounded above by one. This shows that C is an element of the set C_2.

Finally, we show that C satisfies the equation $\mathsf{C}\mathsf{P}^{\text{row}} = \mathsf{P}^{\text{row}}$. Using the block matrix representation of A

$$\mathsf{P}^{\text{row}} = \begin{pmatrix} (A^*W)^{\text{row}} \\ \left(\tilde{A}W\right)^{\text{row}} \end{pmatrix}.$$

The definition of C in Equation (63) implies

$$\begin{aligned} CP^{\text{row}} &= \begin{pmatrix} (A^*W)^{\text{row}} \\ (\mathcal{R}_{\tilde{A}W})^{-1}\tilde{A}T\mathcal{R}_{T^{-1}W}\,(A^*W)^{\text{row}} \end{pmatrix}, \\ &= \begin{pmatrix} (A^*W)^{\text{row}} \\ (\mathcal{R}_{\tilde{A}W})^{-1}\tilde{A}T\mathcal{R}_{T^{-1}W}\left((A^*T)\left(T^{-1}W\right)\right)^{\text{row}} \end{pmatrix}. \end{aligned}$$

By construction, A^*T is a diagonal matrix, which implies

$$\left((A^*T)\left(T^{-1}W\right)\right)^{\text{row}} = \left(\left(T^{-1}W\right)\right)^{\text{row}} = \mathcal{R}_{T^{-1}W}T^{-1}W.$$

Thus, we conclude that $CP^{\text{row}} = P^{\text{row}}$, and thus $C \in \mathcal{C}_2(P)$. Theorem 1 thus implies that any matrix P of rank $K = 2$ admits an anchor word factorization.

B.4.2 Explicit anchor word factorization when $K = 2 \leqslant \min\{V, D\}$

The proof of Theorem 1 gives a simple formula to obtain the anchor word factorization of $\mathbb{P}$ from $C \in \mathcal{C}_2(P)$. In particular, if we start out with the factors (A, W) that were used in the previous subsection, the proof of Theorem 1 implies that the column-normalized version of the $V \times K$ matrix

$$\begin{bmatrix} \mathbb{I}_K \\ \tilde{A}T\mathcal{R}_{T^{-1}W}\mathcal{R}^{-1}_{A^*W} \end{bmatrix} \tag{65}$$

provides an anchor word factorization of P. Since A^*T is diagonal and column stochastic, then the matrix in (65) equals

$$\begin{bmatrix} A^*T \\ \tilde{A}T \end{bmatrix} (A^*T)^{-1},$$

where we have used

$$\mathcal{R}_{A^*W} = \mathcal{R}_{A^*TT^{-1}W} = A^*T\mathcal{R}_{T^{-1}W}.$$

Thus,

$$A_0 = \begin{bmatrix} A^*T \\ \tilde{A}T \end{bmatrix}$$

and $W_0 \equiv T^{-1}W$ provide an anchor word factorization of P.

B.5 Anchor word factorization does not always exists $\mathsf{V} = 4$, $\mathsf{K} = \mathsf{D} = 3$

B.5.1 Example

In this section we show that any matrix P of the form

$$\mathsf{P} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \gamma & 0 \\ 0 & 1-\gamma & 1-\beta \\ 1-\alpha & 0 & \beta \end{pmatrix},$$

for $\alpha, \beta, \gamma \in (0, 1)$ does not admit an anchor word factorization.

The row-normalized version of P is given by:

$$\mathsf{P}^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \end{pmatrix},$$

We define the set $\tilde{\mathcal{C}}_{\mathsf{K}}$ to be the set of $\mathsf{V} \times \mathsf{V}$ matrices of the form

$$\begin{bmatrix} \mathbb{I}_{\mathsf{K}} & 0_{\mathsf{K} \times \mathsf{V}-\mathsf{K}} \\ \mathsf{M} & 0_{\mathsf{V}-\mathsf{K} \times \mathsf{K}} \end{bmatrix},$$

where $\mathsf{M} \geqslant 0$ is a row-normalized matrix (with rows different from zero, so that row-normalization is always well defined). From Lemma 1, we want to show there does not exist $\mathsf{C} \in \tilde{\mathcal{C}}_{\mathsf{K}}$ and a row permutation matrix Π such that $\mathsf{C}\Pi\mathsf{P}^{\text{row}} = \Pi\mathsf{P}^{\text{row}}$.

Since $\mathsf{K} = 3$ we can argue that it is only relevant to focus on four classes of permutations (which are indexed by the row of P^{row} that is placed at the bottom of the permuted matrix). Without loss of generality, we can focus on

$$\mathsf{P}_1^{\text{row}} = \begin{pmatrix} \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & 1 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ 1 & 0 & 0 \end{pmatrix},$$

$$\mathsf{P}_2^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & 1 & 0 \end{pmatrix},$$

$$\mathsf{P}_3^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \end{pmatrix},$$

$$\mathsf{P}_4^{\text{row}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1-\gamma}{2-\gamma-\beta} & \frac{1-\beta}{2-\gamma-\beta} \\ \frac{1-\alpha}{1-\alpha+\beta} & 0 & \frac{\beta}{1-\alpha+\beta} \end{pmatrix}.$$

Note there is no $\mathsf{C} \in \tilde{\mathcal{C}}_{\mathsf{K}}$ such that $\mathsf{C}\mathsf{P}_{\mathfrak{i}}^{\text{row}} = \mathsf{P}_{\mathfrak{i}}^{\text{row}}$ for $\mathfrak{i} = 1, 2$, since this would require some elements of M to be strictly above one.

Consider now the matrices $\mathsf{P}_3^{\text{row}}$ and $\mathsf{P}_4^{\text{row}}$. We can focus on $\mathsf{P}_3^{\text{row}}$, since the argument for the other matrix is entirely analogous. Let the elements of M, which is a 1×3 matrix, be denoted as $[\mathfrak{m}_1, \mathfrak{m}_2, \mathfrak{m}_3]$. In order for the first element of the last row of $\mathsf{P}_3^{\text{row}}$ (which equals zero) to be a convex combination of the first three rows it is necessary to have $\mathfrak{m}_1 = \mathfrak{m}_3 = 0$. However, this implies that the last element of the fourth row of $\mathsf{P}_3^{\text{row}}$ (which equals $1 - \beta/2 - \gamma - \beta$) cannot be obtained as a convex combination of the first three rows, whenever $\beta \in (0, 1)$. Therefore there does not exist $\mathsf{C} \in \tilde{\mathcal{C}}_{\mathsf{K}}$ such that $\mathsf{C}\mathsf{P}_3^{\text{row}} = \mathsf{P}_3^{\text{row}}$. Since the argument for $\mathsf{P}_4^{\text{row}}$ is analogous, we conclude that the anchor word factorization does not exist for P.

B.6 Upper bound for $\mathsf{q}^*_{1-\alpha}(\mathsf{V}, \mathsf{K}, \mathsf{D}, \overline{\mathsf{N}}_{\mathsf{D}})$

Lemma 4. *Let $\|\cdot\|$ denote the Frobenius norm. For any $\alpha \in (0, 1)$*

$$\mathsf{q}^*_{1-\alpha}(\mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}}) \leqslant \sup_{\mathsf{C} \in \mathcal{C}_{\mathsf{K}}} \|\mathsf{C} - \mathbb{I}_{\mathsf{V}}\| \cdot \tilde{\mathsf{q}}^*_{1-\alpha}(\mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}}), \tag{66}$$

where

$$\tilde{\mathsf{q}}^*_{1-\alpha}(\mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}}) = \sup_{(\mathsf{A}, \mathsf{W}) \in \Theta_0} \tilde{\mathsf{q}}_{1-\alpha}(\mathsf{A}\mathsf{W}, \mathsf{V}, \mathsf{D}, \mathsf{K}, \overline{\mathsf{N}}_{\mathsf{D}})$$

and

$$\tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf\left\{ q \in \mathbb{R}_+ \;\middle|\; \mathbb{P}_{AW}\left(\|\widehat{P}^{row} - (AW)^{row}\| \leqslant q \right) \geqslant 1-\alpha. \right\}.$$

Proof. By definition—see Section 3.2— $q_{1-\alpha}(AW, V, D, K, \overline{N}_D)$ is the $1-\alpha$ quantile of the test statistic $T(Y)$ under the distribution $P = AW$, $(A, W) \in \Theta_0$. Thus:

$$q_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf\left\{ q \in \mathbb{R}_+ \;\middle|\; \mathbb{P}_{AW}\left(T(Y) < q \right) \geqslant 1-\alpha \right\}$$

Let $C_P \in \mathcal{C}_K$ be the matrix for which $CP^{row} - (AW)^{row} = \mathbf{0}$ (such a matrix exists by Theorem 1). Since the test statistic $T(Y)$ equals $\min_{C\in\mathcal{C}_K} \|C\widehat{P}^{row} - \widehat{P}^{row}\|$, it follows that

$$\begin{aligned} T(Y) &\leqslant \|C_P\widehat{P}^{row} - \widehat{P}^{row}\| \\ &= \|C_P\widehat{P}^{row} - C_P P^{row} + C_P P^{row} - P^{row} + P^{row} - \widehat{P}^{row}\| \\ &= \|\left(C_P - \mathbb{I}_V\right)\left(\widehat{P}^{row} - P^{row}\right)\| \\ &\leqslant \sup_{C\in\mathcal{C}_K} \|C - \mathbb{I}_V\| \cdot \|\widehat{P}^{row} - P^{row}\|, \end{aligned}$$

where the last inequality follows from the submultiplicativity of Frobenius norm. This inequality implies that

$$Q_1 \equiv \left\{ q \in \mathbb{R}_+ \;\middle|\; \mathbb{P}_{AW}\left(\sup_{C\in\mathcal{C}_K} \|C - \mathbb{I}_V\| \cdot \|\widehat{P}^{row} - P^{row}\| \leqslant q \right) \geqslant 1-\alpha \right\}$$

is a subset of

$$Q_0 \equiv \left\{ q \in \mathbb{R}_+ \;\middle|\; \mathbb{P}_{AW}\left(T(Y) < q \right) \geqslant 1-\alpha \right\}.$$

Therefore,

$$q_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf Q_0 \leqslant \inf Q_1. \tag{67}$$

Define $C^*(V, K) \equiv \sup_{C\in\mathcal{C}_K} \|C - \mathbb{I}_V\|$. We want to show that

$$\inf Q_1 \leqslant C^*(V, K) \cdot \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D).$$

Let

$$Q_2 \equiv \left\{ q \in \mathbb{R}_+ \;\middle|\; \mathbb{P}_{AW}\left(\|\widehat{P}^{row} - P^{row}\| \leqslant q \right) \geqslant 1-\alpha. \right\},$$

and note that, by definition,

$$\tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D) = \inf Q_2.$$

By definition of infimum, there exists a sequence $\{q_n\}_{n\in\mathbb{N}} \subseteq Q_2$ such that

$$\lim_{n\to\infty} q_n = \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D). \tag{68}$$

For each q_n we have that

$$\left(C^*(V, K) \cdot q_n\right) \in Q_1.$$

Consequently,

$$\inf Q_1 \leqslant C^*(V, K) \cdot q_n$$

for all $n \in \mathbb{N}$. We thus conclude by (68) that

$$\inf Q_1 \leqslant C^*(V, K) \cdot \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D)$$

and by (67) that

$$q_{1-\alpha}(AW, V, D, K, \overline{N}_D) \leqslant C^*(V, K) \cdot \tilde{q}_{1-\alpha}(AW, V, D, K, \overline{N}_D).$$

Taking the supremum on both sides over $(A, W) \in \Theta_0$ gives the desired result.

□

B.7 Estimation error of different estimators

In this section we discuss two alternative estimators for P^{row}. Here is a description of the estimators and the results we derive:

1. *Nuclear-Norm Minimizer:* Let $\widehat{P}_{\text{nuc}}$ be the estimator suggested by McRae & Davenport (2021), Section 2.3, Theorem 2.2, p. 712. The following proposition follows from their Theorem 2.2:

   Proposition 3. *Let* $0 < \gamma < 1$ *be an arbitrary scalar. For any* (A, W) *such that* $p_\nu(A, W)/D \geqslant \gamma/V$

   $$\|\widehat{P}^{row}_{nuc} - (AW)^{row}\|_F \leqslant 4\sqrt{\frac{16}{\gamma^2} \cdot \frac{V^{3/2} \cdot \ln\left((D+V)/\epsilon\right) \cdot K}{N_{min}}} \tag{69}$$

   with probability at least $1 - \epsilon$.

2. *Minimax Estimator for the columns:* Let $\widehat{P}_{\text{min}}$ the $V \times D$ matrix with (ν, d)-entry given by

$(\sqrt{\mathsf{N_d}}/\mathsf{V} + \mathfrak{n}_{\nu\mathrm{d}})/(\sqrt{\mathsf{N_d}} + \mathsf{N_d})$. Let $\widehat{\mathsf{P}}^{\text{row}}_{\text{min}}$ the row-normalized version of this estimator. In Section B.7.2 below we establish the following proposition:

Proposition 4. *Let* $0 < \gamma < 1$ *be arbitrary scalars. For any* (A, W) *such that* $\mathfrak{p}_\nu(\mathsf{A}, \mathsf{W})/\mathsf{D} \geqslant \gamma/\mathsf{V}$

$$\|\widehat{\mathsf{P}}^{\textit{row}}_{\mathtt{min}} - (\mathsf{AW})^{\textit{row}}\|_{\mathsf{F}} \leqslant \sqrt{\frac{8\left(1 - \frac{1}{\mathsf{V}}\right)}{\gamma^2 \cdot \epsilon} \cdot \frac{\mathsf{V}^2}{\mathsf{N}_{\textit{min}} + 2\mathsf{N}_{\textit{min}}^{1/2} + 1}} \tag{70}$$

with probability at least $1 - \epsilon$.

The estimator that row-normalizes that minimax estimator is expected to satisfy the high-level assumption in (26) provided

$$\frac{\mathsf{V}^2}{\mathsf{N}_{\text{min}} + 2\mathsf{N}_{\text{min}}^{1/2} + 1}$$

is small. Here, we rely on the same technique as Proposition 3 to derive the rate. We can also provide better rates with an order of

$$\frac{\mathsf{V}^2}{\mathsf{D} \cdot (\mathsf{N}_{\text{min}} + 2\mathsf{N}_{\text{min}}^{1/2} + 1)}$$

with other assumptions about probability design and other techniques.

Outline for this section: Let $\hat{\mathsf{P}}$ be an arbitrary estimator of the population term-document frequency matrix, P. Just as we did in the main body of the paper, define $\hat{\mathsf{P}}^{\text{row}} \equiv \mathcal{R}_{\hat{\mathsf{P}}}^{-1}\hat{\mathsf{P}}$ and $\mathsf{P}^{\text{row}} \equiv \mathcal{R}_{\mathsf{P}}^{-1}\mathsf{P}$. We establish a series of results that will allow us to provide finite-sample bounds for $\|\hat{\mathsf{P}}^{\text{row}} - \mathsf{P}^{\text{row}}\|_{\mathsf{F}}$.

Lemma 5 below shows that in order to upper-bound the estimation error $\|\hat{\mathsf{P}}^{\text{row}} - \mathsf{P}^{\text{row}}\|_{\mathsf{F}}$ we can analyze the terms

$$\|\mathcal{R}_{\mathsf{P}}^{-1}(\mathsf{P} - \hat{\mathsf{P}})\|_{\mathsf{F}} \tag{71}$$

and

$$\|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}}. \tag{72}$$

Lemma 6 uses Markov's inequality to provide an upper bound for the term in (71). Lemma 7 provides an upper bound for the term in (72). The bounds do not depend on the specific form of $\widehat{\mathsf{P}}$ as long as the second moments of the estimator exist.

Lemma 5. *If* $\|\mathcal{R}_{\mathsf{P}}^{-1}(\mathsf{P} - \hat{\mathsf{P}})\|_{\mathsf{F}} \leqslant \delta_1$ *with probability at least* $1 - \epsilon/2$*, and* $\|(\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}\|_{\mathsf{F}} \leqslant \delta_2$ *with*

probability at least $1-\epsilon/2$*, then with probability at least* $1-\epsilon$*,*

$$||\hat{\mathsf{P}}^{row} - \mathsf{P}^{row}||_{\mathsf{F}} \leqslant 2\max\{\delta_1, \delta_2\}$$

Proof. Algebra shows that

$$\begin{aligned}
||\hat{\mathsf{P}}^{\delta\text{row}} - \mathsf{P}^{\text{row}}||_{\mathsf{F}} &= ||\mathcal{R}_{\hat{\mathsf{P}}}^{-1}\hat{\mathsf{P}} - \mathcal{R}_{\mathsf{P}}^{-1}\mathsf{P}||_{\mathsf{F}} \\
&= ||\mathcal{R}_{\hat{\mathsf{P}}}^{-1}\hat{\mathsf{P}} - \mathcal{R}_{\mathsf{P}}^{-1}\hat{\mathsf{P}} + \mathcal{R}_{\mathsf{P}}^{-1}\hat{\mathsf{P}} - \mathcal{R}_{\mathsf{P}}^{-1}\mathsf{P}||_{\mathsf{F}} \\
&\leqslant ||\mathcal{R}_{\hat{\mathsf{P}}}^{-1}\hat{\mathsf{P}} - \mathcal{R}_{\mathsf{P}}^{-1}\hat{\mathsf{P}}||_{\mathsf{F}} + ||\mathcal{R}_{\mathsf{P}}^{-1}\hat{\mathsf{P}} - \mathcal{R}_{\mathsf{P}}^{-1}\mathsf{P}||_{\mathsf{F}} \\
&= ||\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})||_{\mathsf{F}} + ||(\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}||_{\mathsf{F}},
\end{aligned}$$

where the inequality comes from the triangle inequality.

The inequality above implies that for any constant c we have

$$\mathsf{P}(||\hat{\mathsf{P}}^{\text{row}} - \mathsf{P}^{\text{row}}||_{\mathsf{F}} > \mathsf{c}) \leqslant \mathsf{P}(||\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})||_{\mathsf{F}} + ||(\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}||_{\mathsf{F}} > \mathsf{c}).$$

Moreover, the right-hand side of the equation above is upper-bounded by

$$\mathsf{P}(||\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})||_{\mathsf{F}} > \mathsf{c}/2 \text{ or } ||(\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}||_{\mathsf{F}} > \mathsf{c}/2).$$

The subadditivity of probability measures then implies

$$\begin{aligned}
\mathsf{P}(||\hat{\mathsf{P}}^{\text{row}} - \mathsf{P}^{\text{row}}||_{\mathsf{F}} > \mathsf{c}) \leqslant &\mathsf{P}(||\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})||_{\mathsf{F}} > \mathsf{c}/2) \\
&+ \mathsf{P}(||(\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}}||_{\mathsf{F}} > \mathsf{c}/2).
\end{aligned}$$

Take $\mathsf{c} = 2\max\{\delta_1, \delta_2\}$ and note that

$$\mathsf{P}(||\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})||_{\mathsf{F}} > \max\{\delta_1, \delta_2\}) \leqslant \mathsf{P}(||\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})||_{\mathsf{F}} > \delta_1) < \epsilon/2,$$

and analogously $\mathsf{P}((\mathcal{R}_{\hat{\mathsf{P}}}^{-1} - \mathcal{R}_{\mathsf{P}}^{-1})\hat{\mathsf{P}} > \max\{\delta_1, \delta_2\}) < \epsilon/2$. □

Lemma 6. *Suppose that the second moments of* $\hat{\mathsf{p}}_{\nu\mathsf{d}}$ *exist for* $\nu = 1, ..., \mathsf{V}$ *and* $\mathsf{d} = 1, ..., \mathsf{D}$*. Then with probability at least* $1-\epsilon$

$$||\mathcal{R}_{\mathsf{P}}^{-1}(\hat{\mathsf{P}} - \mathsf{P})||_{\mathsf{F}} \leqslant \frac{1}{\mathsf{p}_{\nu min}}\sqrt{\frac{\sum_{\nu=1}^{\mathsf{V}}\sum_{\mathsf{d}=1}^{\mathsf{D}}\mathbb{E}\left[(\hat{\mathsf{p}}_{\nu\mathsf{d}} - \mathsf{p}_{\nu\mathsf{d}})^2\right]}{\epsilon}},$$

where the expectation $\mathbb{E}$ *is taken under the true data generating process* P*.*

Proof. The definition of Frobenius norm implies that for any $x > 0$

$$\begin{aligned}
P(\|\mathcal{R}_P^{-1}(\hat{P} - P)\|_F > x) &= P\left(\sum_v \sum_d \frac{1}{p_v^2}(p_{vd} - \hat{p}_{vd})^2 > x^2\right) \\
&\leqslant P\left(\frac{1}{p_{vmin}^2}\sum_v \sum_d (p_{vd} - \hat{p}_{vd})^2 > x^2\right) \\
&\leqslant \frac{\sum_v \sum_d \mathbb{E}(p_{vd} - \hat{p}_{vd})^2}{p_{vmin}^2 x^2},
\end{aligned}$$

where the last step follows from Markov's inequality. Taking x to be

$$\sqrt{\frac{\sum_{v=1}^{V}\sum_{d=1}^{D}\mathbb{E}\left[(\hat{p}_{vd} - p_{vd})^2\right]}{p_{vmin}^2 \epsilon}}$$

completes the proof. □

Lemma 7. *Suppose that the second moments of* $\hat{p}_{vd}$ *exist for* $v = 1, ..., V$ *and* $d = 1, ..., D$. *Then with probability at least* $1 - \epsilon$

$$\|(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_P^{-1})\hat{P}\|_F \leqslant \frac{1}{p_{vmin}}\sqrt{\frac{\sum_{v=1}^{V}\mathbb{E}\left[(p_v - \hat{p}_v)^2\right]}{\epsilon}}$$

where the expectation $\mathbb{E}$ *is taken under the true data generating process* P, *and* $p_v \equiv \sum_{d=1}^{d} p_{vd}$, $\hat{p}_v \equiv \sum_{d=1}^{d} \hat{p}_{vd}$.

Proof.

$$\begin{aligned}
\|(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_P^{-1})\hat{P}\|_F &= \left[\sum_v \sum_d (\frac{1}{p_v} - \frac{1}{\hat{p}_v})^2 \hat{p}_{vd}^2\right]^{1/2} \\
&= \left[\sum_v \sum_d \frac{(\hat{p}_v - p_v)^2}{p_v^2 \hat{p}_v^2}\hat{p}_{vd}^2\right]^{1/2} \\
&= \left[\sum_v \frac{(\hat{p}_v - p_v)^2}{p_v^2 \hat{p}_v^2}\sum_d \hat{p}_{vd}^2\right]^{1/2} \\
&\leqslant \left[\sum_v \frac{(\hat{p}_v - p_v)^2}{p_v^2 \hat{p}_v^2}\hat{p}_v^2\right]^{1/2}
\end{aligned}$$

$$\leqslant \left[\frac{1}{p_{\nu \min}^2} \sum_{\nu} (\hat{p}_{\nu} - p_{\nu})^2 \right]^{1/2}.$$

The inequality above holds since $(\sum_d \hat{p}_{\nu d}^2)^{1/2} \leqslant \sum_d \hat{p}_{\nu d} = \hat{p}_{\nu}$.

Then, for any $x > 0$

$$\begin{aligned} P(\|(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_{P}^{-1})\hat{P}\|_F > x) &\leqslant P\left(\frac{1}{p_{\nu \min}^2} \sum_{\nu} (\hat{p}_{\nu} - p_{\nu})^2 > x^2 \right) \\ &\leqslant \frac{\sum_{\nu} \mathbb{E}((\hat{p}_{\nu} - p_{\nu})^2)}{p_{\nu \min}^2 x^2}, \end{aligned}$$

where the last line follows by Markov's inequality. Taking

$$x = \frac{1}{p_{\nu \min}} \sqrt{\frac{\sum_{\nu} \mathbb{E}(p_{\nu} - \hat{p}_{\nu})^2}{\epsilon}},$$

yields the desired result. □

B.7.1 Estimation error of $P_{\text{freq}}^{\text{row}}$

Proof of Proposition 2. In a slight abuse of notation, let $\widehat{P}$ denote the $V \times D$ matrix with (ν, d)-entry given by $n_{\nu d}/N_d$. Let $\widehat{P}^{\text{row}}$ the row-normalized version of this estimator.

Note that

$$\begin{aligned} \sum_{\nu} \sum_{d} \mathbb{E}\left[(\hat{p}_{\nu d} - p_{\nu d}) \right]^2 &= \sum_{\nu} \sum_{d} \frac{p_{\nu d}(1 - p_{\nu d})}{N_d} \\ &\leqslant \sum_{\nu} \sum_{d} \frac{p_{\nu d}(1 - p_{\nu d})}{N_{\min}} \\ &= \sum_{d} \frac{1 - \sum_{\nu} p_{\nu d}^2}{N_{\min}} \\ &\leqslant \frac{D(1 - \frac{1}{V})}{N_{\min}}. \end{aligned}$$

The first equality holds because $n_{\nu d}$ is a binomial distribution with parameter N_d and $p_{\nu d}$. The second equality holds since the $\sum_{\nu} p_{\nu d} = 1$. The second inequality comes from the fact that

$$\min_{p_{1d}, \cdots, p_{Vd}} \sum_{\nu} p_{\nu d}^2 \quad \text{s.t.} \quad \sum_{\nu} p_{\nu d} = 1$$

equals $1/V$. Therefore, by Lemma 6 with probability at least $1 - \epsilon/2$

$$\|\mathcal{R}_P^{-1}(P - \hat{P})\|_F \leqslant \frac{1}{p_{\nu min}} \sqrt{\frac{2D(1 - \frac{1}{V})}{N_{min}\epsilon}}.$$

Moreover, since by assumption, $p_{\nu min}/D \geqslant \gamma/V$, we have that

$$\|\mathcal{R}_P^{-1}(P - \hat{P})\|_F \leqslant \sqrt{\frac{2V^2(1 - \frac{1}{V})}{\gamma^2 D N_{min}\epsilon}}.$$

Lemma 7 implies that with probability at least $1 - \epsilon/2$

$$\begin{aligned}
\|(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_P^{-1})\hat{P}\|_F &\leqslant \frac{1}{p_{\nu min}} \sqrt{\frac{2\sum_\nu E(p_\nu - \hat{p}_\nu)^2}{\epsilon}} \\
&= \frac{1}{p_{\nu min}} \sqrt{\frac{2\sum_\nu \sum_d \mathbb{E}\left[(\hat{p}_{\nu d} - p_{\nu d})\right]^2}{\epsilon}} \\
&= \frac{1}{p_{\nu min}} \sqrt{\frac{2D(1 - \frac{1}{V})}{N_{min}\epsilon}} \\
&\leqslant \sqrt{\frac{2V^2(1 - \frac{1}{V})}{\gamma^2 D N_{min}\epsilon}},
\end{aligned}$$

where the second equality holds because the estimators $\hat{p}_{\nu d}$ are unbiased and they are also independent across documents.

Finally, Lemma 5, implies that if $\hat{P}^{row}$ is based on the row-normalization of the empirical frequencies then

$$\|\widehat{P}^{row} - (AW)^{row}\|_F \leqslant \sqrt{\frac{8\left(1 - \frac{1}{V}\right)}{\gamma^2 \cdot \epsilon} \cdot \frac{V^2}{N_{min} \cdot D}}$$

with probability at least $1 - \epsilon$. □

B.7.2 Estimation error of P_{min}^{row}

Proof of Proposition 4. In a slight abuse of notation, let $\widehat{P}$ denote the $V \times D$ matrix with (ν, d)-entry given by $(\sqrt{N_d}/V + n_{\nu d})/(\sqrt{N_d} + N_d)$. Let $\widehat{P}^{row}$ be the row-normalized version of this estimator.

As above, we show that

$$\sum_\nu \sum_d \mathbb{E}\left[(\hat{p}_{\nu d} - p_{\nu d})\right]^2 = \sum_\nu \sum_d \frac{N_d p_{\nu d} - \frac{2N_d p_{\nu d}}{V} + \frac{N_d}{V^2}}{(\sqrt{N_d} + N_d)^2}$$

$$
\begin{aligned}
&\leqslant \sum_{\nu}\sum_{d}\frac{p_{\nu d}-\frac{2p_{\nu d}}{V}+\frac{1}{V^2}}{N_{min}+2N_{min}^{1/2}+1}\\
&=\sum_{d}\sum_{\nu}\frac{p_{\nu d}-\frac{2p_{\nu d}}{V}+\frac{1}{V^2}}{N_{min}+2N_{min}^{1/2}+1}\\
&=\frac{D(1-\frac{1}{V})}{N_{min}+2N_{min}^{1/2}+1}
\end{aligned}
$$

The first equality holds because $n_{\nu d}$ is a binomial distribution with parameter N_d and $p_{\nu d}$. The third equality holds since the $\sum_\nu p_{\nu d}=1$.

Therefore, by Lemma 6 with probability at least $1-\epsilon/2$

$$
\|\mathcal{R}_P^{-1}(P-\hat{P})\|_F \leqslant \frac{1}{p_{\nu min}}\sqrt{\frac{2D(1-\frac{1}{V})}{(N_{min}+2N_{min}^{1/2}+1)\epsilon}}.
$$

Moreover, since by assumption, $p_{\nu min}/D \geqslant \gamma/V$, we have that

$$
\|\mathcal{R}_P^{-1}(P-\hat{P})\|_F \leqslant \sqrt{\frac{2V^2(1-\frac{1}{V})}{\gamma^2 D(N_{min}+2N_{min}^{1/2}+1)\epsilon}}.
$$

Note that

$$
\sum_{\nu}\mathbb{E}\left[\sum_{d}(\hat{p}_{\nu d}-p_{\nu d})^2\right]=\sum_{\nu}\sum_{d}E(\hat{p}_{\nu d}-p_{\nu d})^2+\sum_{\nu}\sum_{d\neq d'}\mathbb{E}(\hat{p}_{\nu d}-p_{\nu d})\mathbb{E}(\hat{p}_{\nu d'}-p_{\nu d'}).
$$

We use the bound for the first term again and for the second term, we know

$$
\mathbb{E}(\hat{p}_{\nu d}-p_{\nu d})=\frac{\frac{1}{V}-p_{\nu d}}{\sqrt{N_d}+1}.
$$

So

$$
\begin{aligned}
\sum_{\nu}\sum_{d\neq d'}\mathbb{E}(\hat{p}_{\nu d}-p_{\nu d})\mathbb{E}(\hat{p}_{\nu d'}-p_{\nu d'}) &= \sum_{\nu}\sum_{d\neq d'}\frac{1}{(\sqrt{N_d}+1)^2}\left(\frac{1-V(p_{\nu d}+p_{\nu d'})}{V^2}+p_{\nu d}p_{\nu d'}\right)\\
&=\sum_{d\neq d'}\frac{1}{(\sqrt{N_d}+1)^2}\sum_{\nu}\left(\frac{1-V(p_{\nu d}+p_{\nu d'})}{V^2}+p_{\nu d}p_{\nu d'}\right)\\
&=\sum_{d\neq d'}\frac{1}{(\sqrt{N_d}+1)^2}\left(\sum_{\nu}p_{\nu d}p_{\nu d'}-\frac{1}{V}\right)
\end{aligned}
$$

$$\leqslant \sum_{d \neq d'} \frac{1}{(\sqrt{N_d}+1)^2}\left(1-\frac{1}{V}\right)$$
$$\leqslant \frac{D^2\left(1-\frac{1}{V}\right)}{N_{min}+2N_{min}^{1/2}+1}$$

The third equality holds since the $\sum_\nu p_{\nu d} = 1$. The first inequality comes from the fact that

$$\max \sum_\nu p_{\nu d} p_{\nu d'} \quad \text{s.t.} \quad \sum_\nu p_{\nu j} = 1 \quad \text{and} \quad p_{\nu j} \geqslant 0 \quad \text{for } j = d \text{ or } d'$$

equals to 1 by Kuhn-Tucker conditions. Therefore,

$$\sum_\nu \mathbb{E}\left[\sum_d (\hat{p}_{\nu d} - p_{\nu d})^2\right] \leqslant \frac{D(D+1)\left(1-\frac{1}{V}\right)}{N_{min}+2N_{min}^{1/2}+1}.$$

Lemma 7 implies that with probability at least $1-\epsilon/2$

$$\begin{aligned}\|(\mathcal{R}_{\hat{P}}^{-1} - \mathcal{R}_P^{-1})\hat{P}\|_F &\leqslant \frac{1}{p_{\nu min}}\sqrt{\frac{2\sum_\nu E(p_\nu - \hat{p}_\nu)^2}{\epsilon}} \\ &\leqslant \frac{1}{p_{\nu min}}\sqrt{2\frac{D(D+1)\left(1-\frac{1}{V}\right)}{N_{min}+2N_{min}^{1/2}+1}} \\ &\leqslant \sqrt{\frac{2(D+1)V^2(1-\frac{1}{V})}{\gamma^2 D\left(N_{min}+2N_{min}^{1/2}+1\right)\epsilon}},\end{aligned}$$

Finally, Lemma 5, implies that if $\hat{P}^{row}$ is based on the row-normalization of the minimax estimator then

$$\|\widehat{P}^{row} - (AW)^{row}\|_F \leqslant \sqrt{\frac{8\left(1-\frac{1}{V}\right)}{\gamma^2 \cdot \epsilon} \cdot \frac{V^2}{N_{min}+2N_{min}^{1/2}+1}}$$

with probability at least $1-\epsilon$. □

B.8 Topic estimation of FOMC1 corpus using Arora, Ge, Kannan & Moitra (2012), Ke & Wang (2022) and LDA

(a) Topic 1: wage

(b) Topic 2: recoveri

(c) Topic 3: kind

(d) Topic 4: uncertainti

Figure 12: Arora, Ge & Moitra (2012)'s estimator of A in the FOMC1 corpus. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largetst weights are colored. The estimated anchor words for each topic is in the caption.

(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

Figure 13: Ke & Wang (2022)'s estimator of A in the FOMC1 corpus. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largetst weights are colored. The estimated anchor words for each topic is in the caption.

(a) Topic 1

(b) Topic 2

(c) Topic 3

(d) Topic 4

Figure 14: Latent Dirichlet Allocation estimator of A in the FOMC1 corpus with uniform priors. Each panel shows the word cloud of words of a topic (column in A matrix), where the font size is proportional to term's weight in the topic, and the top 5 terms with largetst weights are colored. The estimated anchor words for each topic is in the caption.