Communication-Efficient Federated Learning of Latent Patient Representations from Multi-Institutional EHRs

Zhiyu Yan[*]
Department of Biostatistics
Harvard University
romanyan@g.harvard.edu

Shiao Liu[*]
Department of Biostatistics
Harvard University
liushiao1@gmail.com

Rahul Mahajan
Department of Neurology
Mass General Brigham
rmahajan@mgb.org

Anand Viswanathan
Department of Neurology
Mass General Brigham
aviswanathan1@mgb.org

Christopher D. Anderson[†]
Department of Neurology
Mass General Brigham
cdanderson@mgb.org

Rui Duan[†]
Department of Biostatistics
Harvard University
rduan@hsph.harvard.edu

Abstract

Electronic health records (EHRs) contain rich health information for uncovering latent patient representations and clinically meaningful subgroups. Learning from EHR data across institutions is valuable because representations learned at a single site may reflect local data patterns and lack generalizability, yet pooling patient-level records across institutions is often impractical because of privacy constraints and communication costs. We propose a communication-efficient, non-iterative federated spectral method for learning shared latent representation components from multi-institutional EHRs, in which each institution shares aggregate marginal frequencies and a low-rank spectral summary while keeping patient-level records local. A central server aggregates these summaries with the pooled marginal frequencies to recover both the shared low-rank structure and the canonical row-normalization direction required for interpretable representation recovery. We prove row-wise recovery guarantees and error bounds, with a leading federated variance term matching a corresponding lower bound under explicit regularity conditions. Simulations show that our method closely matches pooled analysis in estimating latent representation components, improves as more sites contribute summaries, and outperforms single-site and existing federated benchmarks across a range of data settings. Applied to real-world multi-facility EHR data for stroke patients, the method learns interpretable patient representations that highlight clinically distinct subgroups and could help inform subgroup-specific care.

1 Introduction

Electronic health records (EHRs) provide a broad view of patient care and are increasingly used to generate real-world evidence, characterize disease phenotypes, support risk stratification, and inform clinical decisions. Shaped by complex care processes, EHR data are inherently high-dimensional, heterogeneous, and sparse. Patient representation learning offers a way to summarize these complex records into lower-dimensional latent variables or distributions that can support downstream subgroup discovery, outcome prediction, and generative modeling, which help guide more personalized care strategies [4, 11, 15–17, 22].

[*]Equal contribution. [†]Equal contribution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Arranging EHRs into concept-by-patient counts provides a clean input format for patient representation learning. Here *concepts* are derived clinical data fields that can be counted within a patient record. Structured data may contribute diagnosis, medication, procedure, or laboratory concepts, while unstructured notes may contribute concepts after tokenization or other natural-language preprocessing. This count-based format naturally fits topic modeling, in which each topic is a shared probability distribution over concepts and each patient is represented by nonnegative topic-membership weights that sum to one [2, 6, 20, 21]. We refer to the topics as shared representation components and to a patient's membership-weight vector as the patient representation. This parameterization gives a direct interpretation to both the concept-level loadings and the patient-level coordinates.

Learning patient representations from EHR data collected at multiple institutions is important for improving generalizability, increasing sample size, and studying rare clinical events. Representation components learned at a single hospital may instead reflect local patient case mix or care processes rather than more general disease patterns. However, directly pooling patient-level EHR records across institutions is often infeasible due to data-governance restrictions, privacy concerns, infrastructure barriers, and the computational and communication costs of centralizing large clinical datasets.

We propose Federated Topic-SCORE, a communication-efficient non-iterative spectral method for learning shared latent patient representation components from multi-institutional EHRs. The method builds on the singular value decomposition (SVD)-based Topic-SCORE framework [10], which uses low-rank spectral coordinates, row normalization, and simplex geometry to estimate topic loadings. Standard Topic-SCORE, however, assumes access to the pooled concept-by-patient matrix, which is precisely what multi-institutional EHR settings often cannot provide. In Federated Topic-SCORE, each site keeps patient-level records local and sends only a marginal concept-frequency vector and a low-rank SVD summary. The server combines these summaries to recover the shared low-rank structure and the normalization needed to estimate the global representation component matrix.

Federated Topic-SCORE is distinguished from existing related work in two respects. First, existing federated representation learning and topic modeling methods, such as federated latent Dirichlet allocation (LDA) and federated neural topic modeling, typically rely on iterative updates within collaborating sites, iterative communication across sites, or both [3, 8, 12, 13, 19]. In iterative federated learning settings, each site repeats local computation, summary preparation and transfer, personnel coordination, and waiting for other sites in each iteration, which can significantly impede the practical deployment of the federated algorithm [9, 18]. Federated Topic-SCORE instead estimates the shared topic-loading matrix through a non-iterative spectral scheme, reducing both computation and communication costs. Second, a related line of work on distributed principal component analysis and shared singular-subspace estimation also aggregates local spectral information across sites [5, 14]. However, these methods only focus on subspace estimation, whereas Federated Topic-SCORE further recovers the row-normalized spectral coordinates and simplex vertices needed for interpretable nonnegative topic loadings.

The resulting federated estimator of topic loadings comes with row-wise recovery guarantees for the normalized spectral coordinates used in the simplex step, as well as error bounds whose leading federated variance term matches a corresponding lower bound under explicit regularity conditions. Simulations show that our method closely matches the pooled Topic-SCORE without sharing patient-level records, improves as more sites participate, and outperforms single-site and existing federated topic modeling benchmarks under varying parameter and data settings. In a multi-facility EHR application, Federated Topic-SCORE recovers representation components closer to the results obtained from the pooled data than any single-site analysis and produces interpretable representations whose dominant memberships highlight patient subgroups with distinct survival patterns.

Our contributions are:

- We develop a communication-efficient federated spectral algorithm for learning shared latent patient representation components from multi-institutional EHR data, without sharing patient-level records or requiring iterative cross-site communication.
- We establish finite-sample guarantees for our method, including row-wise control of the normalized spectral coordinates used for latent component recovery and error bounds under heterogeneous site-level membership mixtures.
- We demonstrate the improved performance of our method over comparable benchmarks through simulations, and illustrate the utility of our method in a real-world multi-facility EHR application.

2 Problem setup and method

2.1 Federated latent representation model

Suppose there are M institutions sharing a common, consistently indexed vocabulary of p EHR concepts and a central server that aggregates summaries from all sites. Site m has a sample size of n_m and forms the concept-by-patient frequency matrix $D^{(m)} = [d_1^{(m)}, \ldots, d_{n_m}^{(m)}] \in \mathbb{R}^{p \times n_m}$, where each column is the patient's raw concept-count vector normalized by its total count. Write $D = [D^{(1)}, \ldots, D^{(M)}]$ and $n_{\text{tot}} = \sum_m n_m$.

Let K denote the prescribed number of latent representation components. The shared components are parameterized by the topic-loading matrix $A = [A_{\cdot 1}, \ldots, A_{\cdot K}] \in \mathbb{R}^{p \times K}$, whose columns are probability distributions over EHR concepts. Patient i at site m has membership weights $w_i^{(m)}$. Specifically, $A_{\cdot k} \in \Delta^{p-1}$ and $w_i^{(m)} \in \Delta^{K-1}$, where $\Delta^{q-1} = \{x \in \mathbb{R}^q : x \geq 0, \mathbf{1}^\top x = 1\}$. For simplicity, in the analysis below, we assume a common total concept count N per patient, so that $N d_i^{(m)}$ is the raw concept-count vector. Conditional on these weights,

$$N d_i^{(m)} \sim \text{Multinomial}\{N, A w_i^{(m)}\}, \tag{1}$$

independently across all patients and sites. The topic-loading matrix A is shared across sites, while the membership vectors $w_i^{(m)}$ may differ across sites. Thus, institutions may have different patient mixtures while sharing the same latent representation components. Remark 20 discusses the modifications needed for unequal count lengths.

Let $W^{(m)} = [w_1^{(m)}, \ldots, w_{n_m}^{(m)}]$ and $W = [W^{(1)}, \ldots, W^{(M)}]$. Throughout the theoretical analysis, A and the membership matrices are fixed unknown parameters, and all probabilities and expectations are conditional on W. Define $D_0^{(m)} = \mathbb{E}[D^{(m)} \mid W] = AW^{(m)}$ and $D_0 = [D_0^{(1)}, \ldots, D_0^{(M)}] = AW$. Define the pooled and local membership second moments as $\Sigma_W = n_{\text{tot}}^{-1} W W^\top$ and $\Sigma_{W,m} = n_m^{-1} W^{(m)} W^{(m)\top}$. Each site first sends only its marginal concept-frequency vector $\bar{d}^{(m)} = n_m^{-1} D^{(m)} \mathbf{1}$ to the server. The server aggregates $\bar{d} = n_{\text{tot}}^{-1} \sum_{m=1}^M n_m \bar{d}^{(m)}$ and $M_d = \text{diag}(\bar{d})$. At the population level, with $w = n_{\text{tot}}^{-1} W \mathbf{1}$, define $m_0 = Aw$ and $M_0 = \text{diag}(m_0)$.

The spectral step targets the marginal-frequency normalized loading space

$$\mathcal{S} = \text{col}(M_0^{-1/2} A). \tag{2}$$

The pooled marginal normalization is what places all sites in this common spectral coordinate system. If every $\Sigma_{W,m}$ is nonsingular, then the columns of $M_0^{-1/2} D_0^{(m)}$ span $\mathcal{S}$. Its left singular vectors may nevertheless differ across sites because $\Sigma_{W,m}$ varies with m. Thus, local patient mixtures may differ, but the population representation component subspace is shared.

We next specify the regularity conditions used in our analysis.

Assumption 1 (Regularity: anchor concepts and locally rich mixtures)**.** For $a_j^\top$, the j-th row of A, let $h_j = \|a_j\|_1$, $h_{\min} = \min_j h_j$, $H = \text{diag}(h_1, \ldots, h_p)$, and $\Sigma_A = A^\top H^{-1} A$. There exists a constant $c > 0$, independent of M, N, p and the site sample sizes $n_1, \ldots, n_M$, such that: (i) each latent representation component k has at least one anchor concept, defined as a row j with $A(j,k) > 0$ and $A(j,\ell) = 0$ for all $\ell \neq k$; (ii) $h_{\min} \geq cK/p$; and (iii) $\lambda_{\min}(\Sigma_A) \geq c$, $\lambda_{\min}(\Sigma_W) \geq c$, and $\min_{m \leq M} \lambda_{\min}(\Sigma_{W,m}) \geq c$.

The anchor-concept condition enables simplex-based topic recovery. The lower bound on $\lambda_{\min}(\Sigma_A)$ prevents degeneracy of the topic-loading geometry, while the bounds on Σ_W and $\Sigma_{W,m}$ ensure sufficiently rich pooled and site-specific membership mixtures. The local condition requires enough variation within each site to identify the common signal space. The setting with near-violation of this condition is examined in Appendices A.4.7 and B.3.

2.2 Canonical coordinate geometry and vertex hunting

Under Assumption 1, we fix a strictly positive first basis vector of $\mathcal{S}$ using the pooled marginal.

Proposition 1 (Pooled-marginal canonical direction)**.** *Under Assumption 1,*

$$\xi_1 = M_0^{1/2}\mathbf{1} \in \mathcal{S}, \qquad \|\xi_1\|_2 = 1,$$

and ξ_1 is strictly positive. Let $\Xi_\perp$ be any orthonormal basis of $\mathcal{S} \cap \xi_1^\perp$ and set $\Xi = [\xi_1, \Xi_\perp]$. Then there is a unique nonsingular $V \in \mathbb{R}^{K\times K}$ such that

$$\Xi = M_0^{-1/2}AV, \qquad V_{\cdot,1} = w.$$

For each concept j, define the oracle ratio point $r_j \in \mathbb{R}^{K-1}$ by $r_j(\ell) = \Xi_\perp(j,\ell)/\xi_1(j)$, $1 \le \ell \le K-1$. For each component k, define $v_k^\star \in \mathbb{R}^{K-1}$ by $v_k^\star(\ell) = V_{k,\ell+1}/V_{k,1}$. The concept-level barycentric coordinate vector $\pi_j \in \Delta^{K-1}$ has entries $\pi_j(k) = A(j,k)V_{k,1}/\sum_{s=1}^K A(j,s)V_{s,1} = A(j,k)w(k)/m_0(j)$. These quantities satisfy

$$r_j = \sum_{k=1}^K \pi_j(k)v_k^\star, \qquad A\,\mathrm{diag}(w) = M_0^{1/2}\,\mathrm{diag}(\xi_1)\Pi, \qquad \Pi = [\pi_1,\ldots,\pi_p]^\top. \tag{3}$$

An anchor concept for component k has $r_j = v_k^\star$. The K vertices are affinely independent, so their convex hull is a $(K-1)$-dimensional simplex containing all ratio points. Recovering the barycentric coordinates gives $A\,\mathrm{diag}(w)$ through (3)*. Column normalization removes the positive factors $w(k)$, recovering A up to a column permutation.*

The server therefore estimates the pooled-marginal canonical direction and chooses an orthonormal basis for its complement in the estimated signal space. Different complementary bases rotate the ratio points and their vertices together, leaving the barycentric coordinates unchanged.

After forming empirical ratio points $\widehat{r}_j$, vertex hunting estimates the simplex vertices $v_1^\star,\ldots,v_K^\star$. We use augmented successive projection by default, while any routine satisfying the stability condition stated in Section 3 can also be used. Given the vertices, we then use the standard Topic-SCORE barycentric-coordinate recovery described in Appendix A.3.3, solving the system in (8) and then normalizing the result to obtain the topic loadings.

2.3 Federated Topic-SCORE

We now turn the preceding geometric framework into the federated protocol in Algorithm 1. Patient-level records remain local, while sites share aggregate summaries. Each site first uploads $\bar{d}^{(m)} \in \mathbb{R}^p$. The server then uses these uploads to computes M_d and broadcasts it back to the sites. Each site next uploads $\widehat{\Xi}^{(m)} \in \mathbb{R}^{p\times K}$, whose columns are the leading K left singular vectors of $M_d^{-1/2}D^{(m)}$. Finally, the server broadcasts $\widehat{A}_{\mathrm{fed}}$ for local membership weight estimation. The entire protocol requires no repeated cross-site model-update rounds.

As shown in Algorithm 1, sites use truncated SVD to form their local estimates of $\mathcal{S}$. The server perform a weighted aggregation of the projection matrices based on the local estimates, denoted as $\bar{P}$, using nonnegative weights $\alpha_m \ge 0$, where $\sum_m \alpha_m = 1$. A natural default choice for the weights is $\alpha_m = n_m/n_{\mathrm{tot}}$, which assigns greater weight to larger sites. The projector $\widehat{P}$ represents the orthogonal projector onto the span of eigenvectors associated with the K largest eigenvalues of $\bar{P}$. Aggregating projectors removes dependence on rotations of the local orthonormal bases, since such rotations leave the corresponding projection matrices unchanged. After topic recovery, each site estimates patient memberships by unconstrained weighted least squares followed by truncation and ℓ_1-normalization:

$$\widehat{w}_i^{(m),\star} = \operatorname*{arg\,min}_{b\in\mathbb{R}^K} \left\| M_d^{-1/2}\{d_i^{(m)} - \widehat{A}_{\mathrm{fed}}b\}\right\|_2^2, \qquad \widehat{w}_i^{(m)} = \mathcal{T}(\widehat{w}_i^{(m),\star}), \tag{4}$$

where $\mathcal{T}(b) = [b]_+/\|[b]_+\|_1$, with $\mathcal{T}(b) = K^{-1}\mathbf{1}$ if $[b]_+ = 0$.

Remark 1 (Working local rank). Although our theory is developed in the setting with local upload of rank K, sites may upload a modestly larger rank $L \ge K$ before the server extracts a rank-K aggregate. This can improve finite-sample coverage under noisy or heterogeneous local spectra, but overly large L can add noisy directions. We discuss more details of the local SVD rank in Appendix A.17.

Algorithm 1 Federated Topic-SCORE to estimate latent representation components and memberships

Require: Local concept-by-patient matrices $\{D^{(m)}\}_{m=1}^{M}$ (separately supplied by each site and staying local), number of latent representation components K, weights $\{\alpha_m\}_{m=1}^{M}$, vertex-hunting routine

Ensure: Topic loading estimate $\widehat{A}_{\mathrm{fed}}$ and local memberships $\{\widehat{w}_i^{(m)}\}$

Marginal pre-normalization.

1: Each site sends $\bar{d}^{(m)} = n_m^{-1} D^{(m)}\mathbf{1}$ to the central server.

2: Server forms $\bar{d} = n_{\mathrm{tot}}^{-1} \sum_m n_m \bar{d}^{(m)}$ and sets $M_d = \mathrm{diag}(\bar{d})$.

3: Server broadcasts M_d to all sites.

Local spectral upload.

4: Each site computes $D_f^{(m)} = M_d^{-1/2} D^{(m)}$ and uploads $\widehat{\Xi}^{(m)} \in \mathbb{R}^{p\times K}$, whose columns are its leading K left singular vectors.

Server-side simplex recovery.

5: Form $\bar{P} = \sum_m \alpha_m \widehat{\Xi}^{(m)}\widehat{\Xi}^{(m)\top}$ and extract its top-K projector $\widehat{P}$.

6: Set $\widehat{\xi}_1 = \widehat{P} M_d^{1/2}\mathbf{1} / \|\widehat{P} M_d^{1/2}\mathbf{1}\|_2$.

7: Choose an orthonormal $\widehat{\Xi}_\perp$ spanning $\mathrm{range}(\widehat{P} - \widehat{\xi}_1\widehat{\xi}_1^\top)$, and set $\widehat{r}_j(\ell) = \widehat{\Xi}_\perp(j,\ell)/\widehat{\xi}_1(j)$.

8: Set $\widehat{y}_j = (1, \widehat{r}_j^\top)^\top \in \mathbb{R}^K$ and use vertex hunting to obtain $\widehat{v}_1, \ldots, \widehat{v}_K \in \mathbb{R}^{K-1}$.

9: Form $\widehat{Q} = \begin{pmatrix} 1 & \cdots & 1 \\ \widehat{v}_1 & \cdots & \widehat{v}_K \end{pmatrix}$. For each concept j, solve $\widehat{Q}\widehat{\pi}_j^\star = \widehat{y}_j$ and set $\widehat{\pi}_j = \mathcal{T}(\widehat{\pi}_j^\star)$, with $\mathcal{T}$ defined after (4).

10: Set $\widehat{\Pi} = [\widehat{\pi}_1, \ldots, \widehat{\pi}_p]^\top$ and $\widetilde{A}_{\mathrm{fed}} = M_d^{1/2}\,\mathrm{diag}(\widehat{\xi}_1)\widehat{\Pi}$.

11: Compute $\widehat{A}_{\mathrm{fed},\cdot k} = \widetilde{A}_{\mathrm{fed},\cdot k} / \|\widetilde{A}_{\mathrm{fed},\cdot k}\|_1$ for $k = 1, \ldots, K$.

Local membership estimation.

12: Server broadcasts $\widehat{A}_{\mathrm{fed}}$ to all site, and each site estimates $\widehat{w}_i^{(m)}$ for its own patients using (4).

3 Theoretical guarantees

We study the theoretical guarantees for the estimation accuracy of Federated Topic-SCORE. Specifically, we derive finite-sample error bounds for estimation of the ratio coordinates and the shared topic loading matrix A. These bounds quantify how estimation error depends on the vocabulary size, local sample size, record length, and number of participating sites.

We establish the topic-loading error bounds in three stages. First, projector averaging recovers the shared spectral space up to a federated variance term, a local second-order spectral bias, and a row-wise range term. Second, the pooled marginal identifies the canonical positive direction inside this recovered space. Third, row-wise error bounds for the resulting canonical basis yield bounds for the ratio coordinates and, through vertex hunting, for the topic loading matrix.

Note that the main upper bounds below assume equal site sizes $n_m = n$, equal aggregation weights $\alpha_m = 1/M$, fixed K, and $N \geq 2$. The weighted projector-averaging argument extends to unequal site sizes and general deterministic weights $\{\alpha_m\}_{m=1}^{M}$. Constants C may vary between displays and depend only on K, the regularity constants in Assumption 1, and, where applicable, the stability constants in Assumption 2 below. Appendix A provides the full finite-sample statements, constant dependencies, normalization-bridge decomposition, and proofs, including the extension to settings with unequal site sizes.

For column-normalized topic loading matrices, we measure estimation error using the permutation-aligned ℓ_1 distance

$$L(\widehat{A}, A) = \min_{\kappa \in S_K} \sum_{k=1}^{K} \|\widehat{A}_{\cdot,\kappa(k)} - A_{\cdot,k}\|_1,$$

where S_K denotes the set of permutations of $\{1, \ldots, K\}$. To state the finite-sample bounds, fix a failure probability $\delta \in (0,1)$ and let $\ell = \ell_{n,M}(\delta) = \log(2pM/\delta)$. We distinguish three contributions to the error bounds,

$$T_{\mathrm{var}} = \beta_n \sqrt{\frac{p\ell}{NnM}}, \qquad T_{\mathrm{loc}} = \beta_n^2 \frac{p\ell}{Nn}, \qquad T_{\mathrm{row}} = \frac{\sqrt{p}\,\ell}{M}.$$

Here, $\beta_n \geq 1$ is a deterministic factor arising in the concentration bounds for normalized multinomial counts, with its definition given in Appendix A.4.4. The term T_{var} is the federated variance contribution, T_{loc} is the local second-order spectral bias that is not averaged away across sites, and T_{row} arises from bounding the averaged projector error uniformly across rows.

We combine these contributions with the remaining error terms by defining

$$\mathcal{R}_{n,M}(\delta) = T_{\mathrm{var}} + T_{\mathrm{loc}} + T_{\mathrm{row}} + \mathrm{Rem}_{n,M,N,p}(\delta). \tag{5}$$

The remainder accounts for empirical marginal normalization, the diagonal bias of ordinary SVD, and higher-order terms in the projector perturbation expansion. The complete expression for $\mathcal{R}_{n,M}(\delta)$ is given in Appendix A.4.4.

The theorem below uses the deterministic smallness conditions stated in Theorem 5 in Appendix A.4.4. These conditions keep the perturbations sufficiently small for the subspace and ratio constructions to remain stable. Let $e_j \in \mathbb{R}^p$ denote the jth standard basis vector. The following theorem bounds the row-wise errors in the canonical basis and the associated ratio coordinates.

Theorem 1 (Row-normalized spectral coordinate recovery)**.** *Suppose Assumption 1 holds, $n_m = n$, $\alpha_m = 1/M$, and the deterministic smallness conditions in Theorem 5 are satisfied. Then, with probability at least $1 - \delta$, there exists an orthogonal matrix $O_\perp \in \mathbb{R}^{(K-1)\times(K-1)}$ such that*

$$\max_j h_j^{-1/2} \left\| e_j^\top \{[\widehat{\xi}_1, \widehat{\Xi}_\perp O_\perp] - [\xi_1, \Xi_\perp]\} \right\|_2 + \max_j \|O_\perp^\top \widehat{r}_j - r_j\|_2 \leq C\mathcal{R}_{n,M}(\delta).$$

While Theorem 1 above controls the estimation error in the ratio coordinates, recovering the topic-loading matrix also requires the vertex-hunting step to be stable under small uniform perturbations of these coordinates. The following assumption formalizes this stability requirement.

Assumption 2 (Stable vertex hunting)**.** The vertex-hunting routine is stable under uniform ratio perturbations: if $\max_j \|O^\top \widehat{r}_j - r_j\|_2 \leq \varepsilon$ for some orthogonal $O \in \mathbb{R}^{(K-1)\times(K-1)}$ and sufficiently small ε, then, up to a permutation, $\max_k \|O^\top \widehat{v}_k - v_k^\star\|_2 \leq C\varepsilon$.

Note that Proposition 2 in Appendix A shows that the augmented successive-projection routine satisfies Assumption 2 under the anchor-concept and nondegenerate-simplex conditions.

Combining Theorem 1 with Assumption 2 yields the following topic-loading recovery guarantee.

Theorem 2 (Topic loading recovery)**.** *Suppose Assumptions 1 and 2 hold, $n_m = n$, $\alpha_m = 1/M$, and the deterministic smallness conditions in Theorem 5 are satisfied, with the additional vertex-hunting restriction specified in Theorem 6. Then, with probability at least $1 - \delta$, up to a common permutation of latent representation components,*

$$\max_j \frac{\|\widehat{a}_j - a_j\|_1}{h_j} + L(\widehat{A}_{\mathrm{fed}}, A) \leq C\mathcal{R}_{n,M}(\delta).$$

Here $\widehat{a}_j^\top$ is the jth row of $\widehat{A}_{\mathrm{fed}}$ after the common column permutation.

The following theorem establishes a minimax lower bound of order $\sqrt{p/(NnM)}$ for estimating the topic-loading matrix over the stated parameter class.

Theorem 3 (Minimax lower bound for estimating the topic-loading matrix)**.** *Fix $K \geq 2$. There exist constants $c_\star(K) \in (0, 1/(2K)]$ and $p_0(K)$ such that the following holds. For any fixed $c_{\mathrm{reg}} \in (0, c_\star(K)]$, let the supremum below range over all pLSI parameters $(A, W^{(1:M)})$ satisfying Assumption 1 with $c = c_{\mathrm{reg}}$. Suppose $n_m = n \geq K$, $p \geq p_0(K)$, and $\log(nM) \leq \min\{p, N\}$, $p\log(nM) = o(NnM)$. Then there exist constants $c, c' > 0$, depending only on K and c_{reg}, such that, along any sequence satisfying these conditions and for all sufficiently large problem sizes,*

$$\inf_{\widehat{A}} \sup_{(A, W^{(1:M)})} \mathbb{P}_{A,W^{(1:M)}} \left\{ L(\widehat{A}, A) \geq c\sqrt{\frac{p}{NnM}} \right\} \geq c'.$$

The lower bound has the same dependence on p, N, n, M as the federated variance component T_{var}, up to logarithmic and β_n factors. The complete finite-sample upper guarantee is given by (5).

The error bound in (5) also clarifies how the local sample size n, record length N, and number of participating sites M affect estimation accuracy. Ignoring logarithmic and β_n factors, increasing n or N reduces both T_{var} and T_{loc}, while increasing M reduces T_{var} and T_{row}. The local spectral bias T_{loc} arises from the site-level SVDs computed before aggregation. Averaging more site summaries does not remove this bias. Consequently, at fixed n and N, local bias can eventually dominate the error bound as M increases. In our analysis, T_{row} arises when bounding the errors in the estimated ratio coordinates for all concepts simultaneously.

4 Numerical experiments and real-world EHR application

4.1 Simulations

We conduct simulation studies to compare Federated Topic-SCORE with pooled Topic-SCORE, single-site Topic-SCORE, and a federated latent Dirichlet allocation (LDA) benchmark, using the permutation-aligned ℓ_1 distance from the simulated ground-truth topic matrix A as the evaluation metric. The Federated LDA benchmark is computed based on a recent federated LDA topic modeling method [12]. This method avoids repeated cross-site synchronization, but its site-level LDA fits are iterative and computationally heavier than Federated Topic-SCORE. We begin with a base configuration that serves as the base setting for all subsequent simulation variations and sensitivity analyses, where $M = 5, n_m = 200$ for all sites, $p = 1000, N = 2000$, and $K = 5$. In our simulations, we consider both homogeneous and heterogeneous topic membership weight distributions across sites. Appendix B.1 provides details of the full data generating processes, and Appendix D provides a link to the code repository.

Federated and pooled Topic-SCORE achieve similar recovery performance when membership weight distributions are homogeneous across sites. We performed a series of sensitivity analyses by varying one data-generating parameter at a time while holding all others fixed at their base-case values. These results are consistent with the leading terms in $R_{n,M}$ in that increasing p raises the error, whereas increasing n or N reduces the variance and local bias contributions. Across these settings, the federated estimator nearly overlaps with pooled Topic-SCORE and remains more accurate than single-site estimation, including under site-size imbalance (Figure 1). Federated LDA has substantially larger error throughout the homogeneous settings, consistent with the original Topic-SCORE evidence that LDA-type estimators can be less accurate for topic-loading recovery [10]. Additional sensitivity analyses for settings where membership weight distributions are homogeneous across sites are provided in Appendix B.2.

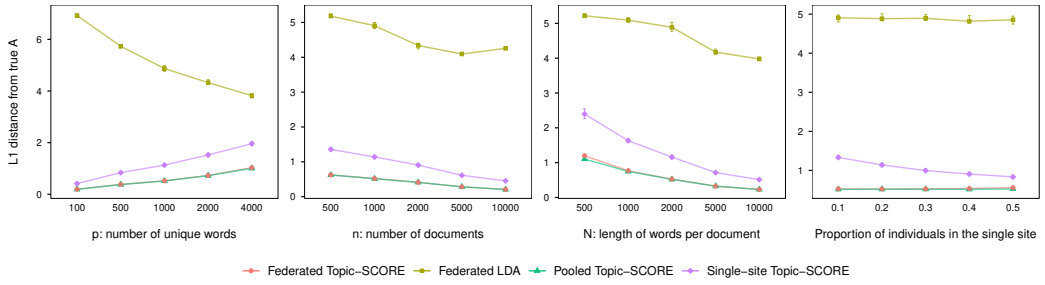


Figure 1: Estimation accuracy of the topic loading matrix A under homogeneous weight distributions across sites. From left to right, panels vary vocabulary size p, total sample size n, count length N, and the fraction of patients at a single site (with the remaining patients distributed equally among the other sites). Points show the median of the ℓ_1 distance across 30 simulation replicates, and error bars show the interquartile ranges. Distinct point shapes and colors identify the methods. The Federated Topic-SCORE and pooled Topic-SCORE curves nearly overlap across the displayed settings.

Federated Topic-SCORE achieves better recovery as more sites participate. As more sites contribute summary statistics, the estimation error of Federated Topic-SCORE decreases across

vocabulary sizes (Figure 2, left). This is consistent with the error bound in Theorem 2, whose federated variance and row-wise terms decrease with M when the per-site sample size is fixed.

Federated Topic-SCORE performs well under heterogeneous membership weight distributions across sites. When membership weights at one site become increasingly concentrated on a single topic while the other sites remain balanced, the single-site estimator fitted to that site becomes less accurate. Such concentration can make $\Sigma_{W,m}$ nearly singular, undermining the local richness condition in Assumption 1. Despite the undermined local topic richness, Federated Topic-SCORE remains close to pooled Topic-SCORE in our experiment (Figure 2, right), suggesting that incorporating summary statistics from the other balanced sites can mitigate the effect of topic imbalance at one site. Federated LDA is relatively stable as the site becomes more imbalanced, but its estimation error remains substantially higher than that of Federated Topic-SCORE. Additional simulations with more severe topic imbalance at one or multiple sites are described in Appendix B.3.

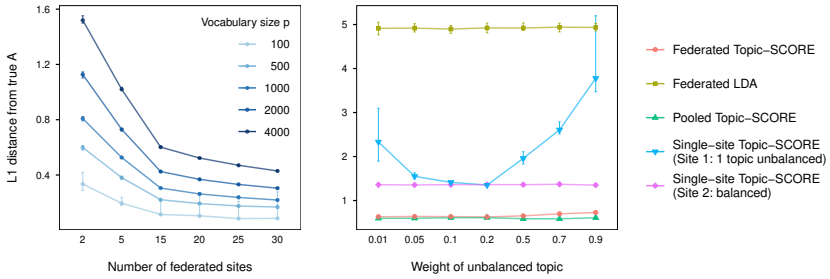


Figure 2: Estimation accuracy of A under varying numbers of participating sites and heterogeneous membership weight distributions across sites. Left: Federated Topic-SCORE error versus the number of participating sites, with per-site sample size fixed as in the base setting. Colors denote vocabulary size p. Right: comparison of methods as the membership weight of one topic for all patients at Site 1 varies, while all other sites remain balanced. Distinct point shapes and colors identify the methods and different single sites. Points show median ℓ_1 distance across 30 simulation replicates, and error bars show interquartile ranges.

4.2 Real-world EHR application: learning stroke patient representations

We applied Federated Topic-SCORE to real-world EHR data of stroke patients from four hospitals in the Mass General Brigham healthcare system, including Massachusetts General Hospital (MGH), Newton–Wellesley Hospital (NWH), Brigham and Women's Hospital (BWH), and Spaulding Rehabilitation Hospital (SRH), with the goal of learning latent representation components of stroke subtypes and characterizing patient subgroups using the representations. The final cohort included 7,057 patients. For each patient, we set their first qualifying acute ischemic stroke admission date from January 1, 2005 through January 31, 2025 as the index date. At each hospital, we constructed a concept-by-patient count matrix using diagnosis codes recorded during the first 30 days after the index date. Patients were followed for 1-year survival, with death as the event and censoring at the last available encounter or the end of follow-up. We used $K = 5$ for interpretability and local working rank $L = 10$, as discussed in Appendix A.17. The study was approved by the Mass General Brigham Institutional Review Board. Appendix C provides details about the cohort description and concept construction.

Federated Topic-SCORE estimates are closer to the pooled benchmark than single-site estimates. Because the ground-truth topic loading matrix is unknown for real EHR data, we use pooled Topic-SCORE as a centralized reference to understand the performance of our method. We compare federated and single-site estimates with the pooled estimate using the permulation-aligned ℓ_1 distance. Federated Topic-SCORE has the smallest distance to the pooled estimate, whereas single-site estimates differ across hospitals and are farther from the reference (Figure 3, left). These differences may reflect variation in local case mix and coding practices.

Learned representation components highlight clinically distinct patient subgroups. The latent components learned by Federated Topic-SCORE yield clinically interpretable post-stroke EHR

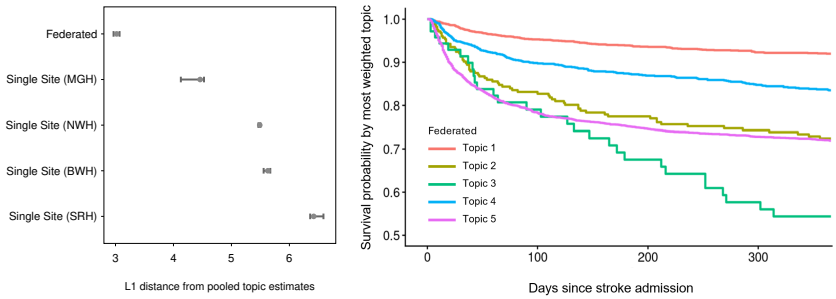


Figure 3: Federated Topic-SCORE estimation and subgroup survival patterns in the stroke EHR cohort. Left: ℓ_1 distance from pooled Topic-SCORE for federated estimation and single-site estimation at Massachusetts General Hospital (MGH), Newton–Wellesley Hospital (NWH), Brigham and Women's Hospital (BWH), and Spaulding Rehabilitation Hospital (SRH). Points and error bars show medians and interquartile ranges across 20 replicates for the vertex hunting procedure. Right: 1-year survival curves for patient subgroups defined by their largest estimated topic membership weights.

concept profiles, as shown by the highest-loading and anchor concepts in Figure 4. We then assigned each patient to the subgroup corresponding to their largest membership weight. Each EHR concept profile is clinically consistent with the corresponding patient subgroup's 1-year survival pattern (Figure 3, right). The subgroup characterized by cancer-care and hemorrhagic-event concepts shows the sharpest early survival decline, while the remaining profiles form a graded survival structure over follow-up. The highest-survival subgroup is dominated by vascular-sequelae and neurological-symptom concepts such as hypertension, aphasia, and hemiplegia, whereas lower-survival groups involve more medically complex multi-system profiles, including infections, cardiac and respiratory complications, vascular insufficiency, and fracture. These profiles should be read as EHR-derived phenotype summaries and require downstream clinical validation, but they may provide hypotheses for more personalized stroke and post-stroke management.

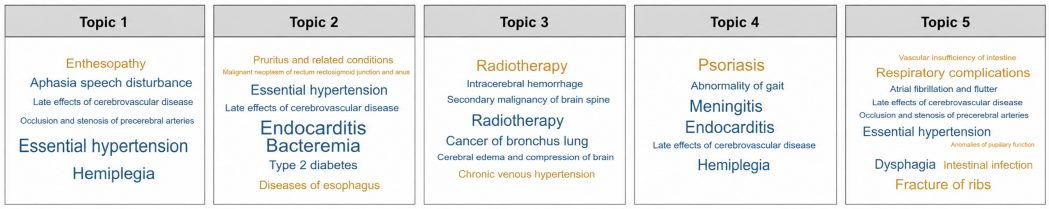


Figure 4: Topics by Federated Topic-SCORE learned from EHR concepts for stroke patients. Each panel summarizes one latent representation component (topic) using diagnosis-code-based concepts. Blue (darker) terms denote the highest-loading concepts, whereas gold (lighter) terms denote anchor concepts.

5 Discussion, limitations, and broader impact

Federated Topic-SCORE provides a non-iterative approach to learn shared latent patient representation components from multi-institutional EHR data. Each site communicates a marginal concept-frequency vector and one low-rank spectral summary, allowing the server to recover latent component structure close to pooled Topic-SCORE without patient-level data sharing. Our simulations show that Federated Topic-SCORE has lower estimation error with larger local sample sizes, longer records, and more participating sites, but higher error with larger vocabularies. These trends are consistent with the bounds that we derived on topic loading estimation error. The method performs well when membership weight distributions differ across sites, although severe topic imbalance at some of the sites can reduce estimation accuracy. The method also substantially outperforms the existing federated LDA benchmark, which is a useful non-iterative cross-site comparator but relies on computationally

intensive local LDA fits and has weaker topic-loading recovery, consistent with prior Topic-SCORE comparisons to LDA-type estimators [10].

In our application to real-world EHR data from stroke patients, the Federated Topic-SCORE estimate was closer to the pooled reference than any single-site estimate. The model then represented each patient by membership weights across topics learned without using any specific outcome labels. These weights form representations for discovery of patient subgroups potentially informative for multiple outcomes. In this study, we assigned each patient to the topic with the largest weight as a simple way to define subgroups for interpretation and comparison and obtained subgroups with distinct survival patterns. In general, patient subgroups can also be defined using the full membership weight vectors, corresponding to each patient's full representation.

Specifying the number of representation components rank K controls representation granularity. Smaller K yields broader and often more stable profiles, whereas larger K may reveal finer subgroups at the cost of interpretability and finite-sample stability. In the EHR application, we use $K = 5$ for interpretability and local working rank $L = 10$ to improve finite-sample subspace coverage before extracting the global rank-K component. Technical discussion of a pilot-stage rule for selecting K and L is deferred to Appendix A.11 and Appendix A.17, respectively. In practice, representation interpretability is also an important consideration when choosing both ranks.

We emphasize that Federated Topic-SCORE can be applied to context-aware EHR concepts based on large language models (LLMs). Recent work on patient-centered topic modeling with LLMs suggests that context-aware models can help construct clinically meaningful concepts or topic-like summaries from unstructured EHR data such as clinical notes. Such LLM-derived embeddings could be created locally and mapped to a shared vocabulary or count data format as described in the existing study [1], after which our method can be applied directly. This application in a multi-institutional setting for unstructured data is an important future step for federated representation learning.

The shared topic loading assumption has the advantage of directly modeling a general population in which shared disease- or biology-related profiles can be studied and expressed, and site-specific $W^{(m)}$ can capture differences in patient case mix and heterogeneity across sites. While the varying data collection approaches across sites could imply different concept meanings, this issue can be mitigated by data harmonization preprocessing or a common data model that aligns concept semantics before modeling [7].

Several limitations should be considered when interpreting the findings of our study. First, the theory assumes a shared-topic pLSI model based on a harmonized EHR concept vocabulary, anchor concept structure, rare-concept preprocessing, and sufficient local representation richness. These conditions are important for identifiability and local spectral recovery. They can be partly assessed through vocabulary checks and anchor diagnostics, and the robustness of our method under near-violations of some assumptions is discussed in the appendix. Second, the federated LDA is a practical topic modeling benchmark for the same non-iterative cross-site communication regime, but we did not provide an exhaustive comparison with iterative federated topic models as they fall into a different communication scheme. Third, the real-world analysis uses data from one multi-facility healthcare network, so its clinical findings may lack generalizability to other health systems and patient populations. We therefore view the application as a feasibility and clinical interpretability demonstration rather than external validation. Fourth, the learned representations and associated survival differences are hypothesis-generating. The survival analysis is associative rather than causal or predictive, and the representations require clinical review and external validation before being used to guide care. Finally, summary sharing reduces patient-level raw-record exposure but does not provide formal differential privacy. Marginal and spectral summaries may still reveal aggregate site-level information, especially for rare concepts or specialized facilities. The current setting is designed for addressing privacy concerns with respect to institutional data-governance restrictions and can be combined with differential privacy approaches as a future direction.

By enabling non-iterative multi-site representation learning without patient-level data pooling, this work may make collaborative patient profiling across broader and more generalizable patient populations practically feasible. The non-iterative summary-sharing design reduces repeated coordination burden, which is often a major barrier to sustained clinical data collaboration. The resulting interpretable EHR concept representations can support hypothesis generation for risk stratification, care-pathway evaluation, and more personalized care.

Acknowledgments and Disclosure of Funding

We thank the reviewers and the area chair for their constructive feedback, which helped improve this work. Rui Duan acknowledges support for this work from the National Institutes of Health (R01 GM148494) and the Patient-Centered Outcomes Research Institute (PCORI) Project Program Award ME-2024C1-37351. Zhiyu Yan is supported by the American Heart Association (26GWT-GDRASp1611353 and 26PRE1563536). Christopher D. Anderson is supported by the National Institutes of Health (U01NS069673 and RF1NS139183) and the Global Brain Care Coalition. Christopher D. Anderson reports sponsored research support from Bayer AG and consulting for Apopharma and MPM BioImpact, unrelated to the present work.

References

[1] Morgane Austern, Yuanchuan Guo, Zheng Tracy Ke, and Tianle Liu. Poisson-process topic model for integrating knowledge from pre-trained language models. *arXiv preprint arXiv:2503.17809*, 2025.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[3] Lorena Calvo-Bartolomé and Jerónimo Arenas-García. Federated neural topic models. *arXiv preprint arXiv:2212.02269*, 2022.

[4] Lingling Ding, Ravikiran Mane, Zhenzhou Wu, Yong Jiang, Xia Meng, Jing Jing, Weike Ou, Xueyun Wang, Yu Liu, Jinxi Lin, et al. Data-driven clustering approach to identify novel phenotypes using multiple biomarkers in acute ischaemic stroke: a retrospective, multicentre cohort study. *eClinicalMedicine*, 53:101639, 2022. doi: 10.1016/j.eclinm.2022.101639.

[5] Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009–3031, 2019.

[6] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[7] George Hripcsak, Ning Shang, Peggy L. Peissig, Luke V. Rasmussen, Cong Liu, Barbara Benoit, Robert J. Carroll, David S. Carrell, Joshua C. Denny, Ozan Dikilitas, Vivian S. Gainer, Kayla Marie Howell, Jeffrey G. Klann, Iftikhar J. Kullo, Todd Lingren, Frank D. Mentch, Shawn N. Murphy, Karthik Natarajan, Jennifer A. Pacheco, Wei-Qi Wei, Ken Wiley, and Chunhua Weng. Facilitating phenotype transfer using a common data model. *Journal of Biomedical Informatics*, 96:103253, 2019. doi: 10.1016/j.jbi.2019.103253.

[8] Di Jiang, Yuanfeng Song, Yongxin Tong, Xueyang Wu, Weiwei Zhao, Qian Xu, and Qiang Yang. Federated topic modeling. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, page 1071–1080. Association for Computing Machinery, 2019. doi: 10.1145/3357384.3357909.

[9] Xilin Jiang, Martin Jinye Zhang, Yidong Zhang, Arun Durvasula, Michael Inouye, Chris Holmes, Alkes L. Price, and Gil McVean. Age-dependent topic modeling of comorbidities in UK Biobank identifies disease subtypes with differential genetic risk. *Nature Genetics*, 55(11): 1854–1865, 2023. doi: 10.1038/s41588-023-01522-8.

[10] Zheng Tracy Ke and Minzhe Wang. Using SVD for topic modeling. *Journal of the American Statistical Association*, 119(545):434–449, 2024. doi: 10.1080/01621459.2022.2123813.

[11] S Khurshid, C Reeder, LX Harrington, P Singh, G Sarma, SF Friedman, P Di Achille, N Diamant, JW Cunningham, AC Turner, et al. Cohort design and natural language processing to reduce bias in electronic health records research. *npj Digital Medicine*, 5(1):47, 2022.

[12] Richa Kushwaha and Parmeet Kaur. A federated LDA-based approach for topic modelling. *Intelligent Decision Technologies*, 19(4):2738–2759, 2025. doi: 10.1177/18724981251319629.

[13] Chengjie Ma, Yawen Li, Meiyu Liang, and Ang Li. Federated topic model and model pruning based on variational autoencoder. In *Proceedings of 2023 Chinese Intelligent Automation Conference*, pages 51–60. Springer Nature Singapore, 2023. ISBN 9789819961870. doi: 10.1007/978-981-99-6187-0_5. URL http://dx.doi.org/10.1007/978-981-99-6187-0_5.

[14] Zhengchi Ma and Rong Ma. Optimal estimation of shared singular subspaces across multiple noisy matrices. *arXiv preprint arXiv:2411.17054*, 2024.

[15] G. Miceli, M.G. Basso, G. Rizzo, C. Pintus, E. Cocciola, A.R. Pennacchio, and A. Tuttolomondo. Artificial intelligence in acute ischemic stroke subtypes according to Toast classification: A comprehensive narrative review. *Biomedicines*, 11(4):1138, Apr 2023. doi: 10.3390/biomedicines11041138.

[16] Wi-Sun Ryu, Dawid Schellingerhout, Hoyoun Lee, Keon-Joo Lee, Chi Kyung Kim, Beom Joon Kim, Jong-Won Chung, Jae-Sung Lim, Joon-Tae Kim, Dae-Hyun Kim, et al. Deep learning-based automatic classification of ischemic stroke subtype using diffusion-weighted images. *Journal of Stroke*, 26(2):300–311, 2024.

[17] Michael Simmons, Ayush Singhal, and Zhiyong Lu. Text mining for precision medicine: bringing structure to EHRs and biomedical literature to understand genes and health. In *Translational Biomedical Informatics: A Precision Medicine Perspective*, pages 139–166. Springer, 2016.

[18] Jiayi Tong, Jenna M. Reps, Chongliang Luo, Yiwen Lu, Lu Li, Juan Manuel Ramirez-Anguita, Milou T. Brand, Scott L. DuVall, Thomas Falconer, Alex Mayer Fuentes, Xing He, Michael E. Matheny, Miguel A. Mayer, Bhavnisha K. Patel, Katherine R. Simon, Marc A. Suchard, Guojun Tang, Benjamin Viernes, Ross D. Williams, Mui van Zandt, Fei Wang, Jiang Bian, Jiayu Zhou, David A. Asch, and Yong Chen. Unlocking efficiency in real-world collaborative studies: a multi-site international study with one-shot lossless GLMM algorithm. *npj Digital Medicine*, 8 (1):457, 2025. doi: 10.1038/s41746-025-01846-1.

[19] Yansheng Wang, Yongxin Tong, and Dingyuan Shi. Federated latent Dirichlet allocation: A local differential privacy based framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6283–6290, 2020. doi: 10.1609/aaai.v34i04.6096.

[20] H. Ye, T. Moreno, A. Alpern, L. Ehwerhemuepha, and A. Qu. Dynamic topic language model on heterogeneous children's mental health clinical notes. *The Annals of Applied Statistics*, 18 (4):3165–3184, Dec 2024.

[21] H. Zhang, C. Zang, Z. Xu, Y. Zhang, J. Xu, J. Bian, D. Morozyuk, D. Khullar, Y. Zhang, A.S. Nordvig, E.J. Schenck, E.A. Shenkman, R.L. Rothman, J.P. Block, K. Lyman, M.G. Weiner, T.W. Carton, F. Wang, and R. Kaushal. Data-driven identification of post-acute SARS-CoV-2 infection subphenotypes. *Nature Medicine*, 29(1):226–235, January 2023. doi: 10.1038/s41591-022-02116-3.

[22] Jingqing Zhang, Luis Daniel Bolanos Trujillo, Ashwani Tanwar, Julia Ive, Vibhor Gupta, and Yike Guo. Clinical utility of automatic phenotype annotation in unstructured clinical notes: intensive care unit use. *BMJ Health & Care Informatics*, 29(1):e100519, 2022.

A Proofs and technical details

A.1 Notation map and proof-appendix scope

This appendix gives a self-contained proof of the theoretical results for the federated Topic-SCORE estimator used in the paper. We use the standard pLSI/topic-model terminology in the proof: words are EHR concepts or codes, documents are patient records, and topics are shared representation components. The topic matrix A is the shared topic-loading matrix in the main text: its columns describe the shared components. Each column w_i of the topic-weight matrix W is a patient representation. Topic-SCORE refers to the SCORE-style spectral recovery chain: estimate a low-rank spectral space, choose a positive first direction, divide the remaining coordinates rowwise by that direction to form ratio coordinates, and use simplex geometry to recover topic loadings. Whitening in this appendix means marginal-frequency pre-normalization by $M_0^{-1/2}$ or $M_d^{-1/2}$, rather than covariance whitening. All probabilities and expectations are conditional on the fixed membership matrices $W^{(1:M)}$, exactly as in the fixed-design Topic-SCORE analysis. The appendix restates the model, regularity assumptions, oracle simplex geometry, projector-averaging bounds, empirical-pre-normalization bridge, row-wise canonical-basis recovery, deterministic simplex propagation, the minimax lower bound, the held-out membership-weight guarantee, and the optional overspecified-rank calculation. The proof-level radius $\varepsilon_{n,M}(\delta)$ defined in (21) is the same quantity denoted $\mathcal{R}_{n,M}(\delta)$ in the main text; the notation ε is used in the appendix only to keep formulas inherited from the spectral perturbation proof compact.

A.2 Model and the federated setting

A.2.1 pLSI model

The vocabulary size is p and the number of topics is $K \ll \min\{p, n_{\mathrm{tot}}\}$. All documents have the common length N. Throughout the upper-bound theory we assume $N \geq 2$, so the signal factor $1 - 1/N$ is bounded away from zero. The empirical word-frequency vector of document i is

$$d_i(j) = \frac{\#\{\text{occurrences of word } j \text{ in document } i\}}{N}, \qquad j = 1, \ldots, p.$$

Let $D = [d_1, \ldots, d_{n_{\mathrm{tot}}}] \in \mathbb{R}^{p \times n_{\mathrm{tot}}}$ be the corpus matrix.

Definition 1 (Topic matrix and topic weights)**.** The topic matrix is $A = [A_1, \ldots, A_K] \in \mathbb{R}^{p \times K}$, where each column is a probability mass function on the vocabulary: $A_k \geq 0$ and $\mathbf{1}^\top A_k = 1$. Document i has topic weights $w_i \in \mathbb{R}^K$ with $w_i \geq 0$ and $\mathbf{1}^\top w_i = 1$. Write

$$W = [w_1, \ldots, w_{n_{\mathrm{tot}}}] \in \mathbb{R}^{K \times n_{\mathrm{tot}}}.$$

Assumption 3 (pLSI generation mechanism)**.** For each document i,

$$N d_i \sim \mathrm{Multinomial}(N, A w_i),$$

and documents are independent conditional on W. Hence the mean matrix is $D_0 := \mathbb{E}[D \mid W] = AW$.

A.2.2 Federated partition and pooled notation

There are M clients. Client m owns n_m documents and forms

$$D^{(m)} = [d_1^{(m)}, \ldots, d_{n_m}^{(m)}] \in \mathbb{R}^{p \times n_m}, \qquad W^{(m)} = [w_1^{(m)}, \ldots, w_{n_m}^{(m)}] \in \mathbb{R}^{K \times n_m}.$$

We work in the shared-topic regime: the topic matrix A is common across clients, while the client-level topic-weight matrices $W^{(m)}$ may differ across sites. Define

$$n_{\mathrm{tot}} := \sum_{m=1}^{M} n_m, \qquad D = [D^{(1)}, \ldots, D^{(M)}], \qquad W = [W^{(1)}, \ldots, W^{(M)}].$$

For any corpus matrix $X \in \mathbb{R}^{p \times n}$, let

$$\bar{x} := \frac{1}{n} X \mathbf{1}, \qquad M_X := \mathrm{diag}(\bar{x}).$$

The pooled population and empirical marginals are

$$m_0 := \frac{1}{n_{\text{tot}}} D_0 \mathbf{1}, \qquad M_0 := \text{diag}(m_0), \qquad \bar{d} := \frac{1}{n_{\text{tot}}} D\mathbf{1}, \qquad M_d := \text{diag}(\bar{d}).$$

For each client,

$$\bar{d}^{(m)} := \frac{1}{n_m} D^{(m)} \mathbf{1}, \qquad M_d^{(m)} := \text{diag}(\bar{d}^{(m)}),$$

and let the corresponding local population marginal be

$$m_{0,m} := \frac{1}{n_m} D_0^{(m)} \mathbf{1}, \qquad M_{0,m} := \text{diag}(m_{0,m}).$$

Throughout, A and $W^{(1:M)}$ are treated as fixed unknown matrices satisfying deterministic regularity conditions. All probabilities and expectations are conditional on $W^{(1:M)}$. This is the same fixed-design convention as in the single-site Topic-SCORE analysis of Ke and Wang's SVD Topic-SCORE paper.

For the topic weights, define the pooled second moment and the client-level second moments

$$\Sigma_W := \frac{1}{n_{\text{tot}}} W W^\top, \qquad \Sigma_{W,m} := \frac{1}{n_m} W^{(m)} W^{(m)\top}, \qquad 1 \le m \le M.$$

When $\lambda_{\min}(\Sigma_{W,m}) > 0$, the columns of

$$M_0^{-1/2} D_0^{(m)} = M_0^{-1/2} A W^{(m)}$$

span the common space

$$\mathcal{S} := \text{col}(M_0^{-1/2} A).$$

This K-dimensional space, rather than any client-specific spectral basis, is the deterministic target of the federated spectral procedure.

Remark 2 (Shared loading matrices and site heterogeneity). The shared-A assumption does not require the participating sites to have identical patient populations. The site-specific membership matrices $W^{(m)}$ may vary arbitrarily and therefore capture differences in case mix, topic prevalence, and patient-level mixtures.

More generally, if a site-specific loading matrix satisfies $A_m = AB_m$ for some column-stochastic matrix B_m, then

$$A_m W^{(m)} = A\big(B_m W^{(m)}\big).$$

Thus, heterogeneity lying within the convex hull of the shared topic dictionary can be absorbed exactly into the membership weights.

What the current model does not capture is topic-conditional coding variation outside this shared span, such as when the same clinical phenotype induces systematically different code distributions across sites. Allowing site-specific loading matrices A_m would address this issue, but would also require additional topic-alignment constraints to preserve cross-site comparability. We will clarify this distinction and identify a shared-plus-local model,

$$A_m = A + \Delta_m,$$

as an important extension for applications involving more heterogeneous healthcare systems.

A.3 Topic-SCORE preliminaries

A.3.1 Anchor words and regularity assumptions

Let $a_j^\top$ denote the jth row of A and define

$$h_j := \|a_j\|_1 = \sum_{k=1}^{K} A(j,k), \qquad h_{\min} := \min_{1 \le j \le p} h_j, \qquad \bar{h} := \frac{1}{p} \sum_{j=1}^{p} h_j = \frac{K}{p}.$$

Also define $H := \text{diag}(h_1, \ldots, h_p)$ and $\Sigma_A := A^\top H^{-1} A$.

Table 1: How the assumptions are used.

Assumption	Single-site role	Federated role
Anchor words	Guarantees the ideal simplex and justifies vertex hunting.	Same role; used only after the ratio cloud has been constructed.
$h_{\min} \gtrsim p^{-1}$	Gives $M_0(j,j) \asymp h_j$ and the row-wise scale.	Same role; also turns pooled marginal concentration into the $h_j^{-1/2}$ row scale.
Pooled Σ_W and $\lambda_{\min}(\Sigma_A)$	Enter oracle geometry, row scales, and the recovery identity.	Enter the canonical positive direction $\xi_1 = M_0^{1/2}\mathbf{1}$, bounded simplex geometry, and the lower bound.
Client-level $\Sigma_{W,m}$	Not present.	Controls local signal gaps and local projector perturbation.

Definition 2 (Anchor word)**.** Word j is an anchor word if row j of A has exactly one nonzero entry. If that nonzero entry is in column k, then word j is an anchor word for topic k.

Assumption 4 (At least one anchor word per topic)**.** For each $k \in [K]$, there exists at least one anchor word for topic k.

Assumption 5 (No extremely rare words)**.** There exists $c_1 \in (0,1)$ such that

$$h_{\min} \geq c_1 \bar{h} = c_1 \frac{K}{p}.$$

Assumption 6 (Pooled weight regularity and topic overlap)**.** There exists $c_2 \in (0,1)$ such that

$$\lambda_{\min}(\Sigma_W) \geq c_2, \qquad \lambda_{\min}(\Sigma_A) \geq c_2.$$

Remark 3 (Why the federated upper bounds use a weaker topic-overlap condition). Ke and Wang's single-site SVD Topic-SCORE analysis assumes the stronger condition $\min_{k,\ell} \Sigma_A(k,\ell) \geq c_2$ together with $\lambda_{\min}(\Sigma_W) \geq c_2$ and $\lambda_{\min}(\Sigma_A) \geq c_2$; compare their condition (10). In that paper the extra entrywise-overlap lower bound is tied to the Perron-based choice of the first singular direction. In the present federated analysis, the first SCORE direction is recovered canonically from the projected pooled marginal,

$$\xi_1 = M_0^{1/2}\mathbf{1},$$

so the oracle/simplex part uses only the positivity of $w = n_{\text{tot}}^{-1} W\mathbf{1}$ and the invertibility of $A^\top M_0^{-1} A$. Accordingly, the federated upper-bound chain requires only $\lambda_{\min}(\Sigma_W) \geq c_2$ and $\lambda_{\min}(\Sigma_A) \geq c_2$. The lower-bound construction below is carried out on a convenient subclass that in fact also has strictly positive off-diagonal overlap, but that stronger property is no longer part of the theorem interface.

Assumption 7 (Client-level weight richness)**.** There exists $c_{\text{loc}} \in (0,1)$ such that

$$\lambda_{\min}(\Sigma_{W,m}) \geq c_{\text{loc}}, \qquad 1 \leq m \leq M.$$

For two topic matrices with columns normalized in ℓ_1, we measure error by the permutation-invariant loss

$$L(\hat{A}, A) := \min_{\kappa \in S_K} \sum_{k=1}^{K} \|\hat{A}_{\kappa(k)} - A_k\|_1. \tag{6}$$

All row-wise statements below are to be read after the same topic-label permutation.

Remark 4 (Scope and relation to existing theory). The theoretical contribution is the federated interface needed for a Topic-SCORE guarantee. The single-site ratio, simplex, and recovery algebra follows the SVD Topic-SCORE chain of Ke and Wang's SVD Topic-SCORE paper; the new ingredients are the federated projector-averaging step, the empirical-whitening bridge, and recovery of the canonical pooled-marginal direction. The mathematical statements below are written so that this proof chain can be checked independently of implementation details and of the empirical study.

A.3.2 The pooled marginal induces the canonical positive direction

The common signal space is

$$\mathcal{S} := \text{col}(M_0^{-1/2} A).$$

Let

$$w := \frac{1}{n_{\mathrm{tot}}} W\mathbf{1} \in \mathbb{R}^K.$$

Since $D_0\mathbf{1} = n_{\mathrm{tot}} m_0$, we have

$$M_0^{-1/2} A w = M_0^{-1/2} m_0 = M_0^{1/2}\mathbf{1}.$$

Because $\mathbf{1}^\top m_0 = 1$, the vector

$$\xi_1 := M_0^{1/2}\mathbf{1}$$

already has Euclidean norm one. It is strictly positive, belongs to $\mathcal{S}$, and is therefore the natural pooled analogue of the positive SCORE denominator used in single-site Topic-SCORE.

Choose any orthonormal basis $\Xi_\perp \in \mathbb{R}^{p\times(K-1)}$ of $\mathcal{S} \cap \xi_1^\perp$ and write

$$\Xi = [\xi_1, \Xi_\perp].$$

Then there exists a unique nonsingular matrix $V = [V_1, \ldots, V_K] \in \mathbb{R}^{K\times K}$ such that

$$\Xi = M_0^{-1/2} A V, \qquad V_1 = w.$$

Here V_ℓ denotes the ℓth column of V, whereas $V_{k\ell}$ denotes the (k,ℓ) entry. Thus the barycentric weights and simplex geometry are built from the pooled topic-weight mean, not from an arbitrary orthonormal basis of the common projector.

Define the oracle ratio matrix

$$R(j,\ell) = \frac{\Xi_\perp(j,\ell)}{\xi_1(j)}, \qquad 1 \le j \le p, \quad 1 \le \ell \le K-1.$$

For $k \in [K]$, define

$$v_k^\star(\ell) := \frac{V_{k,\ell+1}}{V_{k1}}, \qquad 1 \le \ell \le K-1,$$

and for each word j define

$$\pi_j(k) := \frac{A(j,k)V_{k1}}{\sum_{\ell=1}^K A(j,\ell)V_{\ell 1}}.$$

Then each π_j is a probability vector and

$$r_j = \sum_{k=1}^K \pi_j(k) v_k^\star.$$

If word j is an anchor word for topic k, then $r_j = v_k^\star$. Consequently the rows of R lie in a nondegenerate simplex with vertices $v_1^\star, \ldots, v_K^\star$. Moreover,

$$A\,\mathrm{diag}(w) = M_0^{1/2}\,\mathrm{diag}(\xi_1)\Pi, \qquad \Pi = [\pi_1, \ldots, \pi_p]^\top. \tag{7}$$

Remark 5 (Why an arbitrary orthonormal basis is insufficient). Projector averaging recovers a subspace, not a distinguished first vector inside that subspace. If an arbitrary basis is rotated by a full $K \times K$ orthogonal matrix

$$U = \begin{pmatrix} a & b^\top \\ c & D \end{pmatrix},$$

then a row $\xi_1(j)(1, r_j)$ is transformed into

$$\xi_1(j)(a + b^\top r_j,\ c + D r_j),$$

and the induced ratio map is

$$r_j \mapsto \frac{c + D r_j}{a + b^\top r_j},$$

a projective transformation rather than a harmless $(K-1)$-dimensional rotation. This is why the federated procedure must first recover the canonical positive first direction from the pooled marginal.

A.3.3 Vertex hunting and barycentric recovery

Given empirical ratio points $\hat{r}_j$, the Topic-SCORE simplex step first applies a vertex-hunting routine to the augmented points $\hat{y}_j = (1, \hat{r}_j^\top)^\top$ and obtains vertices $\hat{v}_1, \ldots, \hat{v}_K$. The corresponding unconstrained barycentric coordinates are defined by

$$\hat{Q}\hat{\pi}_j^\star = \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix}, \qquad \hat{Q} = \begin{pmatrix} 1 & \cdots & 1 \\ \hat{v}_1 & \cdots & \hat{v}_K \end{pmatrix}. \tag{8}$$

The simplex-normalized coordinate vector is

$$\hat{\pi}_j = \mathcal{T}(\hat{\pi}_j^\star), \qquad \mathcal{T}(b) = \frac{[b]_+}{\|[b]_+\|_1},$$

with $\mathcal{T}(b) = K^{-1}\mathbf{1}$ if $[b]_+ = 0$. In the federated analysis below, the empirical residual basis is identifiable only up to a $(K-1)$-dimensional rotation; the perturbation proof therefore applies the same barycentric step after aligning the empirical ratio cloud by $O_\perp$.

A.4 Federated Topic-SCORE via projector averaging

A.4.1 Protocol and communicated objects

The client-side spectral step uses the same SVD object as ordinary Topic-SCORE. After the server broadcasts the shared empirical whitening matrix M_d, client m forms

$$D_f^{(m)} := M_d^{-1/2} D^{(m)}$$

and computes its leading K left singular vectors. Equivalently, for eigenvector purposes this is the same as diagonalizing the identity-centered Gram matrix

$$\hat{G}_m^{\mathrm{svd}} := \frac{1}{n_m} M_d^{-1/2} D^{(m)} D^{(m)\top} M_d^{-1/2} - \frac{1}{N} I_p. \tag{9}$$

The subtraction of $(1/N)I_p$ is only a scalar eigenvalue shift, so the eigenspace of (9) is exactly the leading left singular subspace of $D_f^{(m)}$.

Let $\hat{\Xi}^{(m)} \in \mathbb{R}^{p\times K}$ contain the leading K left singular vectors of $D_f^{(m)}$ and define the uploaded projector

$$\hat{P}_m := \hat{\Xi}^{(m)} \hat{\Xi}^{(m)\top}.$$

The server averages

$$\bar{P} := \sum_{m=1}^{M} \alpha_m \hat{P}_m, \qquad \alpha_m \geq 0, \qquad \sum_{m=1}^{M} \alpha_m = 1,$$

and lets $\widehat{P}$ denote the top-K spectral projector of $\bar{P}$.

The empirical canonical first direction is

$$u_d := M_d^{1/2}\mathbf{1}, \qquad \hat{\xi}_1 := \frac{\widehat{P} u_d}{\|\widehat{P} u_d\|_2}.$$

Let

$$\widehat{P}_\perp := \widehat{P} - \hat{\xi}_1 \hat{\xi}_1^\top.$$

Choose any orthonormal basis $\hat{\Xi}_\perp$ of $\mathrm{range}(\widehat{P}_\perp)$, form the empirical ratio matrix

$$\hat{R}(j,\ell) = \frac{\hat{\Xi}_\perp(j,\ell)}{\hat{\xi}_1(j)},$$

apply vertex hunting, recover barycentric coordinates $\hat{\Pi}$, and define

$$\tilde{A}_{\mathrm{fed}} := M_d^{1/2} \,\mathrm{diag}(\hat{\xi}_1)\hat{\Pi}.$$

Column-normalizing $\tilde{A}_{\mathrm{fed}}$ gives $\hat{A}_{\mathrm{fed}}$.

Algorithm 2 Practical federated Topic-SCORE via SVD upload and canonical direction recovery

Require: Local corpora $\{D^{(m)}\}_{m=1}^{M}$, topic number K, weights $\{\alpha_m\}_{m=1}^{M}$, a vertex-hunting routine
Ensure: Topic matrix estimate $\hat{A}_{\mathrm{fed}}$

1: Clients send $\bar{d}^{(m)} = n_m^{-1} D^{(m)}\mathbf{1}$ to the server
2: Server aggregates $\bar{d} = n_{\mathrm{tot}}^{-1} D\mathbf{1}$ and broadcasts $M_d = \mathrm{diag}(\bar{d})$
3: **for** each client $m = 1, \ldots, M$ in parallel **do**
4: Form $D_f^{(m)} = M_d^{-1/2} D^{(m)}$
5: Compute the leading K left singular vectors $\hat{\Xi}^{(m)}$ of $D_f^{(m)}$
6: Upload $\hat{\Xi}^{(m)}$, equivalently $\hat{P}_m = \hat{\Xi}^{(m)}\hat{\Xi}^{(m)\top}$
7: **end for**
8: Server forms $\bar{P} = \sum_{m=1}^{M} \alpha_m \hat{P}_m$ and extracts its top-K projector $\widehat{P}$
9: Recover $\hat{\xi}_1 = \widehat{P} M_d^{1/2}\mathbf{1} / \|\widehat{P} M_d^{1/2}\mathbf{1}\|_2$
10: Set $\widehat{P}_\perp = \widehat{P} - \hat{\xi}_1\hat{\xi}_1^\top$ and choose $\hat{\Xi}_\perp$ spanning $\mathrm{range}(\widehat{P}_\perp)$
11: Form $\hat{R}(j,\ell) = \hat{\Xi}_\perp(j,\ell)/\hat{\xi}_1(j)$
12: Apply vertex hunting to the rows of $\hat{R}$ and obtain $\hat{v}_1, \ldots, \hat{v}_K$
13: **for** $j = 1, \ldots, p$ **do**
14: Solve $\begin{pmatrix} 1 & \cdots & 1 \\ \hat{v}_1 & \cdots & \hat{v}_K \end{pmatrix} \hat{\pi}_j^\star = \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix}$
15: Truncate negative coordinates and ℓ_1-normalize to obtain $\hat{\pi}_j$
16: **end for**
17: Set $\hat{\Pi} = [\hat{\pi}_1, \ldots, \hat{\pi}_p]^\top$, $\tilde{A}_{\mathrm{fed}} = M_d^{1/2}\,\mathrm{diag}(\hat{\xi}_1)\hat{\Pi}$
18: Column-normalize $\tilde{A}_{\mathrm{fed}}$ and output $\hat{A}_{\mathrm{fed}}$

Remark 6 (Computing the practical SVD upload). Lines 1–2 of Algorithm 2 are a marginal-aggregation round used only to construct the shared whitening matrix. After M_d is broadcast, the practical spectral upload is obtained by a standard truncated SVD of $D_f^{(m)} = M_d^{-1/2} D^{(m)}$. Equivalently, this is the eigenspace of the identity-centered Gram matrix (9); the identity shift does not affect eigenvectors. Thus the implementation can use the same sparse-matrix SVD routines as single-site Topic-SCORE. Truncated SVD also provides the canonical best rank-(K) approximation of the local matrix and returns an orthonormal basis, making rotation-invariant projector aggregation straightforward. In contrast, methods that directly estimate local topics, such as LDA or NMF, typically require iterative nonconvex optimization and introduce permutation, scaling, and cross-site topic-alignment ambiguities. More robust or weighted subspace estimators could in principle replace SVD under alternative contamination or heteroscedastic noise assumptions, but they require additional modeling assumptions and tuning parameters. Under the noise model considered in this paper, SVD is therefore the simplest and most direct estimator of the model-implied shared subspace, while also enabling a single-round federated implementation.

Remark 7 (Optional centered local spectral upload). The ordinary-SVD upload in Algorithm 2 is computationally convenient because each client can use a standard truncated SVD of $M_d^{-1/2} D^{(m)}$. In sparse or highly heterogeneous settings, however, its fixed-whitening population target contains the deterministic diagonal term $D_{0,m}^{\mathrm{svd}}$. This term is an artifact of the ordinary-SVD identity shift, not of the federated projector-averaging principle. An optional diagonal-debiased variant lets client m form

$$\widehat{G}_m^{\mathrm{deb}} = n_m^{-1} M_d^{-1/2} D^{(m)} D^{(m)\top} M_d^{-1/2} - N^{-1} M_d^{-1/2} M_d^{(m)} M_d^{-1/2},$$

where $M_d^{(m)} = \mathrm{diag}(n_m^{-1} D^{(m)}\mathbf{1})$, and upload the top-$K$ eigenspace of $\widehat{G}_m^{\mathrm{deb}}$. The server-side projector averaging, canonical direction recovery, ratio construction, vertex hunting, and column normalization are unchanged. This variant preserves the same rank-K spectral communication order but removes the ordinary-SVD diagonal-bias residue from the perturbation radius, replacing it with the corresponding centered-Gram concentration terms. The main text reports the ordinary-SVD implementation, so this centered upload is presented as an optional extension rather than as the default algorithm.

Remark 8 (Communication and computation budget). The non-iterative summary-sharing protocol first sends one p-vector $\bar{d}^{(m)}$ from each client, which is needed to form the common whitening matrix.

After the server broadcasts M_d, each client sends a $p \times K$ orthonormal basis $\hat{\Xi}^{(m)}$; the equivalent projector $\hat{P}_m$ need not be transmitted or stored as a dense $p \times p$ matrix. The server can apply the averaged projector to a vector by

$$\bar{P}x = \sum_{m=1}^{M} \alpha_m \hat{\Xi}^{(m)} \{\hat{\Xi}^{(m)\top} x\},$$

so extracting the top-K eigenspace of $\bar{P}$ can also be done matrix-free. Thus the spectral communication after whitening is $O(pK)$ numbers per client, in addition to the one-time marginal vector. The iterative-free and spectral properties of our method have significant advantages in reducing communication and computation costs over iterative approaches. For a representative iterative Federated LDA procedure with R communication rounds and E local inference iterations per round, site m performs approximately $O(REpn_mK)$ local computation and communicates a $p \times K$ topic matrix or sufficient statistic over R rounds, for $O(RpK)$ communication. In contrast, Federated Topic-SCORE performs only one rank-K truncated SVD of the local $p \times n_m$ matrix, with cost $O(pn_mK)$ up to a constant number of matrix passes, and communicates a p-vector and a $p \times K$ spectral summary, for fixed $O(p + pK)$ communication per site.

When site m receives Δn_m new patients, keeping the current $\widehat{A}$ lets Equation (4) estimate their topic weights locally at $O(pK\Delta n_m)$ cost and no communication. Updating A uses one non-iterative refresh of the enlarged matrix, costing $O\{p(n_m+\Delta n_m)K\}$ locally and $O(p+pK)$ in communication. An iterative update repeats this work over R rounds, costing $O\{REp(n_m + \Delta n_m)K\}$ locally and $O(RpK)$ in communication. The non-iterative design therefore avoids repeated computation and communication as cohorts grow.

Remark 9 (Zero empirical frequencies and the definition of $M_d^{-1/2}$). The theoretical statements are made on the high-probability event on which $M_d(j,j) \asymp M_0(j,j)$ for every word j; on this event $M_d^{-1/2}$ is well defined. An implementation can use the harmless floored version

$$M_{d,\tau} := \mathrm{diag}(\bar{d}_1 \vee \tau, \ldots, \bar{d}_p \vee \tau)$$

with τ smaller than a sufficiently small constant multiple of p^{-1}. Under Assumptions 5–6, the floor is inactive on the perturbative marginal event in Lemma 26. Fix measurable rules for all spectral and vertex-hunting ties; each local fixed-whitening spectral rule depends only on that client's matrix. If a required spectral matrix is undefined, assign the fixed rank-K orthogonal projector $P_{\mathrm{fb}} = \sum_{k=1}^{K} e_k e_k^\top$. To define the estimator on every outcome, if a required inverse does not exist, a normalization denominator vanishes, a canonical coordinate is nonpositive, or vertex hunting fails to return an invertible simplex, set

$$\widehat{A}_{\mathrm{fed}} = p^{-1}\mathbf{1}_p\mathbf{1}_K^\top, \qquad \widehat{w}_i = K^{-1}\mathbf{1}_K.$$

Use the same fixed weight fallback if the weighted least-squares matrix is singular or simplex normalization is undefined, and assign fixed finite values to any still undefined intermediate coordinates. These conventions affect only outcomes outside the good events used in the proofs; on those events all original formulas apply.

A.4.2 Fixed-whitening matrices and the identity-centered SVD bridge

The algorithmic projector is exactly the ordinary-SVD projector uploaded in Algorithm 2. Throughout the main text and the proof, set

$$\hat{G}_m := \frac{1}{n_m} M_d^{-1/2} D^{(m)} D^{(m)\top} M_d^{-1/2} - \frac{1}{N} I_p, \tag{10}$$

so that $\hat{P}_m$ is the top-K spectral projector of the same matrix whose eigenvectors are the leading left singular vectors uploaded by client m. The fixed-whitening comparison matrix is

$$\tilde{G}_m := \frac{1}{n_m} M_0^{-1/2} D^{(m)} D^{(m)\top} M_0^{-1/2} - \frac{1}{N} I_p, \tag{11}$$

and the rank-K common-space target is

$$G_{0,m} := \Big(1 - \frac{1}{N}\Big) \frac{1}{n_m} M_0^{-1/2} D_0^{(m)} D_0^{(m)\top} M_0^{-1/2}. \tag{12}$$

Let $\tilde{P}_m$ be the top-K projector of $\tilde{G}_m$.

The identity shift $N^{-1}I_p$ is algorithmically convenient and is exactly what ordinary SVD implements. It leaves an additional deterministic diagonal term in the fixed-whitening population mean:

$$D_{0,m}^{\mathrm{svd}} := \mathbb{E}[\tilde{G}_m \mid W^{(1:M)}] - G_{0,m} = \frac{1}{N}\left(M_0^{-1/2} M_{0,m} M_0^{-1/2} - I_p\right). \tag{13}$$

This is not a centering correction; it is the explicit price of using the ordinary SVD identity shift rather than a local diagonal debiasing term. We measure its row-wise size by

$$\Delta_{\mathrm{svd}} := \max_{1\le m\le M}\left\{\|D_{0,m}^{\mathrm{svd}}\| + \max_{1\le j\le p} h_j^{-1/2}\|D_{0,m}^{\mathrm{svd}} e_j\|_2\right\}. \tag{14}$$

Under Assumptions 5–6, $\Delta_{\mathrm{svd}} \lesssim \sqrt{p}/N$; this deterministic term is included in the final radius.

Because

$$G_{0,m} = \left(1 - \frac{1}{N}\right) B\Sigma_{W,m}B^\top, \qquad B := M_0^{-1/2}A,$$

all $G_{0,m}$ share the same range $\mathcal{S} = \mathrm{col}(B)$. Let P denote the orthogonal projector onto $\mathcal{S}$.

Lemma 1 (Identity-centered fixed-whitening mean and bridge decomposition). Fix $m \in [M]$ and define

$$\hat{S}_m := \frac{1}{n_m} D^{(m)} D^{(m)\top}, \qquad T_m := M_0^{-1/2}\hat{S}_m M_0^{-1/2} = \tilde{G}_m + \frac{1}{N} I_p,$$

$$R_M := M_d^{-1/2} M_0^{1/2} - I_p.$$

Then

$$\mathbb{E}[\tilde{G}_m \mid W^{(1:M)}] = G_{0,m} + D_{0,m}^{\mathrm{svd}}, \tag{15}$$

and

$$\hat{G}_m - \tilde{G}_m = R_M T_m + T_m R_M + R_M T_m R_M. \tag{16}$$

Consequently,

$$\hat{G}_m - G_{0,m} = \underbrace{\{\tilde{G}_m - \mathbb{E}(\tilde{G}_m \mid W^{(1:M)})\}}_{\text{centered sampling fluctuation}} + D_{0,m}^{\mathrm{svd}} + R_M T_m + T_m R_M + R_M T_m R_M. \tag{17}$$

The proof decomposition above follows the algorithmic identity-centered SVD matrix $\hat{G}_m$ throughout. The term $D_{0,m}^{\mathrm{svd}}$ is a deterministic diagonal perturbation of the common-space signal, and the three terms in (16) are the shared-whitening bridge induced by replacing M_0 with M_d.

Define the pooled marginal deviation

$$\Delta_1(D, D_0) := \max_{1\le j\le p} h_j^{-1}|M_d(j,j) - M_0(j,j)|. \tag{18}$$

A.4.3 Abstract projector averaging

The next results hold for any independent rank-K projectors $\{\mathcal{P}_m\}_{m=1}^M$ with common deterministic target P and deterministic averaging weights $\alpha_m \ge 0$ satisfying $\sum_{m=1}^M \alpha_m = 1$. Set

$$\bar{\mathcal{P}} := \sum_{m=1}^M \alpha_m \mathcal{P}_m, \qquad \nu_m := \mathbb{E}\|\mathcal{P}_m - P\|^2.$$

Assumption 8 (Client independence)**.** The projectors being averaged are mutually independent.

Lemma 2 (Davis–Kahan with a projector signal). Let P be a rank-K projector and let $\bar{\mathcal{P}}$ be symmetric. If $\|\bar{\mathcal{P}} - P\| \le 1/4$, then the top-$K$ spectral projector of $\bar{\mathcal{P}}$ is uniquely defined and satisfies

$$\|\widehat{P} - P\| \le 2\|\bar{\mathcal{P}} - P\|.$$

More generally, the same conclusion holds with another absolute constant whenever $\|\bar{\mathcal{P}} - P\| \le c_{\mathrm{DK}} < 1/2$.

Lemma 3 (Variance-bias decomposition). For deterministic weights α_m satisfying $\sum_{m=1}^M \alpha_m = 1$,

$$\|\bar{\mathcal{P}} - P\| \le \Big\| \sum_{m=1}^M \alpha_m (\mathcal{P}_m - \mathbb{E}\mathcal{P}_m) \Big\| + \Big\| \sum_{m=1}^M \alpha_m (\mathbb{E}\mathcal{P}_m - P) \Big\|.$$

No independence is needed for this deterministic decomposition; independence enters only in the concentration step below.

Lemma 4 (Matrix Bernstein for projector averaging). Let $X_m := \alpha_m(\mathcal{P}_m - \mathbb{E}\mathcal{P}_m)$ and $S := \sum_m X_m$. Then $\|X_m\| \le 2\alpha_m$ and

$$\Big\| \sum_{m=1}^M \mathbb{E}[X_m^2] \Big\| \le 4 \sum_{m=1}^M \alpha_m^2 \nu_m.$$

Hence, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|S\| \le 4\sqrt{2\log\frac{2p}{\delta}} \Big(\sum_{m=1}^M \alpha_m^2 \nu_m \Big)^{1/2} + \frac{4}{3} \log\frac{2p}{\delta} \max_m \alpha_m.$$

Lemma 5 (Second-order spectral-projector bias). Suppose $G_m = G + E_m$ with $\mathbb{E}E_m = 0$, and let P and P_m be the spectral projectors onto the same isolated size-K spectral cluster with eigengap $\gamma > 0$. Assume the target cluster has diameter at most $C_0\gamma$ for a fixed constant C_0; in the applications below, all nonzero population eigenvalues are bounded above and below by constants, so this condition is automatic. Assume further that $\|E_m\| \le \gamma/4$ almost surely. With C depending only on C_0, we have

$$\|\mathbb{E}P_m - P\| \le \frac{C}{\gamma^2} \mathbb{E}\|E_m\|^2.$$

If, in addition, $G \succeq 0$ has rank K, P is its nonzero-eigenspace projector, and the remaining spectrum is exactly zero, then for every $j \in [p]$,

$$\|(\mathbb{E}P_m - P)e_j\|_2 \le \frac{C}{\gamma^2} \left[(\mathbb{E}\|E_m\|^2)^{1/2} (\mathbb{E}\|E_m e_j\|_2^2)^{1/2} + \|Pe_j\|_2 \, \mathbb{E}\|E_m\|^2 \right]. \tag{19}$$

Remark 10 (High-probability versions of the second-order bias bound). Lemma 5 is stated with an almost-sure small-perturbation condition. When this condition holds only on $\mathcal{E}_m = \{\|E_m\| \le \gamma/4\}$, apply the Riesz expansion on $\mathcal{E}_m$. Since

$$\mathbb{E}[\mathcal{L}_G(E_m)\mathbf{1}_{\mathcal{E}_m}] = -\mathbb{E}[\mathcal{L}_G(E_m)\mathbf{1}_{\mathcal{E}_m^c}],$$

the truncation argument also controls the linear tail term, in addition to the projector contribution on $\mathcal{E}_m^c$. In the multinomial application, the relevant tail terms are bounded by Lemma 21, while the deterministic ordinary-SVD bias is treated separately, as in the proof of Lemma 8. The columnwise display is intentionally stated only for the rank-K signal case. This is the case used for the fixed-whitening population matrices $G_{0,m}$; for a general isolated cluster we use only the operator bias bound and the operator remainder in Lemma 14.

Remark 11 (Why the columnwise bias bound contains $\|Pe_j\|_2$). The columnwise part of Lemma 5 is deliberately stated only for the rank-K signal case used in Lemma 8. In that case the non-target spectrum is zero, so the resolvent on $P^\perp$ is scalar and the columnwise remainder can retain the factor $\|E_m e_j\|_2$. The term $\|Pe_j\|_2 \, \mathbb{E}\|E_m\|^2$ is essential: a second-order resolvent remainder can act first on the signal component of e_j, and that component is controlled in this paper by the oracle row-scale bound $\|Pe_j\|_2 \lesssim \sqrt{h_j}$.

Theorem 4 (Federated projector averaging bound)**.** *Under Assumption 8 and deterministic weights $\alpha_m \ge 0$ with $\sum_m \alpha_m = 1$, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\bar{\mathcal{P}} - P\| \le \Big\| \sum_{m=1}^M \alpha_m (\mathbb{E}\mathcal{P}_m - P) \Big\| + 4\sqrt{2\log\frac{2p}{\delta}} \Big(\sum_{m=1}^M \alpha_m^2 \nu_m \Big)^{1/2} + \frac{4}{3} \log\frac{2p}{\delta} \max_m \alpha_m. \tag{20}$$

If the right-hand side is at most $1/4$, then the top-K spectral projector $\widehat{P}$ of $\bar{\mathcal{P}}$ is well defined and obeys $\|\widehat{P} - P\| \le 2\|\bar{\mathcal{P}} - P\|$.

Remark 12 (Which projectors are averaged in the theorem). The practical projectors $\hat{P}_m$ are coupled through the shared random matrix M_d. The theorem is applied first to the fixed-whitening ordinary-SVD projectors $\tilde{P}_m$, which are independent conditional on $W^{(1:M)}$. Their population target differs from $G_{0,m}$ by the deterministic diagonal term $D^{\mathrm{svd}}_{0,m}$, and the empirical-whitening bridge $\hat{P}_m - \tilde{P}_m$ is handled deterministically.

Lemma 6 (Vector Bernstein for projector columns). Fix $j \in [p]$ and define

$$r_j := \Big(\sum_{m=1}^{M} \alpha_m \mathcal{P}_m - P\Big) e_j.$$

Let $\tau^2_{m,j} := \mathbb{E}\|(\mathcal{P}_m - P)e_j\|_2^2$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\|r_j - \mathbb{E}r_j\|_2 \le \sqrt{2\log\frac{2}{\delta}}\Big(\sum_{m=1}^{M} \alpha_m^2 \tau^2_{m,j}\Big)^{1/2} + \frac{4}{3}\log\frac{2}{\delta}\max_m \alpha_m.$$

After a union bound over j, the row-normalized version of this display contributes the conservative range term $h_{\min}^{-1/2}\ell_{n,M}(\delta)/M \lesssim \sqrt{p}\,\ell_{n,M}(\delta)/M$ used in Theorem 5.

A.4.4 Specialization to federated Topic-SCORE

For readability we now specialize to equal client sizes and equal averaging weights,

$$n_1 = \cdots = n_M = n, \qquad \alpha_m = 1/M.$$

The abstract projector-averaging results above continue to hold for general $\{n_m, \alpha_m\}$; the same variance–bias decomposition then carries the weights α_m and the local sample sizes n_m explicitly. We do not expand that notation below. Define

$$\beta_n := \begin{cases} 1 + \min\{p/N,\ p^2/N^{3/2}\}, & n \ge \max\{Np^2,\ p^3,\ N^2p^5\}, \\ 1 + p^2/N^{3/2}, & n < \max\{Np^2,\ p^3,\ N^2p^5\}. \end{cases}$$

Also set

$$\ell_{n,M}(\delta) := \log\frac{2pM}{\delta},$$

$$\vartheta_n(\delta) := \beta_n\sqrt{\frac{p\ell_{n,M}(\delta)}{Nn}} + \beta_n^2\frac{p\ell_{n,M}(\delta)}{Nn} + \sqrt{\beta_n p}\,\frac{p\ell_{n,M}(\delta)}{Nn} + \beta_n\sqrt{\frac{p}{N}}\,\frac{\ell_{n,M}(\delta)}{n},$$

$$\eta^{\mathrm{marg}}_{n,M}(\delta) := \sqrt{\frac{p\ell_{n,M}(\delta)}{NnM}} + \frac{p\ell_{n,M}(\delta)}{NnM}.$$

Define also the deterministic ordinary-SVD diagonal-bias radius

$$\zeta^{\mathrm{svd}}_N := \frac{\sqrt{p}}{N}.$$

The linear-in-$\ell_{n,M}$ terms in $\vartheta_n(\delta)$ retain the range contributions in the document-level Bernstein bounds. The term $\sqrt{\beta_n p}\,p\ell/(Nn)$ controls higher moments of the product term in the column decomposition. The term $\beta_n\sqrt{p/N}\,\ell/n$ is used in both the Frobenius/operator and column Bernstein bounds. These terms are kept explicitly in the local concentration interface.

Lemma 7 (Fixed-design heterogeneity envelope). Assume Assumptions 5–6. For every client m and document i, let

$$\pi_i^{(m)} := Aw_i^{(m)}.$$

Then, with a constant depending only on the regularity constants and fixed K,

$$\max_{m,i}\max_{1\le j\le p}\frac{\pi_i^{(m)}(j)}{m_0(j)} \le C, \qquad \max_{m,i}\sum_{j=1}^{p}\frac{\pi_i^{(m)}(j)}{m_0(j)} \le Cp, \qquad \max_{m,i}\sum_{j=1}^{p}\frac{\{\pi_i^{(m)}(j)\}^2}{m_0(j)} \le C.$$

Since $\beta_n \ge 1$, these bounds imply the envelope used in the multinomial concentration lemmas with $C\beta_n$, $C\beta_n p$, and $C\beta_n$ on the right-hand sides. Thus the envelope is not an additional assumption in the regular theorem class.

Proof. Let $\bar{w} = n_{\mathrm{tot}}^{-1} W \mathbf{1}$. From $\lambda_{\min}(\Sigma_W) \geq c_2$ and $0 \leq w_i(k) \leq 1$,

$$\bar{w}_k = \frac{1}{n_{\mathrm{tot}}} \sum_i w_i(k) \geq \frac{1}{n_{\mathrm{tot}}} \sum_i w_i(k)^2 = \Sigma_W(k,k) \geq c_2.$$

Therefore $m_0(j) = a_j^\top \bar{w} \geq c_2 h_j$, while $\pi_i^{(m)}(j) = a_j^\top w_i^{(m)} \leq h_j$. Hence $\pi_i^{(m)}(j)/m_0(j) \leq c_2^{-1}$ uniformly. Summing this inequality over j gives the second display. Finally,

$$\sum_j \frac{\{\pi_i^{(m)}(j)\}^2}{m_0(j)} \leq \Big(\max_j \frac{\pi_i^{(m)}(j)}{m_0(j)}\Big) \sum_j \pi_i^{(m)}(j) \leq C,$$

because each $\pi_i^{(m)}$ is a probability vector. This proves the lemma. □

Remark 13 (Unequal sample sizes across sites). We focused on a setting with equal site size only to simplify the main theorem. For unequal site sizes, a natural choice of weights is $\alpha_m = \frac{n_m}{n_{\mathrm{tot}}}$. With this choice, the leading variance term is $O\left(\sqrt{\frac{p\log(pM/\delta)}{N n_{\mathrm{tot}}}}\right)$, up to the same sparsity factor as in the current theorem. This is the same order as the bound in Theorem 1, because when $n_m = n$ for all sites, $n_{\mathrm{tot}} = nM$ and the term becomes $O\left(\sqrt{\frac{p\log(pM/\delta)}{NnM}}\right)$. The corresponding proof requires only bookkeeping changes. We replace $1/M$ by α_m, use n_m in each site-specific concentration bound, and apply the weighted matrix and vector Bernstein inequalities. The local bias becomes a weighted sum of the site-specific bias terms, and the range term depends on $\max_m \alpha_m$. The remaining projector recovery and Topic-SCORE arguments are unchanged.

Remark 14 (Origin and interpretation of the heavy-tail coefficient). The coefficient β_n is inherited from the single-site SVD Topic-SCORE analysis of Ke and Wang's SVD Topic-SCORE paper. It packages the extra Bernstein/Freedman range terms that arise when multinomial word-count vectors are sparse or heavy-tailed after whitening. In the generic branch, $N \gtrsim p^{4/3}$ ensures $\beta_n = O(1)$. In the first branch, where the displayed additional condition on n holds, $N \gtrsim p$ suffices. In either case the variance component has the familiar sample-size scaling. When $N \ll p$, rare word-level increments have larger bounded range after whitening and β_n grows, reflecting the same heavy-tail correction as in the row-wise singular-vector theory for single-site Topic-SCORE. Lemma 7 is not an algorithmic requirement; it records the deterministic envelope under which the multinomial moment bounds in Appendix A.7 are uniform over clients and documents.

Lemma 8 (Single-client projector bounds for the practical SVD upload). Assume Assumptions 5–7 and

$$\beta_n^2 \frac{p^2}{Nn} \leq c_0, \qquad \zeta_N^{\mathrm{svd}} \leq c_0$$

for a sufficiently small constant $c_0 > 0$. For each client m, let $\gamma_m := \lambda_K(G_{0,m})$. Then $\gamma_m \geq c_{\mathrm{gap}} > 0$ uniformly in m. Moreover, for the fixed-whitening ordinary-SVD projector $\tilde{P}_m$,

$$\mathbb{E}\|(\tilde{P}_m - P)e_j\|_2^2 \leq C h_j \left(\beta_n^2 \frac{p}{Nn} + (\zeta_N^{\mathrm{svd}})^2\right), \qquad \mathbb{E}\|\tilde{P}_m - P\|^2 \leq C\left(\beta_n^2 \frac{p}{Nn} + (\zeta_N^{\mathrm{svd}})^2\right),$$

and

$$\|(\mathbb{E}\tilde{P}_m - P)e_j\|_2 \leq C\sqrt{h_j}\left(\beta_n^2 \frac{p}{Nn} + \zeta_N^{\mathrm{svd}}\right), \qquad \|\mathbb{E}\tilde{P}_m - P\| \leq C\left(\beta_n^2 \frac{p}{Nn} + \zeta_N^{\mathrm{svd}}\right).$$

For later use, define the client-level good events

$$\mathcal{E}_{\mathrm{loc},m} := \{\|\tilde{G}_m - G_{0,m}\| \leq c_{\mathrm{gap}}/8\},$$

$$\mathcal{E}_{\mathrm{col},m}(\delta) := \Big\{\max_{1\leq j\leq p} h_j^{-1/2}\|(\tilde{G}_m - G_{0,m})e_j\|_2 \leq C_{\mathrm{col}}\{\vartheta_n(\delta) + \zeta_N^{\mathrm{svd}}\}\Big\},$$

and

$$\mathcal{E}_{\mathrm{bridge},m} := \{\Delta_1(D, D_0) \leq c_{\mathrm{br}},\ \|\hat{G}_m - \tilde{G}_m\| \leq C_0\Delta_1(D, D_0)\},$$

where $c_{\mathrm{br}} > 0$ is chosen sufficiently small and $C_{\mathrm{col}}, C_0 > 0$ are absolute constants.

On $\mathcal{E}_{\mathrm{loc},m} \cap \mathcal{E}_{\mathrm{bridge},m}$,

$$\|(\hat{P}_m - \tilde{P}_m)e_j\|_2 \le C\sqrt{h_j}\Delta_1(D,D_0) + C\Delta_1(D,D_0)\|(\tilde{G}_m - G_{0,m})e_j\|_2 + C\Delta_1(D,D_0)^2\{1 + \|\tilde{G}_m - G_{0,m}\|\}^2 + C\frac{\Delta_1(D,D_0)}{N}\|e_j\|_2,$$

and

$$\|\hat{P}_m - \tilde{P}_m\| \le C\Delta_1(D,D_0) + C\Delta_1(D,D_0)\|\tilde{G}_m - G_{0,m}\|.$$

Consequently, on $\mathcal{E}_{\mathrm{loc},m} \cap \mathcal{E}_{\mathrm{col},m}(\delta) \cap \mathcal{E}_{\mathrm{bridge},m}$,

$$\|\hat{P}_m - \tilde{P}_m\| \le C\Delta_1(D,D_0),$$

and

$$\max_{1\le j\le p} h_j^{-1/2}\|(\hat{P}_m - \tilde{P}_m)e_j\|_2 \le C\Delta_1(D,D_0)\{1+\vartheta_n(\delta)+\zeta_N^{\mathrm{svd}}\} + C\sqrt{p}\,\Delta_1(D,D_0)^2 + C\frac{\sqrt{p}}{N}\Delta_1(D,D_0).$$

Corollary 1 (Single-client projector operator-norm bound)**.** *Under the assumptions of Lemma 8, for every client m and every $\delta \in (0,1)$,*

$$\|\hat{P}_m - P\| \le C\Bigg[\beta_n\sqrt{\frac{p\log(2p/\delta)}{Nn}} + \beta_n^2\frac{p\log(2p/\delta)}{Nn} + \sqrt{\beta_n p}\,\frac{p\log(2p/\delta)}{Nn} + \beta_n\sqrt{\frac{p}{N}}\,\frac{\log(2p/\delta)}{n} + \Delta_1(D,D_0) + \zeta_N^{\mathrm{svd}}\Bigg]$$

with probability at least $1-\delta$.

Lemma 9 (Projector-signal row extraction). Let $P = \Xi\Xi^\top$ be a rank-K projector satisfying $\|e_j^\top \Xi\|_2 \le C\sqrt{h_j}$ for all j. Let $\bar{P} = P + E$ be symmetric, and let $\widehat{P}$ be the top-K spectral projector of $\bar{P}$. If $\|E\| \le c$ for a sufficiently small constant $c > 0$, then

$$\|\widehat{P} - P\| + \max_{1\le j\le p} h_j^{-1/2}\|(\widehat{P} - P)e_j\|_2 \le C\Big[\|E\| + \max_{1\le j\le p} h_j^{-1/2}\|Ee_j\|_2\Big].$$

Lemma 10 (Canonical first-direction recovery). Assume

$$\Delta_1(D,D_0) + \|\widehat{P} - P\| \le c$$

for a sufficiently small constant $c > 0$. Define $u_d = M_d^{1/2}\mathbf{1}$ and $\hat{\xi}_1 = \widehat{P}u_d/\|\widehat{P}u_d\|_2$. Then

$$\max_{1\le j\le p} h_j^{-1/2}|\hat{\xi}_1(j) - \xi_1(j)| \le C\left[\Delta_1(D,D_0) + \max_{1\le j\le p} h_j^{-1/2}\|(\widehat{P} - P)e_j\|_2 + \|\widehat{P} - P\|\right].$$

Lemma 11 (Residual basis recovery). Let $P_\perp = P - \xi_1\xi_1^\top$ and $\widehat{P}_\perp = \widehat{P} - \hat{\xi}_1\hat{\xi}_1^\top$. Choose orthonormal bases $\Xi_\perp$ and $\hat{\Xi}_\perp$ of their ranges. If the right-hand side in Lemma 10 is sufficiently small, then there exists an orthogonal matrix $O_\perp \in \mathbb{R}^{(K-1)\times(K-1)}$ such that

$$\max_{1\le j\le p} h_j^{-1/2}\|e_j^\top(\hat{\Xi}_\perp O_\perp - \Xi_\perp)\|_2 \le C\left[\Delta_1(D,D_0) + \max_{1\le j\le p} h_j^{-1/2}\|(\widehat{P} - P)e_j\|_2 + \|\widehat{P} - P\|\right].$$

Define the deterministic perturbative radius

$$\begin{aligned}\varepsilon_{n,M}(\delta) := {}& \beta_n\sqrt{\frac{p\ell_{n,M}(\delta)}{NnM}} + \beta_n^2\frac{p\ell_{n,M}(\delta)}{Nn} \\ & + \eta_{n,M}^{\mathrm{marg}}(\delta)\{1 + \vartheta_n(\delta) + \zeta_N^{\mathrm{svd}}\} + \sqrt{p}\,\{\eta_{n,M}^{\mathrm{marg}}(\delta)\}^2 + \zeta_N^{\mathrm{svd}} \\ & + \frac{\sqrt{p}\,\ell_{n,M}(\delta)}{M} + \Big(\beta_n^2\frac{p}{Nn}\Big)^2.\end{aligned} \tag{21}$$

The factor $\eta^{\mathrm{marg}}_{n,M}(\delta)\vartheta_n(\delta)$ is the row-wise bridge interaction created by the term $\Delta_1(D, D_0)\|(\tilde{G}_m - G_{0,m})e_j\|_2$ in Lemma 8. The factor $\sqrt{p}\{\eta^{\mathrm{marg}}_{n,M}(\delta)\}^2$ is the row-normalized quadratic projector-remainder term in the bridge step, and the small term $\sqrt{p}\,\eta^{\mathrm{marg}}_{n,M}(\delta)/N$ is absorbed by the displayed bridge radius under the stated perturbative regime. The deterministic term ζ^{svd}_N is the price of using the ordinary-SVD identity shift: it controls the population diagonal perturbation $D^{\mathrm{svd}}_{0,m}$ in (13). The factor $\sqrt{p}\,\ell_{n,M}(\delta)/M$ is the conservative row-wise Bernstein range term obtained after dividing the vector-Bernstein bound by $\sqrt{h_j}$ and using $h_{\min} \asymp p^{-1}$. The final square term $(\beta_n^2 p/(Nn))^2$ is retained only as a harmless higher-order Riesz-remainder term from the fixed-whitening perturbation expansion; under the perturbative scaling it is absorbed by the displayed first-order local bias term, but keeping it visible makes the expansion auditable. The range term can be improved to $\ell_{n,M}(\delta)/M$ if one proves a localized almost-sure projector-column range bound, but the displayed form is the unconditional row-wise consequence used here.

Theorem 5 (Federated row-wise canonical basis error)**.** *Assume Assumptions 5–7, $n_m = n$, $\alpha_m = 1/M$, and*

$$\beta_n^2 \frac{p^2}{Nn} \le c_0.$$

Suppose also that the local fixed-whitening perturbation radius and the final deterministic radius satisfy

$$\vartheta_n(\delta) \le c_{\mathrm{pert}}, \qquad \zeta^{\mathrm{svd}}_N \le c_{\mathrm{diag}}, \qquad \varepsilon_{n,M}(\delta) \le c_*,$$

for sufficiently small constants $c_{\mathrm{pert}}, c_{\mathrm{diag}}, c_* > 0$. Then, with probability at least $1 - \delta$, there exists an orthogonal matrix $O_\perp \in \mathbb{R}^{(K-1)\times(K-1)}$ such that

$$\max_{1\le j\le p} h_j^{-1/2} \left\| e_j^\top \left([\hat{\xi}_1, \hat{\Xi}_\perp O_\perp] - [\xi_1, \Xi_\perp] \right) \right\|_2 \le C\varepsilon_{n,M}(\delta). \tag{22}$$

Consequently,

$$\max_{1\le j\le p} \|O_\perp^\top \hat{r}_j - r_j\|_2 \le C\varepsilon_{n,M}(\delta).$$

Remark 15 (The perturbative assumption is deterministic). The smallness conditions in Theorem 5 are deterministic. The condition $\vartheta_n(\delta) \le c_{\mathrm{pert}}$ guarantees the clientwise local spectral perturbation event used before projector averaging, while $\zeta^{\mathrm{svd}}_N \le c_{\mathrm{diag}}$ keeps the population diagonal SVD-bias perturbation local and $\varepsilon_{n,M}(\delta) \le c_*$ keeps the averaged projector and subsequent SCORE ratios perturbative. The radius $\varepsilon_{n,M}$ is stated after replacing the random marginal deviation $\Delta_1(D, D_0)$ by the concentration bound from Lemma 26, inserting the local Gram-column radius from Lemma 25 into the bridge term, adding the deterministic ordinary-SVD diagonal-bias radius ζ^{svd}_N, adding the row-normalized quadratic bridge remainder $\sqrt{p}\{\eta^{\mathrm{marg}}_{n,M}(\delta)\}^2$, and retaining the conservative row-wise Bernstein range term $\sqrt{p}\,\ell_{n,M}(\delta)/M$. Thus the theorem does not assume a random event in its hypotheses.

Corollary 2 (Federated projector operator-norm bound)**.** *Under the assumptions of Theorem 5, with probability at least $1 - \delta$,*

$$\|\widehat{P} - P\| + \max_{1\le j\le p} h_j^{-1/2} \|(\widehat{P} - P)e_j\|_2 \le C\varepsilon_{n,M}(\delta).$$

A.4.5 From ratios to topics

The loss $L(\hat{A}, A)$ is defined in (6); all topic-matrix bounds below are understood after the minimizing common permutation of topic labels.

We use the pooled marginal deviation $\Delta_1(D, D_0)$ defined in (18). Define

$$\Delta_2 := \max_{1\le j\le p} h_j^{-1/2} \left\| e_j^\top \left([\hat{\xi}_1, \hat{\Xi}_\perp O_\perp] - [\xi_1, \Xi_\perp] \right) \right\|_2, \tag{23}$$

$$\mathrm{Err}_{\mathrm{VH}}(O_\perp) := \min_{\kappa\in S_K} \max_{1\le k\le K} \|O_\perp^\top \hat{v}_k - v^\star_{\kappa(k)}\|_2. \tag{24}$$

Condition 1 (Rotation-equivariant vertex-hunting stability)**.** Suppose the population ratio cloud lies in a nondegenerate simplex with vertices $v_1^\star, \ldots, v_K^\star$. If there exists an orthogonal matrix

$O \in \mathbb{R}^{(K-1)\times(K-1)}$ and a perturbation radius $\varepsilon_{\mathrm{VH}} \le c_{\mathrm{VH}}$, for a sufficiently small constant depending only on the oracle simplex geometry and fixed K, such that the observed ratio points $\hat{r}_1, \dots, \hat{r}_p$ obey

$$\max_{1\le j\le p} \|O^\top \hat{r}_j - r_j\|_2 \le \varepsilon_{\mathrm{VH}},$$

then the vertex-hunting routine returns vertices $\hat{v}_1, \dots, \hat{v}_K$ such that, up to a permutation of $[K]$,

$$\max_{1\le k\le K} \|O^\top \hat{v}_k - v_k^\star\|_2 \le C_{\mathrm{VH}}\varepsilon_{\mathrm{VH}}.$$

Euclidean vertex-hunting methods such as successive projection and simplex vertex search satisfy this interface when their non-rotated stability theorem holds, because distances, affine independence, and simplex volumes are invariant under a common orthogonal transformation. Under Assumption 4 and the oracle nondegeneracy in Lemma 38, this condition is the standard Topic-SCORE vertex-hunting stability requirement.

Lemma 12 (A deterministic margin for augmented simplex projection). Let $Q = [q_1, \dots, q_K] \in \mathbb{R}^{K\times K}$ be the augmented vertex matrix with $q_k = (1, (v_k^\star)^\top)^\top$. Assume $\sigma_{\min}(Q) \ge c_Q > 0$ and $\max_k \|q_k\|_2 \le C_Q$. For any $S \subsetneq [K]$, let Π_S be the orthogonal projector onto $\mathrm{span}\{q_k : k \in S\}$, with $\Pi_\varnothing = 0$, and set

$$f_S(y) := \|(I - \Pi_S)y\|_2, \qquad R_S := \max_{k\notin S} f_S(q_k), \qquad V_S := \{q_k : k \notin S\}.$$

Then, for every $y = Q\pi$ with $\pi \in \Delta^{K-1}$,

$$\mathrm{dist}(y, V_S) \le \frac{4C_Q^2}{c_Q^2}\big\{R_S - f_S(y)\big\}.$$

In particular, for every $t \ge 0$,

$$R_S - f_S(y) \le t \quad\Longrightarrow\quad \mathrm{dist}(y, V_S) \le Ct, \qquad C = \frac{4C_Q^2}{c_Q^2}.$$

Proof. Fix $S \subsetneq [K]$, write $J = [K] \setminus S$, and set $r_k = (I - \Pi_S)q_k$ for all $k \in [K]$. Thus $r_k = 0$ for $k \in S$. For any $u \in \mathbb{R}^{|J|}$, the singular-value assumption gives

$$\|(I - \Pi_S)Q_J u\|_2 = \min_{v\in\mathbb{R}^{|S|}} \|Q_J u - Q_S v\|_2 \ge c_Q\|u\|_2.$$

Consequently,

$$\|r_k\|_2 \ge c_Q \quad (k \in J), \qquad \|r_i - r_j\|_2 \ge \sqrt{2}\, c_Q \quad (i, j \in J,\ i \ne j).$$

In particular, $c_Q \le R_S \le C_Q$.

Fix $\pi \in \Delta^{K-1}$ and define

$$a := \sum_{k\in S} \pi_k, \qquad s := \sum_{k\in J} \pi_k = 1 - a, \qquad m := \max_{k\in J} \pi_k.$$

By convexity,

$$f_S(y) = \left\|\sum_{k=1}^K \pi_k r_k\right\|_2 \le R_S.$$

The squared-norm variance identity yields

$$\begin{aligned} d &:= R_S^2 - \left\| \sum_{k=1}^{K} \pi_k r_k \right\|_2^2 \\ &= \sum_{k=1}^{K} \pi_k \big(R_S^2 - \|r_k\|_2^2\big) + \frac{1}{2} \sum_{i,j=1}^{K} \pi_i \pi_j \|r_i - r_j\|_2^2 \\ &\geq a R_S^2 + c_Q^2 a s + c_Q^2 \left(s^2 - \sum_{k \in J} \pi_k^2 \right) \\ &= a R_S^2 + c_Q^2 \left(s - \sum_{k \in J} \pi_k^2 \right) \\ &\geq c_Q^2 \left(1 - \sum_{k \in J} \pi_k^2 \right) \\ &\geq c_Q^2 (1 - m). \end{aligned}$$

The first inequality uses the contributions from S in the first sum, the S–J cross terms, and the pairs of distinct indices in J. The last inequality follows from $\sum_{k \in J} \pi_k^2 \leq ms \leq m$.

Choose $k_* \in J$ with $\pi_{k_*} = m$. Since $\|q_k - q_{k_*}\|_2 \leq 2C_Q$, we obtain

$$\begin{aligned} \operatorname{dist}(y, V_S) &\leq \|y - q_{k_*}\|_2 \\ &= \left\| \sum_{k \neq k_*} \pi_k (q_k - q_{k_*}) \right\|_2 \\ &\leq 2C_Q (1 - m) \\ &\leq \frac{2C_Q}{c_Q^2} d. \end{aligned}$$

Finally,

$$d = \big\{R_S - f_S(y)\big\}\big\{R_S + f_S(y)\big\} \leq 2C_Q \big\{R_S - f_S(y)\big\}.$$

Combining the last two displays proves the claim. □

Proposition 2 (A sufficient condition for successive projection)**.** *Suppose the oracle ratio cloud is contained in a nondegenerate simplex $S(v_1^\star, \ldots, v_K^\star)$ and, for each vertex $v_k^\star$, at least one oracle point is exactly equal to that vertex. Let $\sigma_{\min}$ denote the smallest singular value of the augmented vertex matrix*

$$Q = \begin{pmatrix} 1 & \cdots & 1 \\ v_1^\star & \cdots & v_K^\star \end{pmatrix},$$

and assume $\sigma_{\min} \geq c_Q > 0$ and $\max_k \|v_k^\star\| \leq C_Q$. If the observed points satisfy

$$\max_j \|\hat{r}_j - r_j\|_2 \leq \delta$$

with $\delta \leq c(c_Q, C_Q, K)$, then the augmented successive-projection routine—that is, successive projection applied to

$$\hat{y}_j = (1, \hat{r}_j^\top)^\top \in \mathbb{R}^K$$

and returning the last $K - 1$ coordinates of the selected augmented points—returns vertices $\hat{v}_1, \ldots, \hat{v}_K$ such that, after a permutation,

$$\max_k \|\hat{v}_k - v_k^\star\|_2 \leq C(c_Q, C_Q, K)\delta.$$

Consequently successive projection satisfies Condition 1; the same conclusion after a common orthogonal rotation follows by applying the proposition to $O^\top \hat{r}_j$.

Proof. We give the deterministic perturbation argument in the standard robust-SP form. The routine analyzed here is the augmented SP routine stated above; ordinary SP applied directly to the unaugmented ratio points is not the object used in this proposition. Work with augmented points

$$y_j^\star = \begin{pmatrix} 1 \\ r_j \end{pmatrix}, \qquad \hat{y}_j = \begin{pmatrix} 1 \\ \hat{r}_j \end{pmatrix}, \qquad q_k = \begin{pmatrix} 1 \\ v_k^\star \end{pmatrix}.$$

Then $y_j^\star = Q\pi_j$ with $\pi_j \in \Delta^{K-1}$, and for every k there is an anchor index j_k such that $y_{j_k}^\star = q_k$. The assumptions imply that every submatrix of Q formed by a subset of columns is uniformly well-conditioned and that $\|q_k\| \le C$.

For $S \subset [K]$, let Π_S be the orthogonal projector onto $\mathrm{span}\{q_k : k \in S\}$, with $\Pi_\varnothing = 0$, and define $f_S(y) = \|(I - \Pi_S)y\|_2$. If $S \neq [K]$, then for every $y = Q\pi$ in the oracle cloud,

$$f_S(y) = \left\| (I - \Pi_S) \sum_{k \notin S} \pi_k q_k \right\|_2 \le \max_{k \notin S} f_S(q_k) =: R_S,$$

by convexity of the norm. Also $R_S \ge c > 0$, since otherwise a column q_k, $k \notin S$, would be within distance $o(1)$ of the span of the selected columns, contradicting the lower singular-value bound for Q. Let $V_S := \{q_k : k \notin S\}$ denote the set of unselected oracle vertices.

By Lemma 12, for every $t \ge 0$,

$$R_S - f_S(y) \le t \quad \Longrightarrow \quad \mathrm{dist}(y, V_S) \le Ct, \tag{25}$$

where C depends only on the lower singular-value bound and the upper bound on the augmented vertex norms.

We now prove the SP guarantee by induction. Suppose that after t steps the selected noisy augmented columns are within $C_t\delta$ of distinct oracle vertices indexed by a set S_t of size t. Uniform conditioning of the selected oracle columns implies that the projector $\widehat{\Pi}_t$ onto the span of the selected noisy columns satisfies

$$\|\widehat{\Pi}_t - \Pi_{S_t}\| \le C_t\delta. \tag{26}$$

Hence, uniformly over all observed points,

$$\left| \, \|(I - \widehat{\Pi}_t)\hat{y}_j\|_2 - f_{S_t}(y_j^\star) \, \right| \le \|\hat{y}_j - y_j^\star\|_2 + \|\widehat{\Pi}_t - \Pi_{S_t}\| \, \|y_j^\star\|_2 \le C_t\delta.$$

Since an anchor point for every remaining vertex is present, a point attaining R_{S_t} exists in the population cloud. The point selected by SP at step $t+1$ therefore satisfies

$$f_{S_t}(y_{j_{t+1}}^\star) \ge R_{S_t} - C_t\delta.$$

By the margin implication (25), $y_{j_{t+1}}^\star$ lies within $C_t\delta$ of a new vertex $q_{k_{t+1}}$ with $k_{t+1} \notin S_t$. Adding the observation error gives $\|\hat{y}_{j_{t+1}} - q_{k_{t+1}}\| \le C_{t+1}\delta$. This closes the induction; the base case is the same argument with $S_0 = \varnothing$.

After K steps, the selected augmented noisy points can be matched to $q_1, \ldots, q_K$ within $C\delta$. Dropping the first coordinate gives the displayed vertex bound. The argument uses only Euclidean distances and orthogonal projections, so applying it to the rotated cloud $O^\top \hat{r}_j$ proves the rotation-equivariant form. □

Remark 16 (Concrete vertex-hunting routines). Condition 1 is the same stability interface used in the single-site Topic-SCORE proof. The successive-projection routine satisfies this interface under the anchor-word condition and a nondegenerate simplex, and sketched vertex search satisfies it under the standard cluster-denoising regularity conditions; see the vertex-hunting appendix of the single-site Topic-SCORE analysis. The present topic-matrix theorem is conditional on this vertex-hunting stability: the federated contribution is the projector-to-canonical-basis step before vertex hunting, while the final simplex estimation module can be replaced by any routine satisfying Condition 1.

Lemma 13 (Deterministic Topic-SCORE propagation). Assume the oracle simplex is nondegenerate. If

$$\Delta_1(D, D_0) \le c, \qquad \Delta_2 \le c, \qquad \mathrm{Err}_{\mathrm{VH}}(O_\perp) \le c,$$

for a sufficiently small constant $c > 0$, then, up to a common permutation of topic labels,

$$\max_{1 \le j \le p} \frac{\|\hat{a}_j - a_j\|_1}{h_j} \le C\{\Delta_1(D, D_0) + \Delta_2 + \mathrm{Err}_{\mathrm{VH}}(O_\perp)\}. \tag{27}$$

Consequently,

$$L(\hat{A}_{\mathrm{fed}}, A) \le C\{\Delta_1(D, D_0) + \Delta_2 + \mathrm{Err}_{\mathrm{VH}}(O_\perp)\}.$$

Theorem 6 (Federated Topic-SCORE topic-matrix error, conditional on rotation-equivariant vertex hunting)**.** *Assume Assumption 4, the conditions of Theorem 5, and suppose the vertex-hunting routine satisfies the rotation-equivariant stability interface in Condition 1. Choose the small constant c_* in Theorem 5 sufficiently small that the ratio perturbation bound $C\varepsilon_{n,M}(\delta)$ is below the vertex-hunting threshold c_{VH}. In particular, successive projection satisfies this interface under the sufficient conditions in Proposition 2. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$, up to a common permutation of topic labels,*

$$\|\hat{a}_j - a_j\|_1 \le C h_j \varepsilon_{n,M}(\delta), \qquad 1 \le j \le p, \tag{28}$$

and

$$L(\hat{A}_{\mathrm{fed}}, A) \le C\varepsilon_{n,M}(\delta). \tag{29}$$

In particular, after using the pooled-marginal concentration bound for $\Delta_1(D, D_0)$ and the clientwise Gram-column bound from Lemma 25, the shared-whitening bridge residues are

$$\eta^{\mathrm{marg}}_{n,M}(\delta)\{1 + \vartheta_n(\delta) + \zeta^{\mathrm{svd}}_N\} \qquad \textit{and} \qquad \sqrt{p}\,\{\eta^{\mathrm{marg}}_{n,M}(\delta)\}^2,$$

the practical-SVD diagonal-bias residue is ζ^{svd}_N, and the conservative row-wise Bernstein range residue is $\sqrt{p}\,\ell_{n,M}(\delta)/M$. Thus, when the SVD-bias and bridge terms are lower order, the dominant global rate delivered by the row-wise theorem is

$$L(\hat{A}_{\mathrm{fed}}, A) \lesssim \beta_n \sqrt{\frac{p\ell_{n,M}(\delta)}{NnM}} + \beta_n^2 \frac{p\ell_{n,M}(\delta)}{Nn} + \frac{\sqrt{p}\,\ell_{n,M}(\delta)}{M},$$

up to the displayed lower-order bridge terms, the practical-SVD diagonal-bias term, and the fixed-whitening quadratic bias term. If an additional localized projector-column range inequality is imposed or proved, the final range term can be replaced by $\ell_{n,M}(\delta)/M$.

A.4.6 Regime comparison and the three-term balance

Corollary 3 (Regime comparison with the Bernstein range term)**.** *Under the rate simplification after Theorem 6, set $\delta = (nM)^{-3}$ and*

$$\ell_{n,M} := \ell_{n,M}(\delta) = \log(2pn^3M^4).$$

Define

$$T_{\mathrm{var}} := \beta_n \sqrt{\frac{p\ell_{n,M}}{NnM}}, \qquad T_{\mathrm{bias}} := \beta_n^2 \frac{p\ell_{n,M}}{Nn}, \qquad T_{\mathrm{range}} := \frac{\sqrt{p}\,\ell_{n,M}}{M},$$

and

$$T_{\mathrm{br,row}} := \eta^{\mathrm{marg}}_{n,M}(\delta)\{1 + \vartheta_n(\delta) + \zeta^{\mathrm{svd}}_N\}, \qquad T_{\mathrm{br,quad}} := \sqrt{p}\,\{\eta^{\mathrm{marg}}_{n,M}(\delta)\}^2, \qquad T_{\mathrm{svd}} := \zeta^{\mathrm{svd}}_N.$$

Then, up to the fixed-whitening quadratic lower-order term $(\beta_n^2 p/(Nn))^2$,

$$L(\hat{A}_{\mathrm{fed}}, A) \lesssim T_{\mathrm{var}} + T_{\mathrm{bias}} + T_{\mathrm{range}} + T_{\mathrm{br,row}} + T_{\mathrm{br,quad}} + T_{\mathrm{svd}}. \tag{30}$$

The first three terms are the main variance/local-bias/range contributions. The extra bridge terms $T_{\mathrm{br,row}}$ and $T_{\mathrm{br,quad}}$ are inherited from the shared empirical whitening matrix, while T_{svd} is the diagonal-bias price of the practical ordinary-SVD upload. In regimes where these residues are lower order, the qualitative balance is governed by the three main terms.

More explicitly, the variance-bias crossover solves $T_{\mathrm{var}} \asymp T_{\mathrm{bias}}$, giving

$$M \asymp \frac{Nn}{\beta_n^2 p\ell_{n,M}}.$$

The conservative row-wise Bernstein range term is no larger than the bias term once

$$M \gtrsim \frac{Nn}{\beta_n^2\sqrt{p}},$$

and is no larger than the variance term once

$$M \gtrsim \frac{\ell_{n,M} Nn}{\beta_n^2}.$$

Thus, unlike the operator-norm range residue, the unconditional row-wise range residue can be the limiting finite-sample term unless M is fairly large. A sharper localized projector-column range inequality would replace $\sqrt{p}\,\ell_{n,M}/M$ by $\ell_{n,M}/M$ and recover the milder balance in the operator-norm theorem.

Remark 17 (Why the range term remains in the statement). The familiar two-term decomposition $T_{\mathrm{var}} + T_{\mathrm{bias}}$ (called $T_{\mathrm{var}} + T_{\mathrm{loc}}$ in the main text) is recovered only when T_{range}, $T_{\mathrm{br,row}}$, $T_{\mathrm{br,quad}}$, and T_{svd} are lower order. In finite samples, the row-wise projector-average Bernstein inequality contributes the range term, the shared empirical whitening contributes the two bridge residues, and practical SVD upload contributes T_{svd}. The cleanest nonasymptotic statement keeps all of them visible. The displayed row-wise theorem is most informative in many-client regimes; when M is small, a direct pooled or single-site row-wise analysis can give sharper constants and avoid interpreting the conservative $\sqrt{p}/M$ range residue as a statistical limitation.

A.4.7 Heterogeneity-sensitive interpretation

Define the local signal gaps

$$\rho_m := \lambda_K(G_{0,m}), \qquad 1 \le m \le M.$$

Since

$$G_{0,m} = \Big(1 - \frac{1}{N}\Big) B \Sigma_{W,m} B^\top, \qquad B = M_0^{-1/2} A,$$

we have

$$\rho_m \ge \Big(1 - \frac{1}{N}\Big) \lambda_{\min}(B^\top B) \lambda_{\min}(\Sigma_{W,m}). \tag{31}$$

Thus extreme client imbalance can shrink local gaps even when the pooled matrix Σ_W is well behaved.

Proposition 3 (Heterogeneity-sensitive fixed-whitening projector bound for ordinary SVD)**.** *Let*

$$\bar{P}^{(0)} := \sum_{m=1}^{M} \alpha_m \tilde{P}_m, \qquad \rho_m := \lambda_K(G_{0,m}).$$

Assume $\rho_m > 0$ for every m and deterministic weights $\alpha_m \ge 0$ with $\sum_m \alpha_m = 1$. For the identity-centered fixed-whitening matrix, decompose

$$\tilde{G}_m - G_{0,m} = F_m + B_m, \qquad F_m := \tilde{G}_m - \mathbb{E}[\tilde{G}_m \mid W^{(1:M)}], \qquad B_m := D_{0,m}^{\mathrm{svd}},$$

so that $\mathbb{E}[F_m \mid W^{(1:M)}] = 0$. Define

$$v_m := \mathbb{E}\|F_m\|^2, \qquad b_m := \|B_m\|.$$

Let $\widehat{P}^{(0)}$ denote the top-K spectral projector of $\bar{P}^{(0)}$ whenever it is well defined, and let $\hat{\Xi}^{(0)}$ be any orthonormal basis of $\mathrm{range}(\widehat{P}^{(0)})$*. Assume that the deterministic ordinary-SVD diagonal bias is locally perturbative, $b_m \le \rho_m/8$ for all m. For $\delta \in (0,1)$ define*

$$\mathcal{E}_{\mathrm{loc}}(\delta) := \bigcap_{m=1}^{M} \{\|F_m\| \le \rho_m/8\}.$$

Assume $\mathbb{P}(\mathcal{E}_{\mathrm{loc}}(\delta/2)) \ge 1 - \delta/2$. Then, with probability at least $1-\delta$, on $\mathcal{E}_{\mathrm{loc}}(\delta/2)$ and the unconditional averaging event supplied by Theorem 4,

$$\|\bar{P}^{(0)} - P\| \le C\left[\left[\log\frac{4p}{\delta}\sum_{m=1}^{M}\alpha_m^2\frac{v_m + b_m^2}{\rho_m^2}\right]^{1/2} + \sum_{m=1}^{M}\alpha_m\left\{\frac{v_m}{\rho_m^2} + \frac{b_m}{\rho_m} + \frac{b_m^2}{\rho_m^2}\right\} + \log\frac{4p}{\delta}\max_m \alpha_m\right]. \tag{32}$$

If the right-hand side is at most $1/4$, the extracted top-K subspace obeys the same bound up to an absolute constant; equivalently,

$$\|\sin\Theta(\hat{\Xi}^{(0)}, \Xi)\| \le C\left[\left[\log\frac{4p}{\delta}\sum_{m=1}^{M}\alpha_m^2\frac{v_m + b_m^2}{\rho_m^2}\right]^{1/2} + \sum_{m=1}^{M}\alpha_m\left\{\frac{v_m}{\rho_m^2} + \frac{b_m}{\rho_m} + \frac{b_m^2}{\rho_m^2}\right\} + \log\frac{4p}{\delta}\max_m \alpha_m\right]. \tag{33}$$

Corollary 4 (Recovery of the regular regime)**.** *Assume the conditions of Proposition 3,* $\alpha_m = 1/M$*,*

$$v_m \le C\beta_n^2 \frac{p}{Nn}, \qquad b_m \le C\zeta_N^{\mathrm{svd}}, \qquad \zeta_N^{\mathrm{svd}} \le c\rho_*, \qquad \rho_m \ge \rho_* > 0, \qquad 1 \le m \le M.$$

Then, on the same event,

$$\|\sin\Theta(\hat{\Xi}^{(0)}, \Xi)\| \le C\left[\frac{\beta_n}{\rho_*}\sqrt{\frac{p\log(4p/\delta)}{NnM}} + \frac{\zeta_N^{\mathrm{svd}}}{\rho_*}\sqrt{\frac{\log(4p/\delta)}{M}} + \frac{\beta_n^2}{\rho_*^2}\frac{p}{Nn} + \frac{\zeta_N^{\mathrm{svd}}}{\rho_*} + \frac{\log(4p/\delta)}{M}\right]. \tag{34}$$

Remark 18 (Gap-explicit diagnostic for imbalance). The mild upward bend sometimes seen in the federated curve is not caused by a failure of the anchor-word/simplex geometry. It comes from the local matrices $\Sigma_{W,m}$: if some of them are nearly singular, then the corresponding local gaps ρ_m are small, and local spectral bias can dominate the variance reduction from averaging. This is a diagnostic statement outside the uniform regular theorem class when the gaps approach zero; the main consistency theorem assumes the regular regime $\lambda_{\min}(\Sigma_{W,m}) \ge c_{\mathrm{loc}}$.

A.4.8 A lower bound

Theorem 7 (Minimax lower bound on the matched theorem class)**.** *Fix* $K \ge 2$*. Assume* $n_m = n$*,* $n \ge K$*, and* $p \ge p_0(K)$*, where* $p_0(K)$ *is a sufficiently large constant, and*

$$\log(nM) \le \min\{p, N\}, \qquad p\log(nM) = o(NnM).$$

There exists a constant $c_\star(K) \in (0, 1/(2K)]$ *such that the following holds for any fixed* $c_1, c_2, c_{\mathrm{loc}} \in (0, c_\star(K)]$*.*

Let $\Phi_{n,N,p}^{(M)}(K, c_1, c_2, c_{\mathrm{loc}})$ *be the collection of parameters* $(A, W^{(1)}, \dots, W^{(M)})$ *such that: each topic has at least one anchor word;* $h_{\min} \ge c_1 K/p$*;* $\lambda_{\min}(\Sigma_W) \ge c_2$*,* $\lambda_{\min}(\Sigma_A) \ge c_2$*, and* $\lambda_{\min}(\Sigma_{W,m}) \ge c_{\mathrm{loc}}$ *for every client* m*. Then there exist constants* $c_0, \delta_0 > 0$ *such that, along any sequence* (n, M, N, p) *satisfying these conditions and for all sufficiently large problem sizes,*

$$\inf_{\hat{A}} \sup_{(A, W^{(1:M)}) \in \Phi_{n,N,p}^{(M)}(K, c_1, c_2, c_{\mathrm{loc}})} \mathbb{P}\Big(L(\hat{A}, A) \ge c_0\sqrt{\frac{p}{NnM}}\Big) \ge \delta_0.$$

The least favorable subclass used in the proof is chosen so that every client already satisfies local richness. Each client contains the same balanced collection of pure-topic documents, with either $\lfloor n/K \rfloor$ or $\lceil n/K \rceil$ documents at each topic weight e_k. Consequently, $\lambda_{\min}(\Sigma_{W,m}) \ge 1/(2K)$ for all m; when K divides n, this reduces to $\Sigma_{W,m} = K^{-1} I_K$. The packing therefore lies inside the same theorem class used by the upper bounds.

Remark 19 (Interpretation of the lower bound). Theorem 7 gives a minimax lower bound of order $\sqrt{p/(NnM)}$ on the stated parameter class. The variance contribution in Theorem 6 has the same dependence on p, N, n, M, up to logarithmic and β_n factors. The complete finite-sample upper bound also includes the local spectral bias, row-wise Bernstein range, shared-whitening, and ordinary-SVD contributions.

With $\ell = \ell_{n,M}(\delta)$, the conservative terms displayed in the main text satisfy

$$T_{\mathrm{loc}} T_{\mathrm{row}} = \sqrt{p}\,\ell\, T_{\mathrm{var}}^2.$$

The comparison with the lower bound concerns the variance component and does not assert a matching minimax rate for the complete upper bound; every term in (5) is retained.

A.4.9 Auxiliary estimation of document weights and K

Theorem 8 (Document-level topic weights via sample splitting)**.** *Split the corpus into a training part and a held-out document* d_i^{te}*. Suppose each client contributes* n_{tr} *training documents, and set* $n_{\mathrm{tot}}^{\mathrm{tr}} := Mn_{\mathrm{tr}}$*. Construct* $\hat{A}_{\mathrm{fed}}^{\mathrm{tr}}$ *and* M_d^{tr} *from the training split only. Define the training-split population and empirical pooled marginals by*

$$D_0^{\mathrm{tr}} := AW^{\mathrm{tr}}, \qquad M_0^{\mathrm{tr}} := \mathrm{diag}\big((n_{\mathrm{tot}}^{\mathrm{tr}})^{-1} D_0^{\mathrm{tr}} \mathbf{1}\big), \qquad M_d^{\mathrm{tr}} := \mathrm{diag}\big((n_{\mathrm{tot}}^{\mathrm{tr}})^{-1} D^{\mathrm{tr}} \mathbf{1}\big),$$

and define

$$\Delta_1^{\mathrm{tr}} := \max_{1\le j\le p} h_j^{-1}\big|M_d^{\mathrm{tr}}(j,j) - M_0^{\mathrm{tr}}(j,j)\big|.$$

Assume that all hypotheses of Theorem 6, including the anchor-word, minimum-row-mass, pooled and client-level richness, vertex-hunting stability, and perturbative conditions, hold on the training split with n replaced by n_{tr}. For the error statements, relabel the training topic estimate by the common permutation supplied by that theorem and express the estimated weights in the same topic order; this permutation depends only on the training data. Define

$$\hat{w}_i^\star := \arg\min_{b\in\mathbb{R}^K} \sum_{j=1}^{p} \frac{1}{M_d^{\mathrm{tr}}(j,j)}\Big(d_i^{\mathrm{te}}(j) - \sum_{k=1}^{K} b(k)\hat{A}_{\mathrm{fed}}^{\mathrm{tr}}(j,k)\Big)^2.$$

Truncate negative coordinates and renormalize to the simplex to obtain $\hat{w}_i$. Let $\hat{a}_j^{\mathrm{tr}}$ denote the jth row of $\hat{A}_{\mathrm{fed}}^{\mathrm{tr}}$, let $\varepsilon_{n_{\mathrm{tr}},M}^{\mathrm{tr}}(\delta_{\mathrm{tr}})$ denote the deterministic radius from Theorems 5 and 6 applied to the training split, and define the training event

$$\mathcal{E}_{\mathrm{tr}}(\delta_{\mathrm{tr}}) := \left\{ \max_{1\le j\le p} \frac{\|\hat{a}_j^{\mathrm{tr}} - a_j\|_1}{h_j} \le C\varepsilon_{n_{\mathrm{tr}},M}^{\mathrm{tr}}(\delta_{\mathrm{tr}}), \qquad \Delta_1^{\mathrm{tr}} \le c \right\},$$

for a sufficiently small deterministic constant $c > 0$. Conditional on the training split, on $\mathcal{E}_{\mathrm{tr}}(\delta_{\mathrm{tr}})$,

$$\|\hat{w}_i - w_i\|_1 \le 2 \wedge C\left\{ \varepsilon_{n_{\mathrm{tr}},M}^{\mathrm{tr}}(\delta_{\mathrm{tr}}) + \sqrt{\frac{\log(2K/\delta_{\mathrm{te}})}{N}} \right\}$$

with probability at least $1-\delta_{\mathrm{te}}$ over the held-out document. Moreover, if $\mathcal{E}_{\mathrm{tr}}(\delta_{\mathrm{tr}})$ holds with probability at least $1-\delta_{\mathrm{tr}}$ over the training split, then the same bound holds unconditionally with probability at least $1-\delta_{\mathrm{tr}}-\delta_{\mathrm{te}}$. The event $\mathcal{E}_{\mathrm{tr}}(\delta_{\mathrm{tr}})$ is supplied by Theorem 6 together with the pooled-marginal concentration lemma on the training split under the training-split regularity conditions just stated.

This training/held-out module is auxiliary to the projector-only core: one first estimates $\hat{A}_{\mathrm{fed}}^{\mathrm{tr}}$ and M_d^{tr} on the training split and then estimates the held-out w_i. For cross-fitting with a fixed number of folds, evaluate each document only using fits trained without that document. If several such predictions are averaged, first align topic columns to a reference fit by minimum total columnwise ℓ_1 matching. This matching is correct whenever every fit has maximum column error at most $e < c_2/2$ after its oracle permutation: for $k \neq l$, Assumption 6 gives

$$\|A_{\cdot k} - A_{\cdot l}\|_1 \ge (e_k - e_l)^\top \Sigma_A (e_k - e_l) \ge 2c_2,$$

so matched columns are at distance at most $2e$ and mismatched columns are at distance at least $2c_2 - 2e > 2e$. For simultaneous guarantees, allocate the desired failure probability across all training events and evaluated held-out documents.

Theorem 9 (Optional pilot-stage rule for K; outside the projector-only core)**.** *Suppose each client forms an auxiliary symmetric pilot sketch $\hat{G}_m^{\mathrm{pil}}$ on an independent pilot pass, with*

$$\mathbb{E}[\hat{G}_m^{\mathrm{pil}} \mid W^{(1:M)}] = G_{0,m},$$

and the server averages

$$G_{\mathrm{pilot}} := \frac{1}{M}\sum_{m=1}^{M} \hat{G}_m^{\mathrm{pil}}, \qquad G_0 := \frac{1}{M}\sum_{m=1}^{M} G_{0,m}.$$

Assume $\lambda_K(G_0) \ge \rho_* > 0$ and $\lambda_{K+1}(G_0) = 0$. For any threshold $t \in (0, \rho_*)$, define

$$\hat{K}_t^{\mathrm{pil}} := \#\{k \in [p] : \lambda_k(G_{\mathrm{pilot}}) \ge t\}.$$

If

$$\mathbb{P}\{\|G_{\mathrm{pilot}} - G_0\| < \min(t, \rho_* - t)\} \ge 1 - o((nM)^{-3}),$$

then $\hat{K}_t^{\mathrm{pil}} = K$ with probability $1 - o((nM)^{-3})$. For a pilot sketch and threshold satisfying these conditions, one may rerun Algorithm 2 with $K = \hat{K}_t^{\mathrm{pil}}$.

A.4.10 Positioning relative to single-site Topic-SCORE and federated subspace averaging

The federated analysis differs from single-site SVD Topic-SCORE in one essential respect. In the single-site proof, a global normalized corpus matrix has a Perron-selected first singular direction, and the remaining analysis proceeds by SCORE ratios, simplex recovery, and column normalization. Here the client-level matrices share only the column space $\mathcal{S} = \mathrm{col}(M_0^{-1/2}A)$; their eigenbases depend on the local second moments $\Sigma_{W,m}$. Projector averaging is therefore used only to recover the common subspace. The pooled marginal then supplies the missing positive direction $\xi_1 = M_0^{1/2}\mathbf{1}$, after which the deterministic Topic-SCORE geometry is the same as in the single-site Topic-SCORE analysis.

The projector-averaging step is also related to distributed principal-subspace estimation, but the statistical target and the loss are different. Distributed PCA analyses typically estimate a principal eigenspace in operator or Frobenius norm. The present problem additionally requires row-wise control, a canonical direction inside the averaged subspace, multinomial centering, and propagation through a simplex map to an ℓ_1 topic-matrix loss. This is why the proof keeps separate the operator variance term, the row-wise Bernstein range term, the local second-order bias, and the shared-whitening bridge terms.

A.4.11 Proof roadmap

The final proof chain is

$$\tilde{G}_m - G_{0,m} \to \tilde{P}_m - P \to \bar{P}^{(0)} - P \to \bar{P} - P \to \widehat{P} - P \to (\hat{\xi}_1, \hat{\Xi}_\perp) \to \hat{R} \to \hat{\Pi} \to \hat{A}.$$

The first arrow is the local identity-centered SVD perturbation, including the deterministic diagonal bias $D_{0,m}^{\mathrm{svd}}$. The second is the variance-bias projector-averaging step for the fixed-whitening projectors. The third is the bridge correction from the shared empirical whitening matrix M_d. The fourth extracts the spectral projector of the averaged matrix. The fifth is the canonical step unique to the federated problem: the projected pooled marginal supplies $\hat{\xi}_1$, and the residual basis is aligned only by an $(K-1)$-dimensional orthogonal matrix. Once the canonical direction is inserted, the ratio, simplex, and topic-matrix propagation mirrors the single-site Topic-SCORE proof.

Proof organization. The remaining subsections give the detailed proofs and supporting calculations.

A.5 Auxiliary linear algebra

A.5.1 Proof of Lemma 1

Proof. For each document,

$$\mathbb{E}[d_i d_i^\top \mid W] = \left(1 - \frac{1}{N}\right)\pi_i\pi_i^\top + \frac{1}{N}\,\mathrm{diag}(\pi_i).$$

Averaging within client m and pre- and post-multiplying by $M_0^{-1/2}$ gives

$$\mathbb{E}[\tilde{G}_m \mid W^{(1:M)}] = G_{0,m} + \frac{1}{N}\left(M_0^{-1/2}M_{0,m}M_0^{-1/2} - I_p\right) = G_{0,m} + D_{0,m}^{\mathrm{svd}},$$

which proves (15). For the bridge, since

$$M_d^{-1/2} = (I_p + R_M)M_0^{-1/2} \quad \text{and} \quad T_m = M_0^{-1/2}\hat{S}_m M_0^{-1/2},$$

we have

$$\hat{G}_m + \frac{1}{N}I_p = (I_p + R_M)T_m(I_p + R_M), \qquad \tilde{G}_m + \frac{1}{N}I_p = T_m.$$

Subtracting the two displays gives

$$\hat{G}_m - \tilde{G}_m = R_M T_m + T_m R_M + R_M T_m R_M.$$

Combining this bridge identity with the fixed-whitening mean decomposition yields (17). □

A.5.2 Proof of Lemma 2

Proof. The projector P has eigenvalues 1 with multiplicity K and 0 with multiplicity $p-K$. If $\left\|\bar{\mathcal{P}}-P\right\| \le 1/4$, Weyl's inequality shows that $\bar{\mathcal{P}}$ has exactly K eigenvalues in $[3/4, 5/4]$ and the remaining eigenvalues in $[-1/4, 1/4]$. Hence the top-K eigenspace is uniquely defined. Applying Davis–Kahan to the gap-one cluster of P gives

$$\left\|\hat{P}-P\right\| = \left\|\sin\Theta(\hat{\Xi},\Xi)\right\| \le 2\left\|\bar{\mathcal{P}}-P\right\|.$$

The same proof gives a constant depending only on c_{DK} whenever $\left\|\bar{\mathcal{P}}-P\right\| \le c_{\mathrm{DK}} < 1/2$. □

A.5.3 Proof of Lemma 3

Proof. The only algebraic point is to use the normalization $\sum_{m=1}^{M}\alpha_m = 1$. Indeed,

$$\bar{\mathcal{P}}-P = \sum_{m=1}^{M}\alpha_m\mathcal{P}_m - \Big(\sum_{m=1}^{M}\alpha_m\Big)P.$$

Adding and subtracting $\sum_{m=1}^{M}\alpha_m\mathbb{E}\mathcal{P}_m$ gives the exact identity

$$\bar{\mathcal{P}}-P = \sum_{m=1}^{M}\alpha_m(\mathcal{P}_m-\mathbb{E}\mathcal{P}_m) + \sum_{m=1}^{M}\alpha_m(\mathbb{E}\mathcal{P}_m-P).$$

Taking operator norms and applying the triangle inequality proves the claim. This lemma itself does not use independence or nonnegativity of the weights. Nonnegativity is used in Lemmas 4 and 6 to write the range terms with $\max_m\alpha_m$ rather than $\max_m|\alpha_m|$. □

A.5.4 Proof of Lemma 4

Proof. Set $A_m := \mathcal{P}_m - P$ and $B_m := \mathbb{E}\mathcal{P}_m - P = \mathbb{E}[A_m]$. Then

$$\mathcal{P}_m-\mathbb{E}\mathcal{P}_m = A_m-B_m, \qquad X_m = \alpha_m(A_m-B_m).$$

Because $\mathcal{P}_m$ and P are orthogonal projectors and $\mathbb{E}\mathcal{P}_m$ is a positive semidefinite contraction, $\|\mathcal{P}_m\| \le 1$ and $\|\mathbb{E}\mathcal{P}_m\| \le 1$. Since the averaging weights are nonnegative,

$$\|X_m\| \le \alpha_m\|\mathcal{P}_m-\mathbb{E}\mathcal{P}_m\| \le 2\alpha_m.$$

For the variance proxy, we use the noncommutative variance identity directly. Since $B_m = \mathbb{E}A_m$ is deterministic and self-adjoint,

$$\begin{aligned}\mathbb{E}[(\mathcal{P}_m-\mathbb{E}\mathcal{P}_m)^2] &= \mathbb{E}[(A_m-B_m)^2]\\ &= \mathbb{E}[A_m^2] - \mathbb{E}[A_m]B_m - B_m\mathbb{E}[A_m] + B_m^2\\ &= \mathbb{E}[A_m^2] - B_m^2.\end{aligned}$$

The left-hand side is positive semidefinite, because $(A_m-B_m)^2 \succeq 0$ pointwise. Since $B_m^2 \succeq 0$, the identity gives

$$0 \preceq \mathbb{E}[(\mathcal{P}_m-\mathbb{E}\mathcal{P}_m)^2] \preceq \mathbb{E}[A_m^2].$$

Consequently,

$$\left\|\mathbb{E}[(\mathcal{P}_m-\mathbb{E}\mathcal{P}_m)^2]\right\| \le \|\mathbb{E}[A_m^2]\| \le \mathbb{E}\|A_m^2\| = \mathbb{E}\|A_m\|^2 = \nu_m.$$

In particular, the weaker bound by $4\nu_m$ used in the statement also holds. Multiplying by α_m^2 and summing over m gives

$$\left\|\sum_{m=1}^{M}\mathbb{E}[X_m^2]\right\| \le \sum_{m=1}^{M}\alpha_m^2\left\|\mathbb{E}[(\mathcal{P}_m-\mathbb{E}\mathcal{P}_m)^2]\right\| \le 4\sum_{m=1}^{M}\alpha_m^2\nu_m.$$

Tropp's self-adjoint matrix Bernstein inequality applied to the independent centered matrices X_m now yields the displayed probability bound. □

A.5.5 Proof of Lemma 5

Proof. Let $R(z) = (zI - G)^{-1}$. Choose a contour $\mathcal{C}$ (or a finite union of contours) enclosing precisely the target cluster and staying at distance at least $\gamma/2$ from the entire spectrum of G. The cluster-diameter assumption permits total length at most $C(C_0)\gamma$, and $\sup_{z\in\mathcal{C}} \|R(z)\| \le 2/\gamma$. The Riesz formulas are

$$P = \frac{1}{2\pi i}\oint_{\mathcal{C}} R(z)\,dz, \qquad P_m = \frac{1}{2\pi i}\oint_{\mathcal{C}} (zI - G - E_m)^{-1}\,dz.$$

If $\|E_m\| \le \gamma/4$, then $\|E_m R(z)\| \le 1/2$ on $\mathcal{C}$, so the resolvent admits the Neumann expansion

$$(zI - G - E_m)^{-1} = R(z) + R(z)E_m R(z) + \sum_{r\ge 2} R(z)(E_m R(z))^r.$$

Subtracting the Riesz formula for P gives

$$P_m - P = \mathcal{L}_G(E_m) + \mathcal{R}_G(E_m), \qquad \mathbb{E}[\mathcal{L}_G(E_m)] = 0,$$

because $\mathbb{E}E_m = 0$. The operator remainder satisfies

$$\|\mathcal{R}_G(E_m)\| \le \frac{C}{\gamma^2}\|E_m\|^2,$$

and hence

$$\|\mathbb{E}P_m - P\| = \|\mathbb{E}\mathcal{R}_G(E_m)\| \le \frac{C}{\gamma^2}\mathbb{E}\|E_m\|^2.$$

We now prove the columnwise display under the additional rank-K signal assumption in the lemma. In this case $R(z)(I - P) = z^{-1}(I - P)$ on the contour and $|z| \asymp \gamma$. Hence, for any vector x and any $z \in \mathcal{C}$,

$$\|E_m R(z)x\|_2 \le C\gamma^{-1}\{\|E_m x\|_2 + \|E_m\|\,\|Px\|_2\}.$$

Indeed,

$$E_m R(z)x = E_m R(z)Px + z^{-1}E_m(I - P)x,$$

and the second term is bounded by $C\gamma^{-1}\{\|E_m x\|_2 + \|E_m\|\|Px\|_2\}$. Using the Neumann-series expression of the remainder,

$$\mathcal{R}_G(E_m)e_j = \frac{1}{2\pi i}\oint_{\mathcal{C}} \sum_{r\ge 2} R(z)(E_m R(z))^{r-1}E_m R(z)e_j\,dz,$$

we therefore obtain, since $\|E_m R(z)\| \le 1/2$,

$$\|\mathcal{R}_G(E_m)e_j\|_2 \le \frac{C}{\gamma^2}\,\|E_m\|\{\|E_m e_j\|_2 + \|E_m\|\,\|Pe_j\|_2\}.$$

Taking expectations and applying Cauchy–Schwarz to the first product gives

$$\|(\mathbb{E}P_m - P)e_j\|_2 \le \frac{C}{\gamma^2}\left[(\mathbb{E}\|E_m\|^2)^{1/2}(\mathbb{E}\|E_m e_j\|_2^2)^{1/2} + \|Pe_j\|_2\,\mathbb{E}\|E_m\|^2\right],$$

which proves (19). □

A.5.6 Projector perturbation expansion

Lemma 14 (Linear term plus quadratic remainder). Let G be symmetric with a size-K isolated spectral cluster separated from the rest of the spectrum by an eigengap $\gamma > 0$, and let P be its associated projector. Assume the target cluster has diameter at most $C_0\gamma$ for a fixed constant C_0. If $\hat{G} = G + E$ and $\|E\| \le \gamma/4$, then, with C depending only on C_0,

$$\hat{P} - P = \mathcal{L}_G(E) + \mathcal{R}_G(E), \qquad \|\mathcal{L}_G(E)\| \le \frac{C}{\gamma}\,\|E\|, \qquad \|\mathcal{R}_G(E)\| \le \frac{C}{\gamma^2}\,\|E\|^2.$$

If, in addition, $G \succeq 0$ has rank K and P projects onto its nonzero eigenspace, then, for each $j \in [p]$,

$$\left\|(\hat{P} - P)e_j\right\|_2 \le \frac{C}{\gamma}\{\|Ee_j\|_2 + \|E\|\,\|Pe_j\|_2\} + \frac{C}{\gamma^2}\,\|E\|^2. \tag{35}$$

Under this additional rank-K signal assumption, the quadratic remainder has the refined column bound

$$\|\mathcal{R}_G(E)e_j\|_2 \le \frac{C}{\gamma^2}\,\|E\|\,\Big\{\|Ee_j\|_2 + \|E\|\,\|Pe_j\|_2\Big\}. \tag{36}$$

In that rank-K signal case, the crude final term in (35) can be replaced by the right-hand side of (36).

Proof. Let $R(z) := (zI - G)^{-1}$. Choose a contour $\mathcal{C}$ (or a finite union of contours) enclosing precisely the target cluster and staying at distance at least $\gamma/2$ from the entire spectrum of G. Its total length is at most $C(C_0)\gamma$ and $\sup_{z\in\mathcal{C}} \|R(z)\| \le 2/\gamma$. The Riesz formulas give

$$P = \frac{1}{2\pi i}\oint_{\mathcal{C}} R(z)\,dz, \qquad \hat{P} = \frac{1}{2\pi i}\oint_{\mathcal{C}} (zI - G - E)^{-1}\,dz.$$

Because $\|E\| \le \gamma/4$, the Neumann expansion is valid on $\mathcal{C}$:

$$(zI - G - E)^{-1} = R(z) + R(z)ER(z) + \sum_{r\ge 2} R(z)(ER(z))^r.$$

Thus

$$\mathcal{L}_G(E) := \frac{1}{2\pi i}\oint_{\mathcal{C}} R(z)ER(z)\,dz, \qquad \mathcal{R}_G(E) := \frac{1}{2\pi i}\oint_{\mathcal{C}} \sum_{r\ge 2} R(z)(ER(z))^r\,dz.$$

The operator-norm bounds follow from $\|R(z)\| \le C/\gamma$ and the geometric series.

For the column statements, assume that $G \succeq 0$ has rank K and P projects onto its nonzero eigenspace. Put $Q = I - P$. Because the complementary spectrum is exactly zero, the residue calculation gives

$$\mathcal{L}_G(E) = G^\dagger EQ + QEG^\dagger,$$

where $G^\dagger$ is the Moore–Penrose inverse and $\|G^\dagger\| \le \gamma^{-1}$. Since $G^\dagger = G^\dagger P$, we have

$$\begin{aligned}\|\mathcal{L}_G(E)e_j\|_2 &\le \gamma^{-1}\|EQe_j\|_2 + \gamma^{-1}\|E\|\|Pe_j\|_2 \\ &\le \gamma^{-1}\{\|Ee_j\|_2 + 2\|E\|\|Pe_j\|_2\}.\end{aligned}$$

Combining this bound with $\|\mathcal{R}_G(E)e_j\|_2 \le C\gamma^{-2}\|E\|^2$ proves (35) under the stated rank-K signal assumption.

It remains to prove the refined columnwise remainder in the rank-K signal case. Assume now that $G \succeq 0$ has rank K and that P projects onto its nonzero eigenspace. Then the complement spectrum is the single point zero, and on the contour

$$R(z)(I - P) = z^{-1}(I - P), \qquad |z|^{-1} \le C\gamma^{-1}.$$

Set

$$y_z := \|E(zI - G - E)^{-1}e_j\|_2.$$

The resolvent identity gives

$$(zI - G - E)^{-1} = R(z) + R(z)E(zI - G - E)^{-1}.$$

Therefore

$$y_z \le \|ER(z)e_j\|_2 + \|E\|\|R(z)\|y_z.$$

Because $\|E\|\|R(z)\| \le c < 1$ on the contour, the second term can be absorbed. For the first term, using $R(z)(I - P) = z^{-1}(I - P)$,

$$\|ER(z)e_j\|_2 \le C\gamma^{-1}\{\|E(I - P)e_j\|_2 + \|E\|\|Pe_j\|_2\} \le C\gamma^{-1}\{\|Ee_j\|_2 + \|E\|\|Pe_j\|_2\}.$$

Thus

$$y_z \le C\gamma^{-1}\{\|Ee_j\|_2 + \|E\|\|Pe_j\|_2\}.$$

Using the identity

$$\mathcal{R}_G(E) = \frac{1}{2\pi i}\oint_{\mathcal{C}} R(z)ER(z)E(zI - G - E)^{-1}\,dz,$$

valid under the same Neumann expansion, we get

$$\|\mathcal{R}_G(E)e_j\|_2 \le C\gamma \cdot \frac{1}{\gamma}\|E\| \cdot \frac{1}{\gamma} \cdot \frac{1}{\gamma}\{\|Ee_j\|_2 + \|E\|\|Pe_j\|_2\} \le \frac{C}{\gamma^2}\|E\|\{\|Ee_j\|_2 + \|E\|\|Pe_j\|_2\},$$

which proves (36). □

A.6 Row-wise perturbation: proofs

This appendix proves the deterministic perturbation statements that convert an averaged projector into the canonical Topic-SCORE basis. The first column is not chosen from an arbitrary top-K eigenbasis: it is reconstructed by projecting the pooled empirical marginal. The only remaining alignment is the residual $(K-1)$-dimensional orthogonal matrix $O_\perp$.

A.6.1 Proof of Lemma 9

Proof. Write $E := \bar{P} - P$ and let $\hat{\Xi}$ be any orthonormal basis of the range of $\widehat{P}$. Since P is a rank-K projector, its nonzero eigenvalues are all equal to one and the eigengap between the signal cluster and the null space is one. Applying Lemma 28 to the pair $(P, P+E)$ gives an orthogonal matrix $O \in \mathbb{R}^{K\times K}$ such that

$$\max_j h_j^{-1/2}\|e_j^\top(\hat{\Xi}O - \Xi)\|_2 \le C\left[\|E\| + \max_j h_j^{-1/2}\|Ee_j\|_2\right], \qquad \|\hat{\Xi}O - \Xi\| \le C\|E\|.$$

Because

$$\widehat{P} - P = (\hat{\Xi}O - \Xi)(\hat{\Xi}O)^\top + \Xi(\hat{\Xi}O - \Xi)^\top,$$

we have, for every j,

$$\|(\widehat{P} - P)e_j\|_2 = \|e_j^\top(\widehat{P} - P)\|_2 \le \|e_j^\top(\hat{\Xi}O - \Xi)\|_2 + \|e_j^\top\Xi\|_2\|\hat{\Xi}O - \Xi\|.$$

The oracle row-scale bound $\|e_j^\top\Xi\|_2 \le C\sqrt{h_j}$ then gives the displayed row bound. The operator-norm bound follows from Davis–Kahan for the projector signal. This proves the lemma. □

A.6.2 Proof of Lemma 10

Proof. Let

$$z := \widehat{P}u_d, \qquad \nu := \Delta_1(D, D_0) + \max_j h_j^{-1/2}\|(\widehat{P} - P)e_j\|_2 + \|\widehat{P} - P\|.$$

Because $P\xi_1 = \xi_1$ and $\|u_d\|_2 = \|\xi_1\|_2 = 1$,

$$z - \xi_1 = (\widehat{P} - P)u_d + P(u_d - \xi_1).$$

For the first term,

$$|e_j^\top(\widehat{P} - P)u_d| \le \|(\widehat{P} - P)e_j\|_2\|u_d\|_2 = \|(\widehat{P} - P)e_j\|_2.$$

For the second term, write $P = \Xi\Xi^\top$. By Lemma 40, $\|e_j^\top\Xi\|_2 \le C\sqrt{h_j}$. On the event $\Delta_1(D, D_0) \le c$, the diagonal entries of M_d and M_0 are comparable and

$$|u_d(j) - \xi_1(j)| = \left|\sqrt{M_d(j,j)} - \sqrt{M_0(j,j)}\right| \le C\Delta_1(D, D_0)\sqrt{h_j}.$$

Since $\sum_j h_j = K$, this gives $\|u_d - \xi_1\|_2 \le C\Delta_1(D, D_0)$. Hence

$$|e_j^\top P(u_d - \xi_1)| \le \|e_j^\top\Xi\|_2\,\|\Xi^\top(u_d - \xi_1)\|_2 \le C\sqrt{h_j}\Delta_1(D, D_0).$$

Therefore

$$\max_j h_j^{-1/2}|z(j) - \xi_1(j)| \le C\Delta_1(D, D_0) + \max_j h_j^{-1/2}\|(\widehat{P} - P)e_j\|_2.$$

Also,

$$\|z - \xi_1\|_2 \le \|\widehat{P} - P\| + C\Delta_1(D, D_0).$$

By the assumed smallness of $\Delta_1(D, D_0) + \|\widehat{P} - P\|$ and the preceding bound, $\|z\|_2 \ge 1/2$. Since $\hat{\xi}_1 = z/\|z\|_2$,

$$\hat{\xi}_1 - \xi_1 = \frac{z - \xi_1}{\|z\|_2} + \left(\frac{1}{\|z\|_2} - 1\right)\xi_1.$$

Using $|\xi_1(j)| \le C\sqrt{h_j}$ and $|\|z\|_2 - 1| \le \|z - \xi_1\|_2$ yields

$$\max_j h_j^{-1/2}|\hat{\xi}_1(j) - \xi_1(j)| \le C\left[\Delta_1(D, D_0) + \max_j h_j^{-1/2}\|(\widehat{P} - P)e_j\|_2 + \|\widehat{P} - P\|\right].$$

□

A.6.3 Proof of Lemma 11

Proof. Set

$$\eta := \Delta_1(D, D_0) + \max_j h_j^{-1/2} \|(\widehat{P} - P)e_j\|_2 + \|\widehat{P} - P\|.$$

Lemma 10 gives

$$\max_j h_j^{-1/2} |\hat{\xi}_1(j) - \xi_1(j)| \le C\eta, \qquad \|\hat{\xi}_1 - \xi_1\|_2 \le C\eta.$$

Now

$$\widehat{P}_\perp - P_\perp = (\widehat{P} - P) - (\hat{\xi}_1\hat{\xi}_1^\top - \xi_1\xi_1^\top),$$

and

$$\hat{\xi}_1\hat{\xi}_1^\top - \xi_1\xi_1^\top = (\hat{\xi}_1 - \xi_1)\hat{\xi}_1^\top + \xi_1(\hat{\xi}_1 - \xi_1)^\top.$$

Consequently,

$$\|\widehat{P}_\perp - P_\perp\| \le C\eta.$$

For the row/column bound, using $\hat{\xi}_1(j) \asymp \sqrt{h_j}$ in the perturbative regime,

$$\|(\widehat{P}_\perp - P_\perp)e_j\|_2 \le \|(\widehat{P} - P)e_j\|_2 + |\hat{\xi}_1(j)|\,\|\hat{\xi}_1 - \xi_1\|_2 + |\hat{\xi}_1(j) - \xi_1(j)|,$$

so

$$\max_j h_j^{-1/2} \|(\widehat{P}_\perp - P_\perp)e_j\|_2 \le C\eta.$$

Apply the same projector-to-basis argument as in Lemma 9 to the rank-$(K-1)$ projectors $P_\perp$ and $\widehat{P}_\perp$. Since $\|e_j^\top \Xi_\perp\|_2 \le C\sqrt{h_j}$, there exists $O_\perp \in \mathbb{R}^{(K-1)\times(K-1)}$ orthogonal such that

$$\max_j h_j^{-1/2} \|e_j^\top(\hat{\Xi}_\perp O_\perp - \Xi_\perp)\|_2 \le C\eta.$$

□

A.6.4 Proof of Lemma 6

Proof. Fix j and write

$$r_j - \mathbb{E}r_j = \sum_{m=1}^{M} v_{m,j}, \qquad v_{m,j} := \alpha_m \Big((\mathcal{P}_m - P)e_j - \mathbb{E}[(\mathcal{P}_m - P)e_j] \Big).$$

Then $\mathbb{E}v_{m,j} = 0$, $\|v_{m,j}\|_2 \le 4\alpha_m$, and

$$\sum_{m=1}^{M} \mathbb{E}\|v_{m,j}\|_2^2 \le \sum_{m=1}^{M} \alpha_m^2 \tau_{m,j}^2.$$

Put $L = 4\max_m \alpha_m$ and $V = \sum_m \mathbb{E}\|v_{m,j}\|_2^2$. For every integer $q \ge 3$,

$$\sum_m \mathbb{E}\|v_{m,j}\|_2^q \le L^{q-2} V \le \frac{q!}{2} V (L/3)^{q-2},$$

because $q!/2 \ge 3^{q-2}$. The threshold $\sqrt{2Vu} + Bu$ established in the proof of Lemma 23, with $B = L/3$, therefore gives, with $u = \log(2/\delta)$,

$$\|r_j - \mathbb{E}r_j\|_2 \le \sqrt{2u} \left(\sum_{m=1}^{M} \alpha_m^2 \tau_{m,j}^2 \right)^{1/2} + \frac{4}{3} u \max_m \alpha_m$$

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

A.7 Multinomial toolkit

A.7.1 Identity-centered Gram expectation

For a generic client corpus $D = [d_1, \dots, d_n]$ with $d_i = X_i/N$ and $X_i \sim \mathrm{Multinomial}(N, \pi_i)$, let $D_0 = [\pi_1, \dots, \pi_n]$, let M_* be any deterministic positive diagonal matrix, and define the local population marginal

$$M_{0,\mathrm{loc}} := \mathrm{diag}\Big(\frac{1}{n} D_0 \mathbf{1}\Big).$$

Set

$$\tilde{G}^{\mathrm{svd}}(M_*) := \frac{1}{n} M_*^{-1/2} D D^\top M_*^{-1/2} - \frac{1}{N} I_p,$$

and

$$G_0(M_*) := \Big(1 - \frac{1}{N}\Big) \frac{1}{n} M_*^{-1/2} D_0 D_0^\top M_*^{-1/2}.$$

Lemma 15 (Identity-centered fixed-whitening expectation). Under common document length N,

$$\mathbb{E}[\tilde{G}^{\mathrm{svd}}(M_*) \mid W] = G_0(M_*) + \frac{1}{N}\{M_*^{-1/2} M_{0,\mathrm{loc}} M_*^{-1/2} - I_p\}.$$

In particular, with $M_* = M_0$ and $D = D^{(m)}$, this gives

$$\mathbb{E}[\tilde{G}_m \mid W^{(1:M)}] = G_{0,m} + D_{0,m}^{\mathrm{svd}}.$$

Proof. For each i,

$$\mathbb{E}[d_i] = \pi_i, \qquad \mathbb{E}[d_i d_i^\top] = \Big(1 - \frac{1}{N}\Big)\pi_i \pi_i^\top + \frac{1}{N}\,\mathrm{diag}(\pi_i).$$

Therefore

$$\mathbb{E}\left[\frac{1}{n} D D^\top \;\middle|\; W\right] = \Big(1 - \frac{1}{N}\Big)\frac{1}{n} D_0 D_0^\top + \frac{1}{N} M_{0,\mathrm{loc}}.$$

Pre- and post-multiplying by $M_*^{-1/2}$ and subtracting $N^{-1} I_p$ proves the claim. □

Remark 20 (Proof-level note on unequal document lengths). The main text uses the common-length convention to keep notation readable. If document lengths are unequal, the identity-centered ordinary-SVD matrix remains the algorithmic object, while the fixed-whitening expectation gains the corresponding length-weighted diagonal mean. The federated projector and canonical-direction arguments are unchanged; only the deterministic diagonal SVD-bias term and the effective document-length scale in the concentration bounds are altered.

Remark 21 (Why the identity shift creates a diagonal bias). The identity shift $N^{-1} I_p$ is exactly the shift implemented by ordinary SVD, and it does not alter the empirical singular vectors. It is not an exact local multinomial debiasing term under pooled whitening: unless the local population marginal equals the pooled marginal, the fixed-whitening mean contains the diagonal perturbation $D_{0,m}^{\mathrm{svd}}$. This term is kept explicitly in the theorem radius through ζ_N^{svd}.

A.7.2 Moment bounds and bridge control

The following bounds summarize the multinomial concentration ingredients used throughout the main proofs.

Lemma 16 (Whitened Gram fluctuation bounds). Under Assumptions 5–6, for the centered fixed-whitening sampling fluctuation

$$E := \tilde{G}_m - \mathbb{E}[\tilde{G}_m \mid W^{(1:M)}],$$

the column moment bound

$$\mathbb{E}\,\|E e_j\|_2^2 \le C h_j \beta_n^2 \frac{p}{Nn}$$

holds for every $j \in [p]$. If, in addition, the local perturbative scaling

$$\beta_n^2 \frac{p^2}{Nn} \le c_0$$

holds for a sufficiently small absolute constant $c_0 > 0$, then

$$\mathbb{E}\,\|E\|^2 \le C\beta_n^2 \frac{p}{Nn}, \qquad \mathbb{E}\,\|E\|^4 \le C\beta_n^4 \Big(\frac{p}{Nn}\Big)^2.$$

In fact, the direct moment argument in Lemma 20 proves these two operator moment bounds without the additional local scaling. Moreover, with probability at least $1-\delta$,

$$\Delta_1(D, D_0) \le C\sqrt{\frac{p\log(2p/\delta)}{NnM}} + C\frac{p\log(2p/\delta)}{NnM}.$$

In the perturbative regime $p\log(2p/\delta)/(NnM) \le 1$, the second term is absorbed by the first.

Remark 22. Lemma 16 is the federated counterpart of the moment bounds used in the single-site Topic-SCORE analysis of Ke and Wang's SVD Topic-SCORE paper. The first two inequalities control the centered local sampling fluctuation under fixed whitening, while the deterministic identity-shift diagonal bias is tracked separately through $D_{0,m}^{\mathrm{svd}}$. The final display quantifies the extra error induced by the shared empirical whitening matrix.

A.7.3 Basic multinomial moments

Lemma 17 (Second moments of multinomial proportions). Condition on $\pi \in \mathbb{R}^p$ with $\pi \ge 0$ and $\mathbf{1}^\top\pi = 1$, and let $d = X/N$ with $X \sim \mathrm{Multinomial}(N, \pi)$. Then for all $j \ne \ell$,

$$\mathbb{E}[d(j)] = \pi(j), \qquad \mathbb{E}[d(j)^2] = \pi(j)^2 + \frac{1}{N}\pi(j)(1-\pi(j)) \le \pi(j)^2 + \frac{1}{N}\pi(j),$$

$$\mathbb{E}[d(j)d(\ell)] = \pi(j)\pi(\ell) - \frac{1}{N}\pi(j)\pi(\ell).$$

Proof. The identities are standard multinomial moment formulas. Indeed,

$$\mathbb{E}[d(j)] = \frac{1}{N}\mathbb{E}[X_j] = \pi(j), \qquad \mathrm{Var}(d(j)) = \frac{1}{N^2}\mathrm{Var}(X_j) = \frac{1}{N}\pi(j)(1-\pi(j)),$$

so

$$\mathbb{E}[d(j)^2] = \mathrm{Var}(d(j)) + \mathbb{E}[d(j)]^2 = \pi(j)^2 + \frac{1}{N}\pi(j)(1-\pi(j)).$$

Likewise, for $j \ne \ell$,

$$\mathbb{E}[d(j)d(\ell)] = \frac{1}{N^2}\mathbb{E}[X_jX_\ell] = \frac{1}{N^2}\big(\mathrm{Cov}(X_j, X_\ell) + \mathbb{E}[X_j]\mathbb{E}[X_\ell]\big) = \pi(j)\pi(\ell) - \frac{1}{N}\pi(j)\pi(\ell),$$

because $\mathrm{Cov}(X_j, X_\ell) = -N\pi(j)\pi(\ell)$ for multinomial counts. □

Lemma 18 (Fourth moments and mixed fourth moments). Condition on π and let $d = X/N$ with $X \sim \mathrm{Multinomial}(N, \pi)$. Then for $j \ne \ell$,

$$\mathbb{E}[d(j)^2d(\ell)^2] \le C\Big(\pi(j)^2\pi(\ell)^2 + \frac{1}{N}\big(\pi(j)^2\pi(\ell) + \pi(j)\pi(\ell)^2\big) + \frac{1}{N^2}\pi(j)\pi(\ell)\Big),$$

and for each j,

$$\mathbb{E}[d(j)^4] \le C\Big(\pi(j)^4 + \frac{1}{N}\pi(j)^3 + \frac{1}{N^2}\pi(j)^2 + \frac{1}{N^3}\pi(j)\Big),$$

where $C > 0$ is an absolute constant.

Proof. For $j \ne \ell$, write

$$X_j^2X_\ell^2 = \big(X_j(X_j-1) + X_j\big)\big(X_\ell(X_\ell-1) + X_\ell\big)$$

and expand into four terms. Using the factorial-moment identities for multinomial counts,

$$\mathbb{E}[X_j(X_j-1)X_\ell(X_\ell-1)] = N(N-1)(N-2)(N-3)\pi(j)^2\pi(\ell)^2,$$

$$\mathbb{E}[X_j(X_j-1)X_\ell] = N(N-1)(N-2)\pi(j)^2\pi(\ell), \qquad \mathbb{E}[X_jX_\ell(X_\ell-1)] = N(N-1)(N-2)\pi(j)\pi(\ell)^2,$$

and

$$\mathbb{E}[X_jX_\ell] = N(N-1)\pi(j)\pi(\ell).$$

Dividing by N^4 and upper-bounding $(N-r)/N \le 1$ yields the displayed inequality for $\mathbb{E}[d(j)^2d(\ell)^2]$. The bound for $\mathbb{E}[d(j)^4]$ follows similarly by expanding X_j^4 into factorial moments and dividing by N^4. □

A.7.4 Column and operator bounds for the Gram fluctuation

For the identity-centered SVD Gram, define for a generic document

$$Q_i := M_0^{-1/2} d_i d_i^\top M_0^{-1/2}, \qquad d_i = X_i/N,$$

and the centered sampling fluctuation matrix

$$E := \frac{1}{n}\sum_{i=1}^{n}\{Q_i - \mathbb{E}[Q_i \mid W]\}.$$

When the generic corpus is client m, this matrix equals

$$E = \tilde{G}_m - \mathbb{E}[\tilde{G}_m \mid W^{(1:M)}] = \tilde{G}_m - G_{0,m} - D_{0,m}^{\mathrm{svd}}.$$

The deterministic diagonal term $D_{0,m}^{\mathrm{svd}}$ is not included in the centered fluctuation bounds below; it is controlled separately by Lemma 27.

Lemma 19 (Column variance bound for the centered whitened Gram fluctuation). Under Assumptions 5–6, for every $j \in [p]$,

$$\mathbb{E}\,\|Ee_j\|_2^2 \le C h_j \beta_n^2 \frac{p}{Nn}.$$

Proof. All expectations below are conditional on the fixed design. Fix a document with mean $\pi = Aw_i$ and define

$$\mu = M_0^{-1/2}\pi, \quad V_t = M_0^{-1/2}(e_{X_t} - \pi), \quad g = N^{-1}\sum_{t=1}^{N} V_t, \quad \zeta = g_j,$$

where the X_t are independent with distribution π. Regularity gives

$$\|\mu\|_2 \le C, \quad |\mu_j|^2 \le C h_j, \quad \|V_t\|_2 \le C\sqrt{p}, \quad \mathbb{E}\|V_t\|_2^2 \le Cp, \quad \mathbb{E}V_t(j)^2 \le C.$$

Set $H_t = V_t(j)$ and

$$a = \mathbb{E}\|V_t\|_2^2, \quad b = \mathbb{E}H_t^2, \quad c = \mathbb{E}(V_t H_t), \quad d = \mathbb{E}(\|V_t\|_2^2 H_t^2).$$

Then $a \le Cp$, $b \le C$, $\|c\|_2^2 \le ab \le Cp$, and $d \le Cp\,\mathbb{E}H_t^2 \le Cp$. Expanding the four-index sum, all singleton-index terms vanish by independence and centering. Thus

$$\mathbb{E}(\|g\|_2^2\zeta^2) = \frac{Nd + N(N-1)(ab + 2\|c\|_2^2)}{N^4} \le \frac{Cp}{N^2}. \tag{37}$$

The term d/N^3 accounts for coincident indices. Writing $Z_i = Q_i - \mathbb{E}Q_i$, we have

$$Z_i e_j = \mu\zeta + \mu_j g + \{g\zeta - \mathbb{E}(g\zeta)\}.$$

Since $\mathbb{E}\zeta^2 \le C/N$ and $\mathbb{E}\|g\|_2^2 \le Cp/N$, it follows that

$$\mathbb{E}\|Z_i e_j\|_2^2 \le C\left\{\frac{1}{N} + \frac{h_j p}{N} + \frac{p}{N^2}\right\} \le C h_j \beta_n^2 \frac{p}{N}.$$

For the last step, $h_j p \ge c$, and the definition of β_n implies $\beta_n \ge \max\{1, p/N\}$, hence $p/N \le C\beta_n^2$. Indeed, if $N \ge p$ the claim is immediate; if $N < p$, then $p^2/N^{3/2} \ge p/N$, so either branch in its definition is at least p/N. Finally the document summands are independent and centered, and therefore

$$\mathbb{E}\|Ee_j\|_2^2 = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\|Z_i e_j\|_2^2 \le C h_j \beta_n^2 \frac{p}{Nn}.$$

□

Lemma 20 (Operator-norm concentration for the centered Gram fluctuation). Under Assumptions 5–6, for every $u \ge 1$,

$$\mathbb{P}\left\{\|E\| > C\left[\beta_n\sqrt{\frac{pu}{Nn}} + \beta_n\sqrt{\frac{p}{N}}\frac{u}{n}\right]\right\} \le 2e^{-u}.$$

Consequently, for every $\delta \in (0,1)$,

$$\|E\| \le C\left[\beta_n\sqrt{\frac{p\log(2p/\delta)}{Nn}} + \beta_n\sqrt{\frac{p}{N}}\frac{\log(2p/\delta)}{n}\right]$$

with probability at least $1-\delta$. Moreover,

$$\mathbb{E}\|E\|^2 \le C\beta_n^2\frac{p}{Nn}, \qquad \mathbb{E}\|E\|^4 \le C\beta_n^4\left(\frac{p}{Nn}\right)^2.$$

Proof. Work in the Hilbert space of matrices with the Frobenius norm. For each document put $\mu_i = M_0^{-1/2}\pi_i$, $g_i = M_0^{-1/2}(d_i - \pi_i)$, and $Z_i = Q_i - \mathbb{E}Q_i$. Then

$$Z_i = \mu_i g_i^\top + g_i\mu_i^\top + \{g_i g_i^\top - \mathbb{E}(g_i g_i^\top)\}.$$

Using $\|\mu_i\|_2 \le C$ and Lemma 22, for every integer $q \ge 2$,

$$\left(\mathbb{E}\|Z_i\|_F^q\right)^{1/q} \le C\left\{\sqrt{\frac{qp}{N}} + \frac{qp}{N}\right\} \le Cq\beta_n\sqrt{\frac{p}{N}}.$$

Here the definition of β_n gives $\beta_n \ge \max\{1, p/N\}$: the claim is immediate for $N \ge p$, while for $N < p$ one has $p^2/N^{3/2} \ge p/N$, so both branches give the asserted bound. Since $q^q \le e^q q!$, the preceding estimate yields the Bernstein moment condition with per-document variance proxy $C\beta_n^2 p/N$ and scale proxy $C\beta_n\sqrt{p/N}$. Apply Lemma 23 to the independent centered matrices Z_i, divide by n, and use $\|E\| \le \|E\|_F$ to obtain the tail statement. The logarithmic specialization follows by increasing constants if necessary. Integration of the tail gives

$$\mathbb{E}\|E\|^2 \le C\beta_n^2\frac{p}{N}(n^{-1}+n^{-2}), \qquad \mathbb{E}\|E\|^4 \le C\beta_n^4\left(\frac{p}{N}\right)^2(n^{-2}+n^{-4}).$$

Using $n \ge 1$ proves the moment statements. No local perturbative scaling is needed for these moment bounds. □

Lemma 21 (Negligible operator-tail terms for fixed-whitening truncation). Assume Assumptions 5–6, the envelope bound in Lemma 7, and the perturbative scaling $\beta_n^2 p^2/(Nn) \le c_0$ with c_0 sufficiently small. Fix a constant $\gamma > 0$ bounded away from zero, and let

$$F := \tilde{G}_m - \mathbb{E}[\tilde{G}_m \mid W^{(1:M)}], \qquad B := D_{0,m}^{\mathrm{svd}}, \qquad E^{(s)} := F + B,$$

with $\|B\| \le C/N$, $\max_j h_j^{-1/2}\|Be_j\|_2 \le C\zeta_N^{\mathrm{svd}}$, and $\zeta_N^{\mathrm{svd}} \le c_0$. Set

$$a_n^{\mathrm{fw}} := \beta_n^2\frac{p}{Nn} + \{\zeta_N^{\mathrm{svd}}\}^2, \qquad \mathcal{T} := \{\|E^{(s)}\| \le \gamma/4\}.$$

Then, after increasing constants if necessary, uniformly in m and j,

$$\mathbb{P}(\mathcal{T}^c) \le Ch_j a_n^{\mathrm{fw}}, \qquad \mathbb{P}(\mathcal{T}^c) \le Ca_n^{\mathrm{fw}},$$

and

$$\mathbb{E}[\|F\|\mathbf{1}_{\mathcal{T}^c}] \le Ca_n^{\mathrm{fw}}, \qquad \mathbb{E}[\|Fe_j\|_2\mathbf{1}_{\mathcal{T}^c}] \le C\sqrt{h_j}\,a_n^{\mathrm{fw}}.$$

Proof. Put

$$a_F := \beta_n^2\frac{p}{Nn}.$$

Choose the universal constant c_0 in the perturbative assumptions sufficiently small that $\|B\| \le \gamma/16$. Then

$$\mathcal{T}^c \subseteq \{\|F\| > \gamma/8\}.$$

The fourth-moment part of Lemma 20 gives

$$\mathbb{E}\|F\|^4 \le Ca_F^2.$$

Markov's inequality therefore gives

$$\mathbb{P}(\mathcal{T}^c) \le Ca_F^2.$$

Since the perturbative scaling implies $a_F \le c_0/p$ and Assumption 5 gives $h_j \ge c/p$ for fixed K, we have

$$a_F^2 \le C h_j a_F \le C h_j a_n^{\mathrm{fw}}, \qquad a_F^2 \le C a_n^{\mathrm{fw}}.$$

This proves the two displayed probability bounds. Notice that no large-deviation tail at parameter $u \asymp p^2$ is being used here; the fourth-moment estimate is sufficient and avoids any restriction coming from bounded-increment terms in the operator tail.

For the truncated expectations, Cauchy–Schwarz and the second-moment bounds in Lemmas 20 and 19 give

$$\mathbb{E}[\|F\|\mathbf{1}_{\mathcal{T}^c}] \le (\mathbb{E}\|F\|^2)^{1/2}\mathbb{P}(\mathcal{T}^c)^{1/2} \le C a_F^{3/2} \le C a_n^{\mathrm{fw}},$$

because $a_F \le 1$ after decreasing c_0 if needed. Similarly,

$$\mathbb{E}[\|Fe_j\|_2 \mathbf{1}_{\mathcal{T}^c}] \le (\mathbb{E}\|Fe_j\|_2^2)^{1/2}\mathbb{P}(\mathcal{T}^c)^{1/2} \le C\sqrt{h_j}\, a_F^{3/2} \le C\sqrt{h_j}\, a_n^{\mathrm{fw}}.$$

The deterministic diagonal term B is included in a_n^{fw} through $\{\zeta_N^{\mathrm{svd}}\}^2$, and its only role in the tail event is the preliminary inclusion $\mathcal{T}^c \subseteq \{\|F\| > \gamma/8\}$. □

Lemma 22 (Vector moments for whitened multinomial averages). Under Assumptions 5–6, let $d = N^{-1}\sum_{t=1}^N e_{X_t}$ with independent $X_t \sim \pi$, where $\pi = Aw$ for a fixed simplex weight w, and set

$$g = M_0^{-1/2}(d - \pi), \qquad \zeta_j = \frac{d(j) - \pi(j)}{\sqrt{m_0(j)}}.$$

Then

$$\mathbb{E}\|g\|_2^2 \le C\frac{p}{N}, \qquad \mathbb{E}|\zeta_j|^2 \le \frac{C}{N}.$$

For every real $q \ge 2$,

$$\left(\mathbb{E}\|g\|_2^q\right)^{1/q} \le C\sqrt{\frac{qp}{N}}, \qquad \left(\mathbb{E}|\zeta_j|^q\right)^{1/q} \le C\left\{\sqrt{\frac{q}{N}} + \frac{q}{N\sqrt{h_j}}\right\}. \tag{38}$$

Proof. Set $V_t = M_0^{-1/2}(e_{X_t} - \pi)$. The fixed-design envelope and $m_0(j) \asymp h_j \gtrsim p^{-1}$ give

$$\mathbb{E}V_t = 0, \quad \|V_t\|_2 \le C\sqrt{p}, \quad \mathbb{E}\|V_t\|_2^2 \le Cp, \quad \mathbb{E}V_t(j)^2 \le C, \quad |V_t(j)| \le C h_j^{-1/2}.$$

Independence proves the second-moment assertions. For $f(X_1, \dots, X_N) = \|N^{-1}\sum_t V_t\|_2$, changing one coordinate changes f by at most $C\sqrt{p}/N$, while $\mathbb{E}f \le C\sqrt{p/N}$. The bounded-differences inequality gives

$$\mathbb{P}\{f > \mathbb{E}f + t\} \le \exp(-cNt^2/p), \qquad t \ge 0.$$

For clarity, this inequality follows by exposing the word draws in order: each increment of the Doob martingale for f has conditional range at most $C\sqrt{p}/N$; the conditional bounded-variable exponential-moment bound and its product over N draws give $\mathbb{E}\exp\{\lambda(f - \mathbb{E}f)\} \le \exp(C\lambda^2 p/N)$. Chernoff's inequality gives the displayed tail, and integration gives the first inequality in (38).

For $\zeta_j = N^{-1}\sum_t V_t(j)$, scalar Bernstein has total variance at most C/N and summand range at most $C/(N\sqrt{h_j})$. Its tail integration yields

$$\|\zeta_j\|_{L^q} \le C\left\{\sqrt{q/N} + q/(N\sqrt{h_j})\right\},$$

as required. This uses a trace-scale sub-Gaussian bound only for the Euclidean norm of g; it does not assert a variance-only sub-Gaussian bound for every scalar whitened direction. □

Lemma 23 (A Hilbert-space Bernstein criterion). Conditionally on W, let $Z_1, \dots, Z_n$ be independent random vectors in a finite-dimensional Hilbert space with $\mathbb{E}[Z_i \mid W] = 0$. Suppose there are numbers $V \ge 0$ and $B > 0$, deterministic given W, such that

$$\sum_{i=1}^n \mathbb{E}[\|Z_i\|_2^2 \mid W] \le V, \qquad \sum_{i=1}^n \mathbb{E}[\|Z_i\|_2^q \mid W] \le \frac{q!}{2} V B^{q-2}, \qquad q \ge 3.$$

Then, for every $u \ge 0$,

$$\mathbb{P}\left(\left\|\sum_{i=1}^n Z_i\right\|_2 > C\{\sqrt{Vu} + Bu\} \;\middle|\; W\right) \le 2e^{-u}.$$

Proof. All expectations and probabilities in this proof are conditional on W. For $0 < \lambda < B^{-1}$, the moment assumptions and the exponential series give

$$\sum_i \mathbb{E}\big[e^{\lambda\|Z_i\|_2} - 1 - \lambda\|Z_i\|_2 \mid W\big] \le \frac{V}{2}\sum_{q\ge 2}\lambda^q B^{q-2} = \frac{V\lambda^2}{2(1-B\lambda)}.$$

The Hilbert-space case ($D = 1$) of Pinelis (1994, Theorem 3.1)[2] therefore gives

$$\mathbb{P}\left(\left\|\sum_i Z_i\right\|_2 > t \;\middle|\; W\right) \le 2\exp\left\{-\lambda t + \frac{V\lambda^2}{2(1-B\lambda)}\right\}.$$

For $V > 0$ and $u > 0$, take $s = \sqrt{2u/V}$, $\lambda = s/(1+Bs)$, and $t = \sqrt{2Vu} + Bu$. The exponent is $-u$, proving the stronger threshold $\sqrt{2Vu} + Bu$ and hence the displayed claim. If $V = 0$, every Z_i vanishes almost surely conditionally on W; the case $u = 0$ is trivial. □

Lemma 24 (Single-column Euclidean-norm concentration). Under Assumptions 5–6, fix a client m and a word j. Let

$$Z_{i,j}^{(m)} := Q_i^{(m)} e_j - \mathbb{E}[Q_i^{(m)} e_j \mid W^{(1:M)}], \qquad i = 1, \dots, n,$$

where $Q_i^{(m)}$ is the single-document whitened outer-product matrix defined at the beginning of Appendix A.7. Thus $Q_i^{(m)}$ itself is not centered; the vector $Z_{i,j}^{(m)}$ is the centered column summand. Then, for every $u \ge 1$,

$$\mathbb{P}\left(h_j^{-1/2}\left\|\frac{1}{n}\sum_{i=1}^n Z_{i,j}^{(m)}\right\|_2 > C\left[\beta_n\sqrt{\frac{pu}{Nn}} + \beta_n^2\frac{pu}{Nn} + \sqrt{\beta_n p}\,\frac{pu}{Nn} + \beta_n\sqrt{\frac{p}{N}}\frac{u}{n}\right] \;\middle|\; W^{(1:M)}\right) \le 2e^{-u}.$$

Proof. All expectations are conditional on the fixed design. For a generic document with mean $\pi = Aw_i$, write

$$\mu = M_0^{-1/2}\pi, \qquad g = M_0^{-1/2}(d_i - \pi), \qquad \zeta = g_j.$$

Abbreviate $h = h_j$ and $\beta = \beta_n$. Define

$$v = h\beta^2\frac{p}{N}, \qquad b = \sqrt{h}\left\{\beta\sqrt{\frac{p}{N}} + \beta^2\frac{p}{N} + \sqrt{\beta p}\frac{p}{N}\right\}.$$

We will prove for every integer $q \ge 3$ that

$$\mathbb{E}\|(Q_i - \mathbb{E}Q_i)e_j\|_2^q \le \frac{q!}{2}(Cv)(Cb)^{q-2}. \tag{39}$$

Write

$$(Q_i - \mathbb{E}Q_i)e_j = \mu\zeta + \mu_j g + \{g\zeta - \mathbb{E}(g\zeta)\}.$$

The first two terms satisfy

$$\|\mu\zeta\|_{L^q(\ell_2)} \le C\left\{\sqrt{q/N} + q/(N\sqrt{h})\right\} \le Cq\sqrt{v}, \qquad \|\mu_j g\|_{L^q(\ell_2)} \le C\sqrt{qhp/N} \le Cq\sqrt{v}.$$

Indeed $hp \ge c$, $\beta \ge 1$, and $\beta^2 N \ge p$ imply $N^{-1} \le Cv$ and $(N^2 h)^{-1} \le Cv$. Since $b \ge \sqrt{v}$ and $q^q \le e^q q!$, both linear terms have the Bernstein moments in (39).

For the product $Y = g\zeta$, let

$$a = \frac{\sqrt{p}}{N}, \qquad d = \frac{\sqrt{p}}{N^{3/2}\sqrt{h}}, \qquad L = \frac{C\sqrt{p}}{\sqrt{h}}.$$

[2] I. Pinelis, "Optimum Bounds for the Distributions of Martingales in Banach Spaces," *The Annals of Probability*, 22(4):1679–1706, 1994. https://doi.org/10.1214/aop/1176988477.

Hölder's inequality and (38) yield

$$\|Y\|_{L^q(\ell_2)} \le \|g\|_{L^{2q}(\ell_2)}\|\zeta\|_{L^{2q}} \le C(qa + q^{3/2}d).$$

Moreover $\|Y\|_2 \le L$ almost surely, and (37) gives $\mathbb{E}\|Y\|_2^2 \le Cv$. The following comparisons are all deterministic:

$$a^2 \le Cv, \qquad d^2 \le Cv, \qquad b \ge \sqrt{v}, \qquad \frac{b}{d} \ge c\sqrt{\beta N}, \qquad \frac{L}{b} \le \frac{CN}{\sqrt{\beta}}.$$

For example,

$$\frac{a^2}{v} = \frac{1}{h\beta^2 N} \le C, \qquad \frac{d^2}{v} = \frac{1}{h^2\beta^2 N^2} \le C,$$

using $h \ge c/p$, $\beta^2 N \ge p$, and $\beta N \ge p$. The last two comparisons follow from the component $\sqrt{h}\sqrt{\beta p}\, p/N$ of b and $hp \ge c$.

If $3 \le q \le \beta N$, then the $(qa)^q$ part is at most $q!C^q v b^{q-2}$. For the other part, $q^q \le e^q q!$ gives

$$q^{3q/2}d^q \le q!C^q v b^{q-2} q \left(\frac{q}{\beta N}\right)^{(q-2)/2} \le q!C_1^q v b^{q-2},$$

where the final factor q is absorbed into C_1^q. If $q > \beta N$, use instead the deterministic bound and the second moment:

$$\mathbb{E}\|Y\|_2^q \le L^{q-2}\mathbb{E}\|Y\|_2^2 \le Cv(Cqb)^{q-2} \le q!C_2^q v b^{q-2}.$$

Here $L/b \le CN/\sqrt{\beta} \le Cq$ since $q > \beta N$ and $\beta \ge 1$. Thus the required Bernstein moment bound holds for every $q \ge 3$. Centering Y changes constants only, by Jensen's inequality and $\|Y - \mathbb{E}Y\|_{L^q} \le 2\|Y\|_{L^q}$.

Combining the three terms establishes (39). The second-moment bound is

$$\mathbb{E}\|(Q_i - \mathbb{E}Q_i)e_j\|_2^2 \le C\{N^{-1} + hp/N + p/N^2\} \le Cv,$$

as follows from the same decomposition and (37). The Hilbert-space Bernstein criterion for the independent document summands therefore gives

$$\left\|\frac{1}{n}\sum_i (Q_i - \mathbb{E}Q_i)e_j\right\|_2 \le C\left\{\sqrt{\frac{vu}{n}} + \frac{bu}{n}\right\}$$

with probability at least $1 - 2e^{-u}$. Dividing by $\sqrt{h}$ gives the claimed four-term bound. □

Lemma 25 (Uniform clientwise Gram-column concentration). Under Assumptions 5–6, for every $\delta \in (0,1)$,

$$\mathbb{P}\left(\max_{1\le m\le M}\max_{1\le j\le p} h_j^{-1/2}\|(\tilde{G}_m - G_{0,m})e_j\|_2 \le C\{\vartheta_n(\delta) + \zeta_N^{\mathrm{svd}}\}\right) \ge 1 - \delta.$$

Proof. For fixed (m, j),

$$(\tilde{G}_m - G_{0,m})e_j = \frac{1}{n}\sum_{i=1}^n Z_{i,j}^{(m)} + D_{0,m}^{\mathrm{svd}}e_j,$$

where $Z_{i,j}^{(m)}$ is the centered sampling column from Lemma 24. Apply Lemma 24 with $u = \ell_{n,M}(\delta) = \log(2pM/\delta)$ and take a union bound over all $(m, j) \in [M]\times[p]$. Lemma 27 controls the deterministic diagonal term. The resulting right-hand side is $C\{\vartheta_n(\delta) + \zeta_N^{\mathrm{svd}}\}$. □

Lemma 26 (Concentration of the pooled marginal matrix). Under Assumptions 5–6 and the conditional fixed-design setup, suppose all documents have common length N and $n_m = n$. Then for every $\delta \in (0,1)$, with probability at least $1 - \delta$,

$$\Delta_1(D, D_0) \le C\sqrt{\frac{p\log(2p/\delta)}{NnM}} + C\frac{p\log(2p/\delta)}{NnM}.$$

In the perturbative regime $p\log(2p/\delta)/(NnM) \le 1$, the second term is absorbed by the first.

Proof. For each j, write $M_d(j,j)$ as the average of the NnM independent word indicators in the corpus, conditionally on $W^{(1:M)}$. The centered summands have range at most $1/(NnM)$, their total variance is at most $M_0(j,j)/(NnM)$, and the mean is $M_0(j,j)$. A scalar Bernstein inequality combined with a union bound over $j \in [p]$ yields

$$|M_d(j,j) - M_0(j,j)| \le C\sqrt{\frac{M_0(j,j)\log(2p/\delta)}{NnM}} + C\frac{\log(2p/\delta)}{NnM}$$

for all j with probability at least $1-\delta$. Since $M_0(j,j) \asymp h_j$ and $h_j \gtrsim p^{-1}$, division by h_j gives the claim. □

Lemma 27 (Deterministic size of the ordinary-SVD diagonal bias). Under Assumptions 5–6,

$$\Delta_{\mathrm{svd}} \le C\zeta_N^{\mathrm{svd}} = C\frac{\sqrt{p}}{N}.$$

Proof. The matrix $D_{0,m}^{\mathrm{svd}}$ is diagonal, with jth diagonal entry

$$\frac{1}{N}\left\{\frac{M_{0,m}(j,j)}{M_0(j,j)} - 1\right\}.$$

Since $M_{0,m}(j,j) \le h_j$ and $M_0(j,j) \asymp h_j$ under Assumptions 5–6, the absolute value of the ratio in braces is uniformly bounded by a constant. Hence

$$\|D_{0,m}^{\mathrm{svd}}\| \le \frac{C}{N}, \qquad h_j^{-1/2}\|D_{0,m}^{\mathrm{svd}} e_j\|_2 \le \frac{C}{N\sqrt{h_j}} \le C\frac{\sqrt{p}}{N}.$$

Taking the maximum over m and j proves the claim. □

A.7.5 Proof of Lemma 16

Proof. The first display is exactly Lemma 19. The operator-norm second- and fourth-moment bounds follow from Lemma 20. The final display is Lemma 26. Collecting these ingredients proves the claim. □

Remark 23. Lemma 16 follows from elementary multinomial moments and Bernstein concentration for independent documents. The pooled marginal concentration controls the shared empirical whitening matrix M_d in the bridge terms.

A.8 A row-wise eigenvector perturbation lemma

The next lemma is the row-wise perturbation device underlying the singular-vector analysis in Topic-SCORE. It is stated abstractly because it is useful beyond the multinomial model itself.

Lemma 28 (A full-cluster row-wise perturbation bound). Let G_0 and $G = G_0 + Y$ be symmetric $p \times p$ matrices. Assume $G_0 \succeq 0$ has rank K, and let $U_0 \in \mathbb{R}^{p\times K}$ be an orthonormal basis for its range. Let $U \in \mathbb{R}^{p\times K}$ be an orthonormal basis for the leading K-dimensional invariant subspace of G. Suppose that, for some fixed $c \in (0,1)$,

$$\lambda_K(G_0) \ge c\|G_0\|, \qquad \|Y\| \le c\|G_0\|/8.$$

Then there exists an orthogonal matrix $O \in \mathbb{R}^{K\times K}$ such that, for every $j \in [p]$,

$$\|e_j^\top(UO - U_0)\|_2 \le \frac{C}{\|G_0\|}\left\{\|Y\|\,\|e_j^\top U_0\|_2 + \|Ye_j\|_2\right\},$$

where C depends only on c and fixed K.

Proof. This is the full-signal-cluster version of the row-wise perturbation lemma used in single-site Topic-SCORE; it is the only version needed in the present federated proof. Let $P_0 = U_0U_0^\top$ and $P = UU^\top$. Since the nonzero spectrum of G_0 is separated from the null space by at least $c\|G_0\|$ and $\|Y\| \le c\|G_0\|/8$, the leading K-dimensional invariant subspace of G is well defined and Davis–Kahan gives

$$\|P - P_0\| \le C\frac{\|Y\|}{\|G_0\|}.$$

Let O be the orthogonal polar factor of $U^\top U_0$; this choice minimizes $\|UO - U_0\|_F$. Standard projector-to-basis algebra gives

$$\|e_j^\top (UO - U_0)\|_2 \le C \left\{ \|(P - P_0)e_j\|_2 + \|P - P_0\| \, \|e_j^\top U_0\|_2 \right\}. \tag{40}$$

For completeness, one way to see (40) is to write $UO - U_0 = (P - P_0)U_0 + U\{O - U^\top U_0\}$ and use $\|O - U^\top U_0\| \le C\|P - P_0\|^2$, which follows from the singular values of $U^\top U_0$ being within $C\|P - P_0\|^2$ of one.

It remains to control the projector column. Apply the Riesz expansion for the spectral projector of the full nonzero cluster of G_0. Because the cluster is separated from the zero eigenvalue by a gap of order $\|G_0\|$, the first-order term has the cross-cluster form

$$\mathcal{L}(Y) = G_0^\dagger Y(I - P_0) + (I - P_0)YG_0^\dagger \qquad \text{on the range/null decomposition,}$$

where $G_0^\dagger$ denotes the Moore–Penrose inverse on $\mathrm{range}(G_0)$ and $\|G_0^\dagger\| \le C\|G_0\|^{-1}$. Hence

$$\|\mathcal{L}(Y)e_j\|_2 \le \frac{C}{\|G_0\|}\{\|Ye_j\|_2 + \|Y\|\|P_0e_j\|_2\}.$$

The quadratic Riesz remainder satisfies

$$\|\mathcal{R}(Y)e_j\|_2 \le C\frac{\|Y\|}{\|G_0\|}\,\frac{\|Ye_j\|_2 + \|Y\|\|P_0e_j\|_2}{\|G_0\|},$$

which is absorbed by the preceding display because $\|Y\| \le c\|G_0\|/8$. Therefore

$$\|(P - P_0)e_j\|_2 \le \frac{C}{\|G_0\|}\{\|Ye_j\|_2 + \|Y\|\|P_0e_j\|_2\}.$$

Since $\|P_0e_j\|_2 = \|e_j^\top U_0\|_2$, substituting this projector-column bound into (40) proves the lemma. □

Remark 24 (No client-specific canonical SCORE basis). Lemma 28 is a row-wise perturbation device for subspaces and projector signals; it should not be read as saying that every local fixed-whitening eigenbasis is aligned with the pooled canonical basis $[\xi_1, \Xi_\perp]$. In general,

$$G_{0,m} = \left(1 - \frac{1}{N}\right) B\Sigma_{W,m}B^\top, \qquad B = M_0^{-1/2}A,$$

has eigenvectors that depend on the client-specific matrix $\Sigma_{W,m}$. Thus the local population first eigenvector need not equal the pooled canonical direction $\xi_1 = M_0^{1/2}\mathbf{1}$, and local SCORE ratios built directly from a local eigenbasis need not be comparable to the oracle ratio matrix R.

Lemma 28 is used to obtain row-wise control from the averaged-projector signal in Lemma 9. The empirical-whitening projector bridge is established directly by the common-reference resolvent comparison in Step 3 of the proof of Lemma 8. The canonical SCORE denominator is recovered only after server-side projector averaging by projecting the pooled marginal direction. Consequently, we do not assert a local-ratio guarantee: clients compute local SVD eigenspaces only for communication, while topic recovery uses only the server-side canonical ratio cloud.

A.9 Proofs of the main theorems

A.9.1 Proof of Theorem 4

Proof. Apply Lemma 3 to decompose $\bar{\mathcal{P}} - P$ into variance and bias terms. The variance term is controlled by Lemma 4. This yields (20). The subspace conclusion follows from Lemma 2. □

A.9.2 Proof of Lemma 8

Proof. Fix a client m and suppress the client index. Write

$$F^{(s)} := \tilde{G}_m - \mathbb{E}[\tilde{G}_m \mid W^{(1:M)}], \qquad B^{\mathrm{svd}} := D_{0,m}^{\mathrm{svd}}, \qquad E^{(s)} := \tilde{G}_m - G_{0,m} = F^{(s)} + B^{\mathrm{svd}},$$

and

$$E^{(w)} := \hat{G}_m - \tilde{G}_m.$$

Let $\gamma_m := \lambda_K(G_{0,m})$. Lemma 37, Assumption 7, and the standing upper-bound assumption $N \geq 2$ imply $\gamma_m \geq c_{\mathrm{gap}} > 0$ uniformly in m; the factor $1 - 1/N$ is bounded below by $1/2$ and is absorbed into c_{gap}.

By Lemma 16 and Lemma 27,

$$\mathbb{E}\|F^{(s)}e_j\|_2^2 \leq Ch_j\beta_n^2\frac{p}{Nn}, \qquad \mathbb{E}\|F^{(s)}\|^2 \leq C\beta_n^2\frac{p}{Nn}, \qquad \mathbb{E}\|F^{(s)}\|^4 \leq C\beta_n^4\left(\frac{p}{Nn}\right)^2,$$

and

$$\|B^{\mathrm{svd}}\| \leq \frac{C}{N}, \qquad \max_j h_j^{-1/2}\|B^{\mathrm{svd}}e_j\|_2 \leq C\zeta_N^{\mathrm{svd}}.$$

Step 1: fixed-whitening second moments. Set

$$a_n^{\mathrm{fw}} := \beta_n^2\frac{p}{Nn} + \{\zeta_N^{\mathrm{svd}}\}^2.$$

Let $\mathcal{E}_m := \{\|E^{(s)}\| \leq \gamma_m/4\}$. Lemma 21, applied with the constant gap $\gamma_m \geq c_{\mathrm{gap}}$, gives

$$\mathbb{P}(\mathcal{E}_m^c) \leq Ch_j a_n^{\mathrm{fw}}, \qquad \mathbb{P}(\mathcal{E}_m^c) \leq Ca_n^{\mathrm{fw}}.$$

On $\mathcal{E}_m$, Lemma 14 gives

$$\|(\tilde{P}_m - P)e_j\|_2 \leq C\{\|E^{(s)}e_j\|_2 + \|E^{(s)}\|\,\|Pe_j\|_2\} + C\|E^{(s)}\|^2.$$

The oracle row-scale bound gives $\|Pe_j\|_2 \leq C\sqrt{h_j}$. On $\mathcal{E}_m^c$, use the crude bound $\|(\tilde{P}_m - P)e_j\|_2 \leq 2$. Therefore

$$\mathbb{E}\|(\tilde{P}_m - P)e_j\|_2^2 \leq C\mathbb{E}\Big[\{\|E^{(s)}e_j\|_2^2 + h_j\|E^{(s)}\|^2 + \|E^{(s)}\|^4\}\mathbf{1}_{\mathcal{E}_m}\Big] + 4\mathbb{P}(\mathcal{E}_m^c).$$

The first two terms are bounded by

$$\mathbb{E}\|E^{(s)}e_j\|_2^2 + h_j\mathbb{E}\|E^{(s)}\|^2 \leq Ch_j\left(\beta_n^2\frac{p}{Nn} + (\zeta_N^{\mathrm{svd}})^2\right).$$

For the quadratic projector-remainder term, the refined bookkeeping is

$$\mathbb{E}\|E^{(s)}\|^4 \leq C\left(\beta_n^2\frac{p}{Nn}\right)^2 + CN^{-4} \leq Ch_j\left(\beta_n^2\frac{p}{Nn} + (\zeta_N^{\mathrm{svd}})^2\right).$$

The last inequality uses $\beta_n^2 p^2/(Nn) \leq c_0$, Assumption 5, and $\zeta_N^{\mathrm{svd}} = \sqrt{p}/N$ together with the sharper operator bound $\|B^{\mathrm{svd}}\| \leq C/N$. Combining these estimates with Lemma 21 gives

$$\mathbb{E}\|(\tilde{P}_m - P)e_j\|_2^2 \leq Ch_j\left(\beta_n^2\frac{p}{Nn} + (\zeta_N^{\mathrm{svd}})^2\right).$$

The same argument without the column scale gives

$$\mathbb{E}\|\tilde{P}_m - P\|^2 \leq C\left(\beta_n^2\frac{p}{Nn} + (\zeta_N^{\mathrm{svd}})^2\right).$$

Step 2: fixed-whitening projector bias. On $\mathcal{E}_m$, the Riesz expansion gives

$$\tilde{P}_m - P = L_m(F^{(s)}) + L_m(B^{\mathrm{svd}}) + R_m(E^{(s)}).$$

The full, untruncated linear term satisfies $\mathbb{E}L_m(F^{(s)}) = 0$. Hence

$$\mathbb{E}\{L_m(F^{(s)})\mathbf{1}_{\mathcal{E}_m}\} = -\mathbb{E}\{L_m(F^{(s)})\mathbf{1}_{\mathcal{E}_m^c}\}.$$

Lemma 21 and the column form of the linear projector derivative yield

$$\left\|\mathbb{E}\{L_m(F^{(s)})\mathbf{1}_{\mathcal{E}_m^c}\}\right\| \leq Ca_n^{\mathrm{fw}},$$

and

$$\left\|\mathbb{E}\{L_m(F^{(s)})e_j\mathbf{1}_{\mathcal{E}_m^c}\}\right\|_2 \leq C\sqrt{h_j}\,a_n^{\mathrm{fw}}.$$

The linear deterministic diagonal-bias term is bounded by

$$\|L_m(B^{\mathrm{svd}})\| + \max_j h_j^{-1/2}\|L_m(B^{\mathrm{svd}})e_j\|_2 \leq C\zeta_N^{\mathrm{svd}}.$$

The second-order remainder is controlled on $\mathcal{E}_m$ by the rank-K signal version of the refined columnwise Riesz remainder bound (36) in Lemma 14. This specialized version applies here because $G_{0,m} \succeq 0$ has rank K, its non-target spectrum is exactly $\{0\}$, and P projects onto $\mathrm{range}(G_{0,m})$. Equivalently, it is the rank-K column-remainder part of Lemma 5 without using that lemma's mean-zero hypothesis. The mean-zero condition is used only for the full linear fluctuation $L_m(F^{(s)})$. Specifically,

$$\mathbb{E}\{\|R_m(E^{(s)})e_j\|_2 \mathbf{1}_{\mathcal{E}_m}\} \le C\mathbb{E}\big[\|E^{(s)}\|\{\|E^{(s)}e_j\|_2 + \|E^{(s)}\|\|Pe_j\|_2\}\big] \le C\sqrt{h_j}\left(\beta_n^2\frac{p}{Nn} + (\zeta_N^{\mathrm{svd}})^2\right),$$

where the final step uses Cauchy–Schwarz, the column and operator second-moment bounds above, and $\|Pe_j\|_2 \le C\sqrt{h_j}$. The corresponding operator remainder satisfies

$$\mathbb{E}\{\|R_m(E^{(s)})\|\mathbf{1}_{\mathcal{E}_m}\} \le C\mathbb{E}\|E^{(s)}\|^2 \le C\left(\beta_n^2\frac{p}{Nn} + (\zeta_N^{\mathrm{svd}})^2\right).$$

The complement contributes at most the negligible tail terms from Lemma 21. Since $\zeta_N^{\mathrm{svd}} \le c_0$, the term $(\zeta_N^{\mathrm{svd}})^2$ is absorbed by ζ_N^{svd}. Thus

$$\|\mathbb{E}\tilde{P}_m - P\| \le C\left(\beta_n^2\frac{p}{Nn} + \zeta_N^{\mathrm{svd}}\right),$$

and

$$\|(\mathbb{E}\tilde{P}_m - P)e_j\|_2 \le C\sqrt{h_j}\left(\beta_n^2\frac{p}{Nn} + \zeta_N^{\mathrm{svd}}\right).$$

Step 3: empirical-whitening bridge for the ordinary-SVD projector. Work on $\mathcal{E}_{\mathrm{loc},m} \cap \mathcal{E}_{\mathrm{bridge},m}$. Lemma 47 gives

$$\|E^{(w)}\| \le C\Delta_1(D, D_0)\{1 + \|E^{(s)}\|\},$$

and

$$\|E^{(w)}e_j\|_2 \le C\sqrt{h_j}\Delta_1(D, D_0) + C\Delta_1(D, D_0)\|E^{(s)}e_j\|_2 + C\frac{\Delta_1(D, D_0)}{N}.$$

Choosing c_{br} sufficiently small ensures $\|E^{(s)}\| + \|E^{(w)}\| \le c_{\mathrm{gap}}/4$. We compare both perturbed matrices through the common rank-K reference $G_{0,m}$. Choose a contour $\mathcal{C}$ enclosing the nonzero spectrum of $G_{0,m}$, at distance at least $c_{\mathrm{gap}}/2$ from its spectrum, and excluding zero. The population spectral bounds imply that the contour length and $|z|^{-1}$ on $\mathcal{C}$ are bounded by fixed constants. Define

$$\widetilde{R}(z) = (zI - \tilde{G}_m)^{-1}, \qquad \widehat{R}(z) = (zI - \hat{G}_m)^{-1}.$$

Their operator norms are uniformly bounded on $\mathcal{C}$. The Riesz formula and the resolvent identity give

$$\hat{P}_m - \tilde{P}_m = \frac{1}{2\pi i}\oint_{\mathcal{C}} \widehat{R}(z)E^{(w)}\widetilde{R}(z)\,dz.$$

For each column, the exact identity

$$\widetilde{R}(z)e_j = z^{-1}e_j + z^{-1}\widetilde{R}(z)(G_{0,m} + E^{(s)})e_j$$

implies

$$\|E^{(w)}\widetilde{R}(z)e_j\|_2 \le C\left[\|E^{(w)}e_j\|_2 + \|E^{(w)}\|\{\|G_{0,m}e_j\|_2 + \|E^{(s)}e_j\|_2\}\right].$$

Integrating and using $\|G_{0,m}e_j\|_2 \le C\sqrt{h_j}$ therefore gives

$$\|(\hat{P}_m - \tilde{P}_m)e_j\|_2 \le C\left[\|E^{(w)}e_j\|_2 + \|E^{(w)}\|\{\sqrt{h_j} + \|E^{(s)}e_j\|_2\}\right].$$

Substituting the preceding bridge bounds and using $\|E^{(s)}\| \le c_{\mathrm{gap}}/8$ yields

$$\|(\hat{P}_m - \tilde{P}_m)e_j\|_2 \le C\sqrt{h_j}\Delta_1(D, D_0) + C\Delta_1(D, D_0)\|E^{(s)}e_j\|_2 + C\frac{\Delta_1(D, D_0)}{N}.$$

This is stronger than the detailed column bound in the statement, which retains an additional nonnegative quadratic remainder. The same contour integral gives $\|\hat{P}_m - \tilde{P}_m\| \le C\|E^{(w)}\| \le C\Delta_1(D, D_0)$. If $\mathcal{E}_{\mathrm{col},m}(\delta)$ also holds, divide by $\sqrt{h_j}$, take the maximum over j, and use $h_{\min}^{-1/2} \lesssim \sqrt{p}$ to obtain

$$\max_j h_j^{-1/2}\|(\hat{P}_m - \tilde{P}_m)e_j\|_2 \le C\Delta_1(D, D_0)\{1 + \vartheta_n(\delta) + \zeta_N^{\mathrm{svd}}\}.$$

Here the identity-shift contribution $\sqrt{p}\,\Delta_1(D, D_0)/N$ equals $\zeta_N^{\mathrm{svd}}\Delta_1(D, D_0)$. The simplified bridge bound in the statement follows. □

A.9.3 Proof of Corollary 1

Proof. Put $u = \log(2p/\delta)$ and

$$a_\delta = \beta_n \sqrt{\frac{pu}{Nn}} + \beta_n^2 \frac{pu}{Nn} + \sqrt{\beta_n p}\, \frac{pu}{Nn} + \beta_n \sqrt{\frac{p}{N}}\, \frac{u}{n}.$$

The operator concentration bound in Lemma 20 implies, with probability at least $1 - \delta$,

$$\|\tilde{G}_m - G_{0,m}\| \le C\{a_\delta + \zeta_N^{\mathrm{svd}}\}.$$

Work on this event. If $a_\delta + \zeta_N^{\mathrm{svd}} + \Delta_1(D, D_0)$ exceeds a sufficiently small fixed constant, the claimed bound follows from $\|\hat{P}_m - P\| \le 2$ after increasing its constant. Otherwise the local perturbation and marginal deviation are small. Lemma 47 then gives

$$\|\hat{G}_m - \tilde{G}_m\| \le C\Delta_1(D, D_0),$$

because $\|\tilde{G}_m + N^{-1} I_p\|$ is bounded. Davis–Kahan at $G_{0,m}$ consequently gives

$$\|\hat{P}_m - P\| \le C\{a_\delta + \zeta_N^{\mathrm{svd}} + \Delta_1(D, D_0)\},$$

which proves the corollary. □

A.9.4 Proof of Lemma 13

Proof. Suppress the final permutation of topic labels. Write

$$\Delta_1 := \Delta_1(D, D_0), \qquad \Delta_2 := \max_j h_j^{-1/2} \left\| e_j^\top \left([\hat{\xi}_1, \hat{\Xi}_\perp O_\perp] - [\xi_1, \Xi_\perp] \right) \right\|_2, \qquad E_{\mathrm{VH}} := \mathrm{Err}_{\mathrm{VH}}(O_\perp).$$

The canonical-basis error gives

$$|\hat{\xi}_1(j) - \xi_1(j)| \le \Delta_2 \sqrt{h_j}, \qquad \|e_j^\top (\hat{\Xi}_\perp O_\perp - \Xi_\perp)\|_2 \le \Delta_2 \sqrt{h_j}.$$

Since $\xi_1(j) \asymp \sqrt{h_j}$ and the perturbation is small, $\hat{\xi}_1(j) \asymp \sqrt{h_j}$. Therefore

$$O_\perp^\top \hat{r}_j - r_j = \frac{e_j^\top (\hat{\Xi}_\perp O_\perp - \Xi_\perp)}{\hat{\xi}_1(j)} + e_j^\top \Xi_\perp \left(\frac{1}{\hat{\xi}_1(j)} - \frac{1}{\xi_1(j)} \right),$$

and Lemma 40 implies

$$\max_j \|O_\perp^\top \hat{r}_j - r_j\|_2 \le C\Delta_2. \tag{41}$$

Let

$$Q = \begin{pmatrix} 1 & \cdots & 1 \\ v_1^\star & \cdots & v_K^\star \end{pmatrix}, \qquad \hat{Q} = \begin{pmatrix} 1 & \cdots & 1 \\ O_\perp^\top \hat{v}_1 & \cdots & O_\perp^\top \hat{v}_K \end{pmatrix}.$$

The oracle simplex is nondegenerate, so Q is uniformly well conditioned. Since $\|\hat{Q} - Q\| \le C E_{\mathrm{VH}}$, inverse perturbation and (41) give

$$\|\hat{\pi}_j^\star - \pi_j\|_2 \le C(\Delta_2 + E_{\mathrm{VH}}), \qquad \hat{\pi}_j^\star := \hat{Q}^{-1} \begin{pmatrix} 1 \\ O_\perp^\top \hat{r}_j \end{pmatrix}.$$

Truncation and renormalization are locally Lipschitz near the simplex, so

$$\|\hat{\pi}_j - \pi_j\|_1 \le C(\Delta_2 + E_{\mathrm{VH}}). \tag{42}$$

Using the oracle identity $A\,\mathrm{diag}(w) = M_0^{1/2}\,\mathrm{diag}(\xi_1)\Pi$, define

$$\hat{A}^* := M_d^{1/2}\,\mathrm{diag}(\hat{\xi}_1)\hat{\Pi}, \qquad A^* := A\,\mathrm{diag}(w) = M_0^{1/2}\,\mathrm{diag}(\xi_1)\Pi.$$

For each row,

$$\hat{a}_j^* - a_j^* = \sqrt{M_d(j,j)}\hat{\xi}_1(j)(\hat{\pi}_j - \pi_j) + \sqrt{M_d(j,j)}(\hat{\xi}_1(j) - \xi_1(j))\pi_j + \left(\sqrt{M_d(j,j)} - \sqrt{M_0(j,j)}\right)\xi_1(j)\pi_j.$$

Because $M_d(j,j) \asymp M_0(j,j) \asymp h_j$, $\xi_1(j) \asymp \sqrt{h_j}$, and $|\sqrt{M_d(j,j)} - \sqrt{M_0(j,j)}| \le C\Delta_1\sqrt{h_j}$, (42) yields

$$\|\hat{a}_j^* - a_j^*\|_1 \le Ch_j(\Delta_1 + \Delta_2 + E_{\mathrm{VH}}).$$

The coordinates of w are bounded away from zero by Lemma 38. Summing the last display over rows controls the column sums of $\hat{A}^*$ relative to those of $A^* = A\,\mathrm{diag}(w)$. For each column k, write $s_k = \|A_k^*\|_1 = w_k$ and $\hat{s}_k = \|\hat{A}_k^*\|_1$. Then $|\hat{s}_k - s_k| \le C(\Delta_1 + \Delta_2 + E_{\mathrm{VH}})$ and $s_k \asymp 1$. Hence

$$\left\| \frac{\hat{A}_k^*}{\hat{s}_k} - \frac{A_k^*}{s_k} \right\|_1 \le C\|\hat{A}_k^* - A_k^*\|_1 + C|\hat{s}_k - s_k|,$$

and the row-wise version of the same decomposition gives

$$\max_j \frac{\|\hat{a}_j - a_j\|_1}{h_j} \le C(\Delta_1 + \Delta_2 + E_{\mathrm{VH}}).$$

Summing over j gives the global $L(\hat{A}_{\mathrm{fed}}, A)$ bound. □

A.9.5 Proof of Theorem 5

Proof. Write

$$\bar{P} := \frac{1}{M}\sum_{m=1}^{M} \hat{P}_m, \qquad \bar{P}^{(0)} := \frac{1}{M}\sum_{m=1}^{M} \tilde{P}_m.$$

Let $\mathcal{E}_{\mathrm{avg}}(\delta/5)$ be the fixed-whitening averaging event supplied by Theorem 4 and Lemma 6, after the usual union bound over $j \in [p]$. Also define

$$\mathcal{E}_{\mathrm{loc}}(\delta/5) := \bigcap_{m=1}^{M} \mathcal{E}_{\mathrm{loc},m},$$

$$\mathcal{E}_{\mathrm{col}}(\delta/5) := \Big\{ \max_{1\le m\le M} \max_{1\le j\le p} h_j^{-1/2} \|(\tilde{G}_m - G_{0,m})e_j\|_2 \le C\{\vartheta_n(\delta) + \zeta_N^{\mathrm{svd}}\} \Big\},$$

$$\mathcal{E}_{\mathrm{marg}}(\delta/5) := \Big\{ \Delta_1(D, D_0) \le C\eta_{n,M}^{\mathrm{marg}}(\delta) \Big\}, \qquad \mathcal{E}_{\mathrm{bridge}}(\delta/5) := \bigcap_{m=1}^{M} \mathcal{E}_{\mathrm{bridge},m}.$$

Since $\eta_{n,M}^{\mathrm{marg}}(\delta)$ is included as a positive component of $\varepsilon_{n,M}(\delta)$, choosing the perturbative constant c_* sufficiently small ensures that $C\eta_{n,M}^{\mathrm{marg}}(\delta) \le c_{\mathrm{br}}$. Hence, on $\mathcal{E}_{\mathrm{loc}}(\delta/5) \cap \mathcal{E}_{\mathrm{marg}}(\delta/5)$, the smallness condition $\Delta_1(D, D_0) \le c_{\mathrm{br}}$ in $\mathcal{E}_{\mathrm{bridge},m}$ is satisfied and Lemma 47 implies $\mathcal{E}_{\mathrm{bridge}}(\delta/5)$. Thus the bridge event is a deterministic consequence of the local and marginal events in the perturbative regime.

By Theorem 4, Lemma 6, Lemma 20 with a union bound over m, Lemma 25, Lemma 26, Lemma 47, and the deterministic bound in Lemma 27,

$$\mathcal{E}_*(\delta) := \mathcal{E}_{\mathrm{avg}}(\delta/5) \cap \mathcal{E}_{\mathrm{loc}}(\delta/5) \cap \mathcal{E}_{\mathrm{col}}(\delta/5) \cap \mathcal{E}_{\mathrm{marg}}(\delta/5) \cap \mathcal{E}_{\mathrm{bridge}}(\delta/5)$$

has probability at least $1 - \delta$.

On $\mathcal{E}_{\mathrm{avg}}(\delta/5)$, the independent fixed-whitening ordinary-SVD projectors $\tilde{P}_1, \ldots, \tilde{P}_M$ satisfy

$$\|\bar{P}^{(0)} - P\| + \max_j h_j^{-1/2} \|(\bar{P}^{(0)} - P)e_j\|_2 \le C\left[\beta_n \sqrt{\frac{p\ell_{n,M}(\delta)}{NnM}} + \beta_n^2 \frac{p\ell_{n,M}(\delta)}{Nn} + \zeta_N^{\mathrm{svd}} + \frac{\sqrt{p}\,\ell_{n,M}(\delta)}{M} \right]. \tag{43}$$

The vector-Bernstein variance term associated with the deterministic diagonal bias also produces $\zeta_N^{\mathrm{svd}}\sqrt{\ell_{n,M}(\delta)/M}$. This is absorbed into the displayed bound using

$$\zeta_N^{\mathrm{svd}}\sqrt{\ell_{n,M}(\delta)/M} \le \zeta_N^{\mathrm{svd}} + \ell_{n,M}(\delta)/M \le \zeta_N^{\mathrm{svd}} + \sqrt{p}\,\ell_{n,M}(\delta)/M.$$

Next consider the empirical-whitening bridge from $\tilde{P}_m$ to $\hat{P}_m$. On $\mathcal{E}_{\mathrm{loc}}(\delta/5)$, the local operator norm obeys $\|\tilde{G}_m - G_{0,m}\| \le c_{\mathrm{gap}}/8$, and the deterministic bound $\|G_{0,m}\| \le C$ follows from

Assumptions 5–7. Hence $\|\tilde{G}_m\| \le C$. On $\mathcal{E}_{\mathrm{loc}}(\delta/5) \cap \mathcal{E}_{\mathrm{marg}}(\delta/5)$, the preceding paragraph ensures $\mathcal{E}_{\mathrm{bridge}}(\delta/5)$. Hence Lemma 8 gives

$$\|\hat{P}_m - \tilde{P}_m\| \le C\eta_{n,M}^{\mathrm{marg}}(\delta), \qquad 1 \le m \le M, \tag{44}$$

and

$$\max_{1\le j\le p} h_j^{-1/2}\|(\hat{P}_m - \tilde{P}_m)e_j\|_2 \le C\eta_{n,M}^{\mathrm{marg}}(\delta)\{1 + \vartheta_n(\delta) + \zeta_N^{\mathrm{svd}}\} + C\sqrt{p}\,\{\eta_{n,M}^{\mathrm{marg}}(\delta)\}^2. \tag{45}$$

Averaging (44) and (45) over m yields the same bound for $\bar{P} - \bar{P}^{(0)}$.

Combining the fixed-whitening ordinary-SVD term and the empirical-whitening bridge on $\mathcal{E}_*(\delta)$ gives

$$\|\bar{P} - P\| + \max_j h_j^{-1/2}\|(\bar{P} - P)e_j\|_2 \le C\varepsilon_{n,M}(\delta).$$

Under the deterministic perturbative assumption $\varepsilon_{n,M}(\delta) \le c_*$, Lemma 9 transfers this control to the extracted spectral projector:

$$\|\widehat{P} - P\| + \max_j h_j^{-1/2}\|(\widehat{P} - P)e_j\|_2 \le C\varepsilon_{n,M}(\delta).$$

Lemma 10 then gives the row-wise bound for $\hat{\xi}_1 - \xi_1$, and Lemma 11 gives an orthogonal $O_\perp$ such that

$$\max_j h_j^{-1/2}\|e_j^\top(\hat{\Xi}_\perp O_\perp - \Xi_\perp)\|_2 \le C\varepsilon_{n,M}(\delta).$$

Combining the first-column and residual-column bounds proves (22). The ratio consequence follows from the ratio algebra in the proof of Lemma 13. □

A.9.6 Proof of Corollary 2

Proof. The projector bound

$$\|\widehat{P} - P\| + \max_j h_j^{-1/2}\|(\widehat{P} - P)e_j\|_2 \le C\varepsilon_{n,M}(\delta)$$

is obtained explicitly in the proof of Theorem 5 before the canonical-direction step. □

A.9.7 Proof of Theorem 6

Proof. On the event of Theorem 5, $\Delta_2 \le C\varepsilon_{n,M}(\delta)$. Lemma 26 gives

$$\Delta_1(D, D_0) \le C\sqrt{\frac{p\ell_{n,M}(\delta)}{NnM}} + C\frac{p\ell_{n,M}(\delta)}{NnM}$$

after the usual union-bound adjustment. By the choice of c_* in Theorem 5, the ratio consequence of Theorem 5 satisfies the smallness hypothesis $C\varepsilon_{n,M}(\delta) \le c_{\mathrm{VH}}$ in Condition 1. Therefore Condition 1 gives $\mathrm{Err}_{\mathrm{VH}}(O_\perp) \le C\varepsilon_{n,M}(\delta)$. Substitution into Lemma 13 proves the row-wise and global topic-matrix bounds. □

A.9.8 Proof of Corollary 3

Proof. Set

$$T_{\mathrm{var}} := \beta_n\sqrt{\frac{p\ell_{n,M}}{NnM}}, \qquad T_{\mathrm{bias}} := \beta_n^2\frac{p\ell_{n,M}}{Nn}, \qquad T_{\mathrm{range}} := \frac{\sqrt{p}\,\ell_{n,M}}{M},$$

and

$$T_{\mathrm{br,row}} := \eta_{n,M}^{\mathrm{marg}}(\delta)\{1 + \vartheta_n(\delta) + \zeta_N^{\mathrm{svd}}\}, \qquad T_{\mathrm{br,quad}} := \sqrt{p}\,\{\eta_{n,M}^{\mathrm{marg}}(\delta)\}^2, \qquad T_{\mathrm{svd}} := \zeta_N^{\mathrm{svd}}.$$

Theorem 6 and the definition of $\varepsilon_{n,M}(\delta)$ yield

$$L(\hat{A}_{\mathrm{fed}}, A) \lesssim T_{\mathrm{var}} + T_{\mathrm{bias}} + T_{\mathrm{range}} + T_{\mathrm{br,row}} + T_{\mathrm{br,quad}} + T_{\mathrm{svd}}$$

up to the fixed-whitening quadratic term $(\beta_n^2 p/(Nn))^2$, which proves (30).

The variance-bias crossover is obtained by solving $T_{\text{bias}} \asymp T_{\text{var}}$:

$$\beta_n^2 \frac{p\ell_{n,M}}{Nn} \asymp \beta_n \sqrt{\frac{p\ell_{n,M}}{NnM}} \quad \Longleftrightarrow \quad M \asymp \frac{Nn}{\beta_n^2 p \ell_{n,M}}.$$

The conservative row-wise range term is no larger than the bias term when

$$\frac{\sqrt{p}\,\ell_{n,M}}{M} \lesssim \beta_n^2 \frac{p\ell_{n,M}}{Nn} \quad \Longleftrightarrow \quad M \gtrsim \frac{Nn}{\beta_n^2 \sqrt{p}}.$$

It is no larger than the variance term when

$$\frac{\sqrt{p}\,\ell_{n,M}}{M} \lesssim \beta_n \sqrt{\frac{p\ell_{n,M}}{NnM}} \quad \Longleftrightarrow \quad M \gtrsim \frac{\ell_{n,M} Nn}{\beta_n^2}.$$

These are exactly the comparisons stated in the corollary. □

A.9.9 Proof of Proposition 3 and Corollary 4

Proof of Proposition 3. The fixed-whitening projectors $\tilde{P}_1, \ldots, \tilde{P}_M$ are independent under the original conditional-on-$W^{(1:M)}$ probability law. Therefore Theorem 4 is applied unconditionally, before intersecting with the local event $\mathcal{E}_{\text{loc}}(\delta/2)$.

It remains to record the moment and bias inputs. For a fixed client, write

$$\tilde{G}_m = G_{0,m} + F_m + B_m, \qquad \mathbb{E}[F_m \mid W^{(1:M)}] = 0, \qquad B_m = D_{0,m}^{\text{svd}}.$$

Let

$$\mathcal{E}_m := \{\|F_m + B_m\| \le \rho_m/4\}.$$

On $\mathcal{E}_m$, Davis–Kahan gives

$$\|\tilde{P}_m - P\| \le \frac{C}{\rho_m}\|F_m + B_m\|.$$

On $\mathcal{E}_m^c$, use the crude bound $\|\tilde{P}_m - P\| \le 2$. Markov's inequality applied to $F_m + B_m$ gives

$$\mathbb{P}(\mathcal{E}_m^c) \le C\frac{v_m + b_m^2}{\rho_m^2}.$$

Consequently,

$$\nu_m := \mathbb{E}\|\tilde{P}_m - P\|^2 \le C\left\{\frac{v_m}{\rho_m^2} + \frac{b_m^2}{\rho_m^2}\right\}.$$

For the bias, the Riesz expansion on $\mathcal{E}_m$ gives

$$\tilde{P}_m - P = L_m(F_m + B_m) + R_m(F_m + B_m),$$

where L_m is the first-order spectral-projector derivative at $G_{0,m}$ and

$$\|L_m(B_m)\| \le C\frac{b_m}{\rho_m}, \qquad \|R_m(F_m + B_m)\| \le C\frac{\|F_m + B_m\|^2}{\rho_m^2}.$$

The constants in these expansion bounds do not require a bounded ratio $\|G_{0,m}\|/\rho_m$. To verify this, write $G_0 = G_{0,m}$, $H = F_m + B_m$, and $\rho = \rho_m$. On $\mathcal{E}_m$, $\|H\| \le \rho/4$. Along the vertical line $z = \rho/2 + it$, put

$$R_0(z) = (zI - G_0)^{-1}, \qquad R_H(z) = (zI - G_0 - H)^{-1}.$$

Since the spectrum of G_0 is contained in $\{0\} \cup [\rho, \infty)$, Weyl's inequality gives

$$\|R_0(\rho/2 + it)\| \le \{(\rho/2)^2 + t^2\}^{-1/2}, \qquad \|R_H(\rho/2 + it)\| \le \{(\rho/4)^2 + t^2\}^{-1/2}.$$

Close the contour in the right half-plane and let its radius tend to infinity. The integrals for the linear term and remainder converge absolutely, and the resolvent identity yields

$$\mathcal{L}_{G_0}(H) = -\frac{1}{2\pi}\int_{\mathbb{R}} R_0(z)HR_0(z)\,dt, \qquad \mathcal{R}_{G_0}(H) = -\frac{1}{2\pi}\int_{\mathbb{R}} R_0(z)HR_0(z)HR_H(z)\,dt.$$

Therefore

$$\|\mathcal{L}_{G_0}(H)\| \le C\rho^{-1}\|H\|, \qquad \|\mathcal{R}_{G_0}(H)\| \le C\|H\|^2 \int_{\mathbb{R}} \{(\rho/4)^2 + t^2\}^{-3/2}\, dt \le C\rho^{-2}\|H\|^2.$$

The same linear bound applies to B_m. This proves the stated gap dependence with constants independent of the largest population eigenvalue.

The untruncated linear term satisfies $\mathbb{E}L_m(F_m) = 0$, but the expansion is used only on $\mathcal{E}_m$. Therefore

$$\mathbb{E}\{L_m(F_m)\mathbf{1}_{\mathcal{E}_m}\} = -\mathbb{E}\{L_m(F_m)\mathbf{1}_{\mathcal{E}_m^c}\}.$$

By Cauchy–Schwarz and the displayed Markov bound,

$$\left\|\mathbb{E}\{L_m(F_m)\mathbf{1}_{\mathcal{E}_m^c}\}\right\| \le C\rho_m^{-1}(\mathbb{E}\|F_m\|^2)^{1/2}\mathbb{P}(\mathcal{E}_m^c)^{1/2} \le C\left(\frac{v_m}{\rho_m^2} + \frac{b_m^2}{\rho_m^2}\right).$$

The deterministic ordinary-SVD diagonal bias B_m remains at first order through $L_m(B_m)$, and the quadratic remainder contributes $v_m/\rho_m^2 + b_m^2/\rho_m^2$ on $\mathcal{E}_m$. The complement contributes the same order by the crude bound and the preceding tail estimate. Hence

$$\|\mathbb{E}\tilde{P}_m - P\| \le C\left(\frac{v_m}{\rho_m^2} + \frac{b_m}{\rho_m} + \frac{b_m^2}{\rho_m^2}\right). \tag{46}$$

Substituting these variance and bias inputs into Theorem 4 with failure probability $\delta/2$ yields an averaging event of probability at least $1 - \delta/2$ on which (32) holds. Intersecting this averaging event with $\mathcal{E}_{\mathrm{loc}}(\delta/2)$ gives total probability at least $1 - \delta$. If the right-hand side is at most $1/4$, Lemma 2 gives the same bound for the extracted top-K subspace. □

Proof of Corollary 4. Under the assumptions of the corollary and $\alpha_m = 1/M$,

$$\sum_{m=1}^{M} \alpha_m^2 \frac{v_m}{\rho_m^2} \le C\frac{\beta_n^2}{\rho_\star^2}\frac{p}{NnM}, \qquad \sum_{m=1}^{M} \alpha_m^2 \frac{b_m^2}{\rho_m^2} \le C\frac{(\zeta_N^{\mathrm{svd}})^2}{\rho_\star^2 M},$$

$$\sum_{m=1}^{M} \alpha_m \left\{\frac{v_m}{\rho_m^2} + \frac{b_m}{\rho_m} + \frac{b_m^2}{\rho_m^2}\right\} \le C\left\{\frac{\beta_n^2}{\rho_\star^2}\frac{p}{Nn} + \frac{\zeta_N^{\mathrm{svd}}}{\rho_\star}\right\}, \qquad \max_m \alpha_m = \frac{1}{M}.$$

Substitution into Proposition 3 proves (34). □

A.10 Proof of the lower bound

This proof follows the standard pLSI lower-bound construction, with the only change that the total number of independent documents is $n_{\mathrm{tot}} := nM$. We present a self-contained argument with explicit constructions.

A.10.1 A Fano-type lemma (Tsybakov)

We use a standard finite-packing version of Fano's method.

Lemma 29 (Fano for a finite packing). Let $\{P_\theta : \theta \in \Theta\}$ be a family of distributions and let $\ell(\cdot,\cdot)$ be a pseudometric on the parameter space. Assume there exist $\theta_0, \theta_1, \ldots, \theta_J$ such that

$$\ell(\theta_s, \theta_t) \ge 2r \quad \forall s \ne t, \qquad \mathrm{KL}(P_{\theta_j} \| P_{\theta_0}) \le \alpha \log J \quad \forall j \ge 1,$$

for some $\alpha \in (0, 1/8)$. Then for any estimator $\hat{\theta}$,

$$\sup_{\theta \in \{\theta_0, \ldots, \theta_J\}} \mathbb{P}_\theta\big(\ell(\hat{\theta}, \theta) \ge r\big) \;\ge\; c_0,$$

where $c_0 > 0$ is an absolute constant (depending only on α).

Proof. This is a standard application of Fano's inequality. Apply Fano's inequality to the finite parameter set $\{\theta_0, \theta_1, \ldots, \theta_J\}$ endowed with the loss-induced separation $\ell(\theta_s, \theta_t) \ge 2r$. The KL assumption implies that the average divergence from the reference point θ_0 is at most $\alpha \log J$, and since $\alpha < 1/8$, Tsybakov's theorem yields

$$\inf_{\hat{\theta}} \sup_{\theta \in \{\theta_0, \ldots, \theta_J\}} \mathbb{P}_\theta\big(\ell(\hat{\theta}, \theta) \ge r\big) \ge c_0$$

for an absolute constant $c_0 > 0$. This implies the displayed lower bound for every estimator $\hat{\theta}$ and proves the lemma. □

A.10.2 A KL upper bound tailored to pLSI

This is the support-restricted KL estimate needed for the pLSI lower bound.

Lemma 30 (KL bound via support-restricted mean-matrix perturbation). Let $D_0, \tilde{D}_0 \in \mathbb{R}^{p\times n_{\mathrm{tot}}}$ be two matrices whose columns are probability vectors. Let P and $\tilde{P}$ be the joint laws of n_{tot} independent multinomial documents of length N with mean matrices D_0 and $\tilde{D}_0$, respectively. Define the perturbation support

$$\mathcal{S} := \{(j,i) : \tilde{D}_0(j,i) \neq D_0(j,i)\},$$

and assume that $D_0(j,i) > 0$ for every $(j,i) \in \mathcal{S}$. Let

$$\delta_{\mathcal{S}} := \max_{(j,i)\in\mathcal{S}} \frac{|\tilde{D}_0(j,i) - D_0(j,i)|}{D_0(j,i)}.$$

If $\delta_{\mathcal{S}} < 1$, then there exists a universal constant $C > 0$ such that, with

$$Q_{\mathcal{S}} := \sum_{(j,i)\in\mathcal{S}} \frac{|\tilde{D}_0(j,i) - D_0(j,i)|^2}{D_0(j,i)},$$

we have

$$\mathrm{KL}(\tilde{P}\|P) \leq (1 + C\delta_{\mathcal{S}})\, NQ_{\mathcal{S}}.$$

Proof. For a single document, KL between $\mathrm{Mult}(N,\eta)$ and $\mathrm{Mult}(N,\tilde{\eta})$ equals

$$N\sum_{j=1}^{p} \tilde{\eta}_j \log\frac{\tilde{\eta}_j}{\eta_j}.$$

Apply this columnwise with $\eta = D_0(:,i)$ and $\tilde{\eta} = \tilde{D}_0(:,i)$. If $(j,i) \notin \mathcal{S}$, then the corresponding summand is zero because $\tilde{D}_0(j,i) = D_0(j,i)$. Hence only entries in $\mathcal{S}$ matter. For $(j,i) \in \mathcal{S}$, write

$$x_{j,i} := \frac{\tilde{D}_0(j,i) - D_0(j,i)}{D_0(j,i)}, \qquad |x_{j,i}| \leq \delta_{\mathcal{S}} < 1.$$

Using

$$(1+x)\log(1+x) \leq x + \frac{1}{2}x^2 + C|x|^3 \qquad (|x| < 1)$$

and summing only over $(j,i) \in \mathcal{S}$ gives a linear term

$$N\sum_{(j,i)\in\mathcal{S}} D_0(j,i)x_{j,i} = N\sum_{(j,i)\in\mathcal{S}} \{\tilde{D}_0(j,i) - D_0(j,i)\}.$$

This term is zero column by column. Indeed, outside $\mathcal{S}$ the two matrices coincide, and both $D_0(:,i)$ and $\tilde{D}_0(:,i)$ are probability vectors; hence

$$\sum_{j:(j,i)\in\mathcal{S}} \{\tilde{D}_0(j,i) - D_0(j,i)\} = \sum_{j=1}^{p} \{\tilde{D}_0(j,i) - D_0(j,i)\} = 0.$$

The remaining quadratic and cubic terms satisfy

$$\frac{N}{2}\sum_{(j,i)\in\mathcal{S}} D_0(j,i)x_{j,i}^2 + CN\sum_{(j,i)\in\mathcal{S}} D_0(j,i)|x_{j,i}|^3 \leq NQ_{\mathcal{S}} + CN\delta_{\mathcal{S}}Q_{\mathcal{S}},$$

because $Q_{\mathcal{S}} = \sum_{(j,i)\in\mathcal{S}} D_0(j,i)x_{j,i}^2$ and $|x_{j,i}| \leq \delta_{\mathcal{S}}$ on $\mathcal{S}$. Hence

$$\mathrm{KL}(\tilde{P}\|P) \leq (1 + C\delta_{\mathcal{S}})NQ_{\mathcal{S}}.$$

This is exactly the stated bound. □

A.10.3 A least favorable subclass and packing construction

Set $n_{\mathrm{tot}} := nM$. We build a finite parameter set inside the matched theorem class $\Phi^{(M)}_{n,N,p}(K, c_1, c_2, c_{\mathrm{loc}})$.

Step 1: balanced-client pure-topic weights. Choose integers $n_1, \dots, n_K$ satisfying

$$n_k \in \{\lfloor n/K \rfloor, \lceil n/K \rceil\}, \qquad \sum_{k=1}^{K} n_k = n.$$

Use the same allocation at every client. For every client m, set

$$W^{(m)} = [\underbrace{e_1, \dots, e_1}_{n_1}, \underbrace{e_2, \dots, e_2}_{n_2}, \dots, \underbrace{e_K, \dots, e_K}_{n_K}] \in \mathbb{R}^{K \times n}. \tag{47}$$

Let $W = [W^{(1)}, \dots, W^{(M)}]$. Then

$$\Sigma_{W,m} = \Sigma_W = \mathrm{diag}(n_1/n, \dots, n_K/n) \succeq \frac{1}{2K} I_K, \qquad 1 \le m \le M,$$

because $n \ge K$ implies $\lfloor n/K \rfloor / n \ge 1/(2K)$. Thus the least favorable subclass satisfies both the pooled and the client-level richness assumptions uniformly.

Step 2: construct a baseline topic matrix $A^{(0)}$ with many anchor rows. Fix a probability vector $\eta^\star \in \mathbb{R}^K$ in the interior of the simplex, with distinct entries and

$$\frac{1}{2K} \le \eta^\star_k \le \frac{3}{2K}, \qquad k = 1, \dots, K.$$

Define $\eta := K\eta^\star$, so $\eta_k \in [1/2, 3/2]$ and $K^{-1} \sum_k \eta_k = 1$.

Choose constants $b_1 > 0$ and $b_2 \in (0, 1)$ such that $1/3 < b_1 b_2 < 1$, and define

$$\vartheta_k := \frac{1 - (1 - b_1 b_2)\eta_k}{K b_1 b_2}, \qquad k = 1, \dots, K.$$

If $b_1 b_2$ is chosen so that $3(1 - b_1 b_2)/2 < 1$, then each $\vartheta_k > 0$ and $\sum_k \vartheta_k = 1$.

Let $p_{\mathrm{anc}} := \lfloor b_2 p \rfloor$ and $p_{\mathrm{non}} := p - p_{\mathrm{anc}}$. Allocate the anchor rows topicwise by

$$p_{\mathrm{anc},k} := \lfloor \vartheta_k p_{\mathrm{anc}} \rfloor, \quad 1 \le k \le K - 1, \qquad p_{\mathrm{anc},K} := p_{\mathrm{anc}} - \sum_{k=1}^{K-1} p_{\mathrm{anc},k}.$$

Then $\sum_{k=1}^{K} p_{\mathrm{anc},k} = p_{\mathrm{anc}}$ and $p_{\mathrm{anc},k} = \vartheta_k p_{\mathrm{anc}} + O(1)$ for every k; in particular, for all sufficiently large p, each $p_{\mathrm{anc},k} \ge 1$. Define

$$\beta_{\mathrm{non}} := \frac{1 - b_1 b_2}{p_{\mathrm{non}}}, \qquad \alpha_{\mathrm{anc},k} := \frac{1 - (1 - b_1 b_2)\eta_k}{p_{\mathrm{anc},k}}, \qquad k = 1, \dots, K.$$

Partition the first p_{anc} rows into disjoint anchor index sets I^{anc}_k with $|I^{\mathrm{anc}}_k| = p_{\mathrm{anc},k}$. Define $A^{(0)}$ row-wise by

$$a_j^{(0)\top} = \alpha_{\mathrm{anc},k} e_k^\top, \qquad j \in I^{\mathrm{anc}}_k,$$

and, for non-anchor rows,

$$a_j^{(0)\top} = \beta_{\mathrm{non}} \eta^\top.$$

Each topic column now sums exactly to one, because

$$p_{\mathrm{anc},k} \alpha_{\mathrm{anc},k} + p_{\mathrm{non}} \beta_{\mathrm{non}} \eta_k = 1 - (1 - b_1 b_2)\eta_k + (1 - b_1 b_2)\eta_k = 1.$$

Also $A^{(0)} \ge 0$, every topic has $p_{\mathrm{anc},k}$ anchor rows, and all row sums are of order K/p.

Lemma 31 (Baseline parameter lies in the matched class). For fixed K and fixed interior $\eta^\star$, one can choose b_1, b_2 so that $(A^{(0)}, W^{(1:M)})$ satisfies the anchor-word condition, Assumption 5, Assumption 6, and Assumption 7 with constants bounded away from zero. In fact the same construction also yields strictly positive off-diagonal overlap $\min_{k,\ell} \Sigma_A(k, \ell) \ge c_2'$ for some $c_2' > 0$.

Proof. The balanced construction (47) gives $\lambda_{\min}(\Sigma_W), \lambda_{\min}(\Sigma_{W,m}) \geq 1/(2K)$ for all m. For $j \in I_k^{\mathrm{anc}}$, the row sum is $h_j = \alpha_{\mathrm{anc},k}$, while for non-anchor rows $h_j = K\beta_{\mathrm{non}}$. Because $p_{\mathrm{anc},k} = \vartheta_k p_{\mathrm{anc}} + O(1)$, $p_{\mathrm{anc}} \asymp p$, and

$$1 - (1 - b_1 b_2)\eta_k = K b_1 b_2 \vartheta_k,$$

we have $\alpha_{\mathrm{anc},k} \asymp K/p$. Also $K\beta_{\mathrm{non}} = K(1 - b_1 b_2)/p_{\mathrm{non}} \asymp K/p$. Hence $h_{\min} \asymp K/p$, proving Assumption 5.

It remains to check $\Sigma_A = A^{(0)\top} H^{-1} A^{(0)}$. Each anchor row in topic k contributes $\alpha_{\mathrm{anc},k} e_k e_k^\top$ to Σ_A, so the full anchor block contributes

$$\sum_{k=1}^{K} p_{\mathrm{anc},k} \alpha_{\mathrm{anc},k} e_k e_k^\top = \mathrm{diag}\big(1 - (1 - b_1 b_2)\eta_1, \ldots, 1 - (1 - b_1 b_2)\eta_K\big).$$

Each non-anchor row contributes $(\beta_{\mathrm{non}}/K)\eta\eta^\top$, hence the full non-anchor block contributes

$$p_{\mathrm{non}} \frac{\beta_{\mathrm{non}}}{K} \eta\eta^\top = \frac{1 - b_1 b_2}{K} \eta\eta^\top.$$

Therefore

$$\Sigma_A = \mathrm{diag}\big(1 - (1 - b_1 b_2)\eta_1, \ldots, 1 - (1 - b_1 b_2)\eta_K\big) + \frac{1 - b_1 b_2}{K} \eta\eta^\top.$$

Because η is interior and $3(1 - b_1 b_2)/2 < 1$, the diagonal term is uniformly positive and the rank-one term is positive semidefinite, so $\lambda_{\min}(\Sigma_A)$ is bounded below by a positive constant depending only on $(b_1, b_2, K, \eta^\star)$. This proves Assumption 6. The same display also shows that every entry of Σ_A is strictly positive, so in particular $\min_{k,\ell} \Sigma_A(k, \ell) \geq c_2'$ for some $c_2' > 0$; we record this only as a by-product. Assumption 7 already holds from the balanced-client construction. □

Step 3: Varshamov–Gilbert packing on non-anchor rows. Let $p_1 := \lfloor (1 - b_2)p \rfloor$ and $m_{\mathrm{VG}} := \lfloor p_1/2 \rfloor$. Choose $p_0(K)$ large enough that, for every $p \geq p_0(K)$, all $p_{\mathrm{anc},k} \geq 1$, $m_{\mathrm{VG}} \geq 8$, and the anchor counts and p_1 are bounded above and below by fixed positive multiples of p. By the Varshamov–Gilbert bound, there exist $\omega^{(0)}, \omega^{(1)}, \ldots, \omega^{(J)} \in \{0, 1\}^{m_{\mathrm{VG}}}$ such that $\omega^{(0)} = 0$,

$$\sum_{j=1}^{m_{\mathrm{VG}}} \mathbf{1}\{\omega_j^{(s)} \neq \omega_j^{(t)}\} \geq \frac{m_{\mathrm{VG}}}{8} \quad (s \neq t), \qquad J \geq 2^{m_{\mathrm{VG}}/8}.$$

Set

$$\alpha := \frac{C_1}{K p_1} \sqrt{\frac{p}{N n_{\mathrm{tot}}}}, \tag{48}$$

where $C_1 > 0$ is a sufficiently small constant. For each s, let $u^{(s)} \in \mathbb{R}^p$ be supported on any fixed p_1 non-anchor rows and equal to $(\omega^{(s)}, -\omega^{(s)})$ on those rows, with one extra zero if p_1 is odd; set all remaining entries to zero. Define

$$A_k^{(s)} = A_k^{(0)} + \alpha u^{(s)}, \qquad k = 1, \ldots, K.$$

Because $\sum_j u^{(s)}(j) = 0$, all topic columns still sum to one. Since the baseline non-anchor mass is $\asymp p^{-1}$ and $p \log(nM) = o(NnM)$, taking C_1 small ensures $A^{(s)} \geq 0$ and preserves the regularity conditions.

Lemma 32 (Packing perturbations remain in the matched class). For $C_1 > 0$ small enough, every matrix $A^{(s)}$ constructed above is column-stochastic, nonnegative, has the same anchor rows as $A^{(0)}$, and satisfies Assumptions 5 and 6 with constants bounded away from zero. Together with the balanced-client weights (47), each $(A^{(s)}, W^{(1:M)})$ belongs to $\Phi_{n,N,p}^{(M)}(K, c_1, c_2, c_{\mathrm{loc}})$ for all sufficiently large problem sizes under the stated scaling.

Proof. Column sums are unchanged because $\sum_j u^{(s)}(j) = 0$. Anchor rows are not perturbed, so every topic still has at least one anchor row. On non-anchor rows the baseline entries are $\asymp p^{-1}$, whereas

$$p\alpha \lesssim \sqrt{\frac{p}{N n_{\mathrm{tot}}}} = o(1)$$

under $p\log(nM) = o(NnM)$. Choosing C_1 small therefore gives nonnegativity and keeps every non-anchor row sum within a constant factor of K/p; anchor row sums were already $\asymp K/p$. Hence $h_{\min} \gtrsim K/p$ uniformly over the packing.

It remains to check Σ_A. The perturbation satisfies, uniformly in s,

$$\|H^{(0)-1/2}(A^{(s)} - A^{(0)})\| \le C\sqrt{p}\,\alpha\,\sqrt{p_1} \lesssim C_1\sqrt{\frac{p}{Nn_{\text{tot}}}} = o(1),$$

Moreover, uniformly over the packing,

$$\max_j \frac{|h_j^{(s)} - h_j^{(0)}|}{h_j^{(0)}} \le Cp\alpha = o(1),$$

because anchors are unchanged and each non-anchor row sum changes by at most $K\alpha$. Therefore

$$\left\| A^{(s)\top} H^{(s)-1} A^{(s)} - A^{(0)\top} H^{(0)-1} A^{(0)} \right\| = o(1).$$

Lemma 31 gives $\lambda_{\min}(\Sigma_A^{(0)}) \ge c > 0$, so Weyl's inequality gives $\lambda_{\min}(\Sigma_A^{(s)}) \ge c/2$ for all sufficiently large problem sizes under the stated scaling. The balanced construction (47) gives $\lambda_{\min}(\Sigma_W), \lambda_{\min}(\Sigma_{W,m}) \ge 1/(2K)$ for every s and every client m. This proves the lemma. □

Lemma 33 (Separation in $L(\cdot,\cdot)$). For $s \ne t$,

$$L(A^{(s)}, A^{(t)}) \ge c\sqrt{\frac{p}{Nn_{\text{tot}}}}.$$

Proof. Let $s_k := 1 - (1 - b_1 b_2)\eta_k$, $s_{\min} := \min_k s_k > 0$, and $I^{\text{anc}} := \bigcup_{k=1}^K I_k^{\text{anc}}$. For any nonidentity permutation κ, the anchor rows alone contribute

$$\sum_{k=1}^K \sum_{j \in I^{\text{anc}}} |A^{(s)}_{\kappa(k)}(j) - A^{(t)}_k(j)| = 2 \sum_{k:\kappa(k)\ne k} s_k \ge 4 s_{\min}.$$

The identity matching has loss at most $2K\alpha m_{\text{VG}} = o(1)$. Thus it minimizes the loss for all sufficiently large problem sizes. The packing separation now gives

$$L(A^{(s)}, A^{(t)}) = 2K\alpha\|\omega^{(s)} - \omega^{(t)}\|_1 \ge 2K\alpha\frac{m_{\text{VG}}}{8} \asymp \sqrt{\frac{p}{Nn_{\text{tot}}}}.$$

□

Lemma 34 (KL control). For C_1 small enough, for all $s \ge 1$,

$$\text{KL}(P_s \| P_0) \le \alpha_0 \log J$$

for some $\alpha_0 \in (0, 1/8)$.

Proof. Let $D_0^{(s)} = A^{(s)}W$. Since every column of $W^{(m)}$ is pure, each column of $D_0^{(s)}$ is one of the topic columns. The perturbation from $A^{(0)}$ to $A^{(s)}$ is supported entirely on the non-anchor rows, so the matrix difference $D_0^{(s)} - D_0^{(0)}$ vanishes on every anchor row. Consequently the perturbation support $\mathcal{S}$ in Lemma 30 is contained in the non-anchor rows.

On each non-anchor row, the baseline entry of a pure-topic column equals

$$D_0^{(0)}(j,i) = \beta_{\text{non}}\eta_k \asymp p^{-1},$$

uniformly over j and over the topic label k of the column. The perturbation magnitude on $\mathcal{S}$ is at most α, so

$$\sum_{(j,i)\in\mathcal{S}} \left|D_0^{(s)}(j,i) - D_0^{(0)}(j,i)\right|^2 \le n_{\text{tot}} p_1 \alpha^2,$$

and

$$\delta_{\mathcal{S}} = \max_{(j,i)\in\mathcal{S}} \frac{|D_0^{(s)}(j,i) - D_0^{(0)}(j,i)|}{D_0^{(0)}(j,i)} \lesssim p\alpha = o(1)$$

under the theorem's scaling assumption. Applying the support-restricted KL lemma therefore gives

$$\mathrm{KL}(P_s \| P_0) \lesssim N \sum_{(j,i)\in\mathcal{S}} \frac{|D_0^{(s)}(j,i) - D_0^{(0)}(j,i)|^2}{D_0^{(0)}(j,i)} \lesssim Npn_{\mathrm{tot}}p_1\alpha^2 \asymp C_1^2 p.$$

Since $\log J \asymp p$, choosing C_1 sufficiently small proves the claim. □

A.10.4 Conclude the minimax lower bound

Proof of Theorem 7. Restrict the supremum in Theorem 7 to the finite subset $\{(A^{(s)}, W^{(1:M)})\}_{s=0}^{J}$ constructed above, which lies inside $\Phi_{n,N,p}^{(M)}(K, c_1, c_2, c_{\mathrm{loc}})$ by Lemmas 31 and 32. Take loss $\ell((A, W^{(1:M)}), (A', W^{(1:M)})) := L(A, A')$. The triangle inequality for L follows by composing column permutations, which act isometrically on the entrywise ℓ_1 norm.

By Lemma 33, the packing radius is $r \asymp \sqrt{p/(Nn_{\mathrm{tot}})} = \sqrt{p/(NnM)}$. By Lemma 34, the KL condition for Fano holds. Apply Lemma 29 to conclude that

$$\inf_{\hat{A}} \sup_{s\in\{0,\dots,J\}} \mathbb{P}_s\Big(L(\hat{A}, A^{(s)}) \geq c\sqrt{\frac{p}{NnM}}\Big) \;\geq\; \delta_0,$$

for constants $c, \delta_0 > 0$. Since the supremum over the full parameter class is larger than the supremum over this finite subset, the theorem follows. □

A.11 Proofs of the document-weight and pilot-K results

Lemma 35 (Local Lipschitzness of truncation-and-renormalization). Let $\pi \in \mathbb{R}^K$ be a probability vector and let $x \in \mathbb{R}^K$ satisfy $\|x - \pi\|_1 \leq 1/2$. Define

$$\mathcal{T}(x) := \frac{[x]_+}{\|[x]_+\|_1}, \qquad [x]_+ := (\max\{x_1, 0\}, \dots, \max\{x_K, 0\})^\top.$$

Then

$$\|\mathcal{T}(x) - \pi\|_1 \leq 4\|x - \pi\|_1.$$

Proof. Since coordinatewise truncation cannot increase ℓ_1 distance to a nonnegative vector,

$$\|[x]_+ - \pi\|_1 \leq \|x - \pi\|_1.$$

Also

$$\big|\|[x]_+\|_1 - 1\big| = \big|\|[x]_+\|_1 - \|\pi\|_1\big| \leq \|[x]_+ - \pi\|_1 \leq \frac{1}{2},$$

so $\|[x]_+\|_1 \geq 1/2$. Therefore,

$$\left\| \frac{[x]_+}{\|[x]_+\|_1} - [x]_+ \right\|_1 = \frac{|1 - \|[x]_+\|_1|}{\|[x]_+\|_1} \|[x]_+\|_1 \leq 2\|[x]_+ - \pi\|_1.$$

Combining the last two displays with the triangle inequality gives

$$\|\mathcal{T}(x) - \pi\|_1 \leq 3\|[x]_+ - \pi\|_1 \leq 3\|x - \pi\|_1.$$

Replacing 3 by 4 leaves a clean statement. □

Lemma 36 (Conditioning of the weighted least-squares Gram matrix). Let

$$B := A^\top M_d^{-1} A, \qquad \hat{B} := \hat{A}^\top M_d^{-1} \hat{A}, \qquad \hat{h}_j := \sum_{k=1}^{K} \hat{A}(j,k).$$

Assume $\Delta_1(D, D_0) \le c$ for a sufficiently small constant c, and suppose moreover that

$$\max_{1\le j\le p} \frac{\|\hat{a}_j - a_j\|_1}{h_j} \le \varepsilon_0$$

for a sufficiently small $\varepsilon_0 > 0$. Then

$$cI_K \preceq B \preceq CI_K, \qquad \lambda_{\min}(\hat{B}) \ge c/2, \qquad \|\hat{B}^{-1}\| \le 2/c,$$

and

$$\max_{1\le j\le p} \frac{\hat{h}_j}{M_d(j,j)} \le C.$$

The constants depend only on the regularity constants and on K.

Proof. By Lemma 37 and the assumption $\Delta_1(D, D_0) \le c$, one has $M_d(j,j) \asymp M_0(j,j) \asymp h_j$ uniformly in j. Hence

$$B = A^\top M_d^{-1} A \asymp A^\top M_0^{-1} A.$$

Lemma 37 shows that $A^\top M_0^{-1} A$ is well-conditioned, so the same is true for B.

Next,

$$\hat{B} - B = (\hat{A} - A)^\top M_d^{-1}(\hat{A} - A) + (\hat{A} - A)^\top M_d^{-1} A + A^\top M_d^{-1}(\hat{A} - A).$$

Thus

$$\|\hat{B} - B\| \le \|M_d^{-1/2}(\hat{A} - A)\|^2 + 2\|M_d^{-1/2}(\hat{A} - A)\|\,\|M_d^{-1/2} A\|.$$

Because $M_d(j,j) \asymp h_j$ and $\|\hat{a}_j - a_j\|_1 \le \varepsilon_0 h_j$,

$$\|M_d^{-1/2}(\hat{A} - A)\|^2 \le \|M_d^{-1/2}(\hat{A} - A)\|_F^2 \le C \sum_{j=1}^{p} \frac{\|\hat{a}_j - a_j\|_2^2}{h_j} \le C\varepsilon_0^2 \sum_{j=1}^{p} h_j \le C_K \varepsilon_0^2.$$

Thus this operator norm can be made arbitrarily small by choosing ε_0 small. Since $\|M_d^{-1/2} A\| \le C$, the decomposition above yields $\|\hat{B} - B\| \le c/2$ for small enough ε_0, and Weyl's inequality gives $\lambda_{\min}(\hat{B}) \ge c/2$. The inverse bound follows immediately.

Finally,

$$|\hat{h}_j - h_j| \le \|\hat{a}_j - a_j\|_1 \le \varepsilon_0 h_j,$$

so $\hat{h}_j \asymp h_j$. Since $M_d(j,j) \asymp h_j$ as well, the ratio $\hat{h}_j / M_d(j,j)$ is uniformly bounded. □

A.11.1 Proof of the document-weight theorem

Proof. Condition throughout on the training split D^{tr}. Then

$$\hat{A} := \hat{A}^{\mathrm{tr}}_{\mathrm{fed}}, \qquad M := M_d^{\mathrm{tr}}$$

are deterministic for the held-out analysis, while d_i^{te} is independent of them and has mean Aw_i. Write

$$d_i^{\mathrm{te}} = Aw_i + \eta_i, \qquad \mathbb{E}[\eta_i \mid D^{\mathrm{tr}}, w_i] = 0.$$

The unconstrained WLS estimator is

$$\hat{w}_i^\star = (\hat{A}^\top M^{-1} \hat{A})^{-1} \hat{A}^\top M^{-1} d_i^{\mathrm{te}}.$$

Work on the training event

$$\mathcal{E}_{\mathrm{tr}}(\delta_{\mathrm{tr}}) = \left\{ \max_{1\le j\le p} \frac{\|\hat{a}_j - a_j\|_1}{h_j} \le C\varepsilon^{\mathrm{tr}}_{n_{\mathrm{tr}},M}(\delta_{\mathrm{tr}}), \qquad \Delta_1^{\mathrm{tr}} \le c \right\}.$$

On this event, Lemma 36 gives

$$\lambda_{\min}(\hat{A}^\top M^{-1} \hat{A}) \ge c, \qquad \|(\hat{A}^\top M^{-1} \hat{A})^{-1}\| \le C, \qquad \max_j \frac{\hat{h}_j}{M(j,j)} \le C.$$

Using $d_i^{\mathrm{te}} = Aw_i + \eta_i$ and the identity $(\hat{A}^\top M^{-1}\hat{A})^{-1}\hat{A}^\top M^{-1}\hat{A} = I$, we obtain

$$\hat{w}_i^\star - w_i = (\hat{A}^\top M^{-1}\hat{A})^{-1}\hat{A}^\top M^{-1}\{(A-\hat{A})w_i + \eta_i\}.$$

For any $v \in \mathbb{R}^p$,

$$\|\hat{A}^\top M^{-1} v\|_1 \le \sum_j \frac{\hat{h}_j}{M(j,j)}|v(j)| \le C\|v\|_1.$$

Since w_i is a simplex vector and $\mathcal{E}_{\mathrm{tr}}(\delta_{\mathrm{tr}})$ supplies the rowwise topic bound,

$$\|(A-\hat{A})w_i\|_1 \le \max_{1\le j\le p}\frac{\|\hat{a}_j - a_j\|_1}{h_j}\sum_{j=1}^p h_j \le C\varepsilon^{\mathrm{tr}}_{n_{\mathrm{tr}},M}(\delta_{\mathrm{tr}}).$$

It remains to bound the held-out noise term. For each k,

$$T_k := e_k^\top \hat{A}^\top M^{-1}\eta_i = \sum_{j=1}^p \frac{\hat{A}(j,k)}{M(j,j)}\eta_i(j).$$

Conditional on D^{tr} and w_i, the coefficients are deterministic and uniformly bounded: indeed, $0 \le \hat{A}(j,k) \le \hat{h}_j$ after the truncation, simplex renormalization, and column normalization steps, and Lemma 36 gives $\hat{h}_j/M(j,j) \le C$. Writing the held-out document as an average of N independent word draws, T_k is a mean-zero average of bounded variables. Hoeffding's inequality and a union bound over fixed K give

$$\|\hat{A}^\top M^{-1}\eta_i\|_1 \le C\sqrt{\frac{\log(2K/\delta_{\mathrm{te}})}{N}}$$

with probability at least $1-\delta_{\mathrm{te}}$ over the held-out document. Therefore, with

$$R_{\mathrm{W}} := C\left\{\varepsilon^{\mathrm{tr}}_{n_{\mathrm{tr}},M}(\delta_{\mathrm{tr}}) + \sqrt{\frac{\log(2K/\delta_{\mathrm{te}})}{N}}\right\},$$

we have $\|\hat{w}_i^\star - w_i\|_1 \le R_{\mathrm{W}}$. If $R_{\mathrm{W}} \le 1/2$, Lemma 35 gives the same order for the truncated and renormalized estimator after enlarging constants. If $R_{\mathrm{W}} > 1/2$, the trivial simplex bound $\|\hat{w}_i - w_i\|_1 \le 2$ gives the displayed $2 \wedge R_{\mathrm{W}}$ form after another enlargement of the universal constant. For cross-fitting, apply this argument only to predictions from fits that exclude the target document. Intersect the training and held-out events by a union bound and use the topic matching described above. After alignment, convexity of the ℓ_1 norm bounds the error of any average of these predictions by the average of their errors; independence between folds is not required. □

A.11.2 Proof of the pilot-K theorem

Proof. Let

$$r_t := \min\{t, \rho_\star - t\}.$$

On the event $\|G_{\mathrm{pilot}} - G_0\| < r_t$, Weyl's inequality gives

$$|\lambda_k(G_{\mathrm{pilot}}) - \lambda_k(G_0)| < r_t, \qquad 1 \le k \le p.$$

For $1 \le k \le K$,

$$\lambda_k(G_{\mathrm{pilot}}) \ge \lambda_K(G_0) - r_t \ge \rho_\star - r_t \ge t.$$

For $k \ge K+1$,

$$\lambda_k(G_{\mathrm{pilot}}) < \lambda_{K+1}(G_0) + r_t = r_t \le t.$$

Thus the thresholding rule returns exactly K. The assumed probability bound for the operator-norm event gives $\hat{K}_t^{\mathrm{pil}} = K$ with probability $1 - o((nM)^{-3})$. This pilot step is separate from Algorithm 2; after selecting $\hat{K}_t^{\mathrm{pil}}$, the main projector-only protocol is rerun with that value. □

A.12 Probability space and conditioning convention

Because the use of the shared empirical whitening matrix M_d can obscure what is random and what is deterministic, we record the underlying probability space explicitly. For each client m, the matrix $W^{(m)}$ is regarded as fixed. Conditional on $W^{(1:M)}$, the document columns of $D^{(m)}$ are independent and satisfy

$$Nd_i^{(m)} \sim \mathrm{Multinomial}\big(N, Aw_i^{(m)}\big).$$

Therefore the product probability measure is

$$\mathbb{P}_{A,W^{(1:M)}} = \bigotimes_{m=1}^{M} \bigotimes_{i=1}^{n_m} \mathrm{Law}\big(d_i^{(m)} \mid A, w_i^{(m)}\big).$$

All expectations and probabilities in the main text are taken with respect to this conditional law. In particular,

$$D_0^{(m)} = AW^{(m)}, \qquad D_0 = AW, \qquad M_0 = \mathrm{diag}\Big(\frac{1}{n_{\mathrm{tot}}} D_0 \mathbf{1}\Big)$$

are deterministic once $W^{(1:M)}$ is fixed. This is the fixed-design pLSI setup used throughout the analysis.

The only genuinely random object appearing in the whitening step is M_d. Because the same M_d is shared by all clients, the local matrices $\hat{G}_m$ are not independent. This is why the fixed-whitening matrices $\tilde{G}_m$ are essential:

$$\hat{G}_m \xleftarrow[\text{bridge}]{M_d \leftrightarrow M_0} \tilde{G}_m \xrightarrow{\mathbb{E}[\cdot|W^{(1:M)}]} G_{0,m} + D_{0,m}^{\mathrm{svd}}.$$

The fixed-whitening matrices $\tilde{G}_m$ remain independent across clients, so the averaging argument can still be driven by matrix Bernstein inequalities. The dependence induced by the shared M_d is absorbed through $\hat{P}_m - \tilde{P}_m$, while the deterministic ordinary-SVD diagonal bias $D_{0,m}^{\mathrm{svd}}$ is included in the fixed-whitening perturbation radius. This separation is the key probabilistic simplification in the federated analysis.

If one instead postulates a hierarchical prior on $W^{(1:M)}$, the results may be applied conditionally on the realized design whenever the stated assumptions hold. Let $\mathcal{G}_W$ denote the event that all required design assumptions hold with fixed regularity constants. Integrating the conditional guarantee yields an unconditional failure probability at most

$$\delta + \mathbb{P}(\mathcal{G}_W^c).$$

The same $1-\delta$ guarantee therefore follows when the prior is supported on the regular class; otherwise, the probability of $\mathcal{G}_W^c$ must be controlled separately.

A second subtle point concerns the choice of target subspace. Because each client may have a different $W^{(m)}$, the population matrices $G_{0,m}$ are not numerically identical. Nevertheless, as long as each $\Sigma_{W,m}$ is full rank, all $G_{0,m}$ share the same leading eigenspace:

$$\mathrm{col}(G_{0,m}) = \mathrm{col}(M_0^{-1/2} A W^{(m)}) = \mathrm{col}(M_0^{-1/2} A).$$

Hence there is still a common deterministic projector P. This is the correct object to average around. The bridge decomposition therefore solves two issues simultaneously:

1. it isolates the conditional sampling fluctuation from the deterministic ordinary-SVD diagonal bias $D_{0,m}^{\mathrm{svd}}$;
2. it reveals a common deterministic target projector even though the client-wise population Gram matrices have different eigenvalues.

The main text uses this convention uniformly. Formulas involving conditional expectations are statements about the conditional law of the document noise with the entire fixed design, and hence the pooled matrix M_0, held fixed. No contradiction arises from the fact that one may also place a higher-level probabilistic model on $W^{(1:M)}$.

A.13 Oracle geometry under common whitening

We record the oracle geometry under the common whitening M_0. The first Topic-SCORE direction is fixed canonically by the pooled marginal, not by an arbitrary eigenbasis.

Lemma 37 (The pooled marginal matrix is comparable to H). Recall $M_0 = \mathrm{diag}(m_0)$ with $m_0 = n_{\mathrm{tot}}^{-1} D_0 \mathbf{1}$ and $D_0 = AW$. Under Assumption 6,

$$ch_j \le M_0(j,j) \le h_j, \qquad 1 \le j \le p.$$

Consequently,

$$cI_K \preceq A^\top M_0^{-1} A \preceq CI_K.$$

Proof. Write $w = n_{\mathrm{tot}}^{-1} W\mathbf{1}$. Then $M_0(j,j) = a_j^\top w$. Since every coordinate of w is at most one, $M_0(j,j) \le \|a_j\|_1 = h_j$. Also,

$$w(k) = \frac{1}{n_{\mathrm{tot}}} \sum_{i=1}^{n_{\mathrm{tot}}} w_i(k) \ge \frac{1}{n_{\mathrm{tot}}} \sum_{i=1}^{n_{\mathrm{tot}}} w_i(k)^2 = \Sigma_W(k,k) \ge \lambda_{\min}(\Sigma_W) \ge c.$$

Therefore $M_0(j,j) \ge ch_j$. The matrix inequalities follow by comparing M_0 with H entrywise and using $\Sigma_A = A^\top H^{-1} A$. □

Set

$$\xi_1 := M_0^{1/2} \mathbf{1}.$$

Because $\mathbf{1}^\top m_0 = 1$, this vector has Euclidean norm one. Choose any orthonormal basis $\Xi_\perp$ of $\mathrm{col}(M_0^{-1/2} A) \cap \xi_1^\perp$ and define

$$\Xi = [\xi_1, \Xi_\perp].$$

Lemma 38 (Canonical basis and change of coordinates). There exists a unique nonsingular matrix $V = [V_1, \ldots, V_K] \in \mathbb{R}^{K \times K}$ such that

$$\Xi = M_0^{-1/2} AV.$$

Its first column is exactly the pooled mean topic-weight vector

$$V_1 = w := \frac{1}{n_{\mathrm{tot}}} W\mathbf{1}.$$

Moreover,

$$VV^\top = (A^\top M_0^{-1} A)^{-1}.$$

In particular, every coordinate of V_1 is bounded away from zero and infinity.

Proof. Since $M_0^{-1/2} A$ has full column rank and the same column space as Ξ, the matrix V exists uniquely. Also

$$M_0^{-1/2} Aw = M_0^{-1/2} m_0 = M_0^{1/2} \mathbf{1} = \xi_1,$$

so uniqueness forces $V_1 = w$. Multiplying $\Xi = M_0^{-1/2} AV$ on the left by $\Xi^\top$ yields

$$I_K = \Xi^\top \Xi = V^\top A^\top M_0^{-1} AV,$$

equivalently $VV^\top = (A^\top M_0^{-1} A)^{-1}$. By Lemma 37, $A^\top M_0^{-1} A$ is uniformly well conditioned. Also each coordinate of w is at least $\lambda_{\min}(\Sigma_W)$, proving strict positivity and boundedness. □

For each topic k, define

$$v_k^\star(\ell) = \frac{V_{k,\ell+1}}{V_{k1}}, \qquad 1 \le \ell \le K-1,$$

and for each word j, define

$$\pi_j(k) = \frac{A(j,k) V_{k1}}{\sum_{\ell=1}^K A(j,\ell) V_{\ell 1}}.$$

Lemma 39 (Oracle simplex geometry). Writing the oracle ratio matrix as

$$R(j,\ell) = \frac{\Xi_\perp(j,\ell)}{\xi_1(j)},$$

one has

$$r_j = \sum_{k=1}^{K} \pi_j(k) v_k^\star, \qquad A\,\mathrm{diag}(V_1) = M_0^{1/2}\,\mathrm{diag}(\xi_1)\Pi, \qquad \Pi = [\pi_1, \ldots, \pi_p]^\top.$$

If word j is an anchor word for topic k, then $r_j = v_k^\star$. Hence the oracle ratio cloud lies in a nondegenerate simplex with vertices $v_1^\star, \ldots, v_K^\star$.

Proof. Because $\Xi = M_0^{-1/2} AV$,

$$[1, R] = \mathrm{diag}(\xi_1)^{-1} M_0^{-1/2} AV.$$

Write $V = \mathrm{diag}(V_1)[1, V^\star]$, where $V^\star(k,\ell) = V_{k,\ell+1}/V_{k1}$. Then

$$[1, R] = \Pi[1, V^\star], \qquad \Pi := \mathrm{diag}(\xi_1)^{-1} M_0^{-1/2} A\,\mathrm{diag}(V_1).$$

This proves the convex-combination formula and the identity $A\,\mathrm{diag}(V_1) = M_0^{1/2}\,\mathrm{diag}(\xi_1)\Pi$. If row j of A has only one nonzero entry at topic k, then $\pi_j = e_k$ and $r_j = v_k^\star$. Finally,

$$\begin{pmatrix} 1 & \cdots & 1 \\ v_1^\star & \cdots & v_K^\star \end{pmatrix}^\top = \mathrm{diag}(V_1)^{-1} V,$$

so the simplex is nondegenerate because both factors on the right are nonsingular. □

Lemma 40 (Row scales). For every $j \in [p]$,

$$\xi_1(j) = \sqrt{M_0(j,j)} \asymp \sqrt{h_j}, \qquad \|e_j^\top \Xi\|_2 \le C\sqrt{h_j}, \qquad \|r_j\|_2 \le C.$$

Proof. The first statement is immediate from Lemma 37. For the second,

$$e_j^\top \Xi = M_0(j,j)^{-1/2} a_j^\top V,$$

so $\|e_j^\top \Xi\|_2 \le M_0(j,j)^{-1/2}\|a_j\|_1\|V\| \le C\sqrt{h_j}$. The last bound follows because each ratio point lies in the convex hull of the bounded vertices $v_1^\star, \ldots, v_K^\star$. □

Interface to the deterministic propagation. Appendix A.14 uses exactly three oracle inputs: the identity $A\,\mathrm{diag}(w) = M_0^{1/2}\,\mathrm{diag}(\xi_1)\Pi$; the row-scale bounds $\xi_1(j) \asymp \sqrt{h_j}$ and $\|e_j^\top \Xi\|_2 \le C\sqrt{h_j}$; and the bounded geometry of the simplex matrix. Once these facts are in place, the remainder of the Topic-SCORE propagation is deterministic.

A.14 Additional technical details for topic-matrix recovery

This appendix records the deterministic Topic-SCORE propagation from the canonical basis error of Theorem 5 to the topic-matrix error of Theorem 6. Throughout, the empirical spectral basis is the canonical pair $[\hat\xi_1, \hat\Xi_\perp O_\perp]$, not an arbitrary top-$K$ eigenbasis. We use the shorthand

$$\Delta_2 := \max_{1\le j\le p} h_j^{-1/2} \Big\| e_j^\top \big([\hat\xi_1, \hat\Xi_\perp O_\perp] - [\xi_1, \Xi_\perp]\big) \Big\|_2.$$

Lemma 41 (Scale of the canonical positive direction). Under Assumptions 5–6, there exist constants $c_-, c_+ > 0$ such that

$$c_-\sqrt{h_j} \le \xi_1(j) \le c_+\sqrt{h_j}, \qquad 1 \le j \le p.$$

Moreover, if $\Delta_1(D, D_0)$ and Δ_2 are sufficiently small, then $\hat\xi_1(j) \asymp \sqrt{h_j}$ uniformly in j.

Proof. The deterministic bounds are Lemma 40. For the empirical vector, use $|\hat\xi_1(j) - \xi_1(j)| \le \Delta_2\sqrt{h_j}$ and the smallness of Δ_2. □

Lemma 42 (Ratio perturbation). Under the same assumptions,

$$\max_{1\le j\le p} \|O_\perp^\top \hat{r}_j - r_j\|_2 \le C\Delta_2.$$

Proof. By definition,

$$O_\perp^\top \hat{r}_j - r_j = \frac{e_j^\top(\hat{\Xi}_\perp O_\perp - \Xi_\perp)}{\hat{\xi}_1(j)} + e_j^\top \Xi_\perp \Big(\frac{1}{\hat{\xi}_1(j)} - \frac{1}{\xi_1(j)}\Big).$$

Use $\hat{\xi}_1(j) \asymp \xi_1(j) \asymp \sqrt{h_j}$, $\|e_j^\top \Xi_\perp\|_2 \le C\sqrt{h_j}$, and the definition of Δ_2. □

Lemma 43 (Barycentric perturbation). Let

$$Q = \begin{pmatrix} 1 & \cdots & 1 \\ v_1^\star & \cdots & v_K^\star \end{pmatrix}, \qquad \hat{Q} = \begin{pmatrix} 1 & \cdots & 1 \\ O_\perp^\top \hat{v}_1 & \cdots & O_\perp^\top \hat{v}_K \end{pmatrix}.$$

Then for all j,

$$\|\hat{\pi}_j - \pi_j\|_1 \le C\{\Delta_2 + \mathrm{Err}_{\mathrm{VH}}(O_\perp)\}.$$

Proof. The unconstrained barycentric coordinates satisfy

$$\hat{\pi}_j^\star = \hat{Q}^{-1}\begin{pmatrix} 1 \\ O_\perp^\top \hat{r}_j \end{pmatrix}, \qquad \pi_j = Q^{-1}\begin{pmatrix} 1 \\ r_j \end{pmatrix}.$$

The simplex matrix Q is uniformly well conditioned by Lemma 39, and $\|\hat{Q} - Q\| \le C\,\mathrm{Err}_{\mathrm{VH}}(O_\perp)$. Therefore inverse perturbation plus Lemma 42 gives $\|\hat{\pi}_j^\star - \pi_j\|_2 \le C(\Delta_2 + \mathrm{Err}_{\mathrm{VH}})$. Truncation and renormalization are locally Lipschitz near the simplex, yielding the same order in ℓ_1. □

Lemma 44 (Pre-normalized row perturbation). Let

$$\hat{A}^* := M_d^{1/2}\,\mathrm{diag}(\hat{\xi}_1)\hat{\Pi}, \qquad A^* := A\,\mathrm{diag}(w) = M_0^{1/2}\,\mathrm{diag}(\xi_1)\Pi.$$

Then

$$\|\hat{a}_j^* - a_j^*\|_1 \le Ch_j\{\Delta_1(D, D_0) + \Delta_2 + \mathrm{Err}_{\mathrm{VH}}(O_\perp)\}.$$

Proof. Write

$$\hat{a}_j^* - a_j^* = \sqrt{M_d(j,j)}\hat{\xi}_1(j)(\hat{\pi}_j - \pi_j) + \sqrt{M_d(j,j)}(\hat{\xi}_1(j) - \xi_1(j))\pi_j + (\sqrt{M_d(j,j)} - \sqrt{M_0(j,j)})\xi_1(j)\pi_j.$$

Use $M_0(j,j) \asymp M_d(j,j) \asymp h_j$, $\xi_1(j) \asymp \hat{\xi}_1(j) \asymp \sqrt{h_j}$, and Lemma 43. □

Lemma 45 (Column normalization perturbation). Let $\hat{A}_{\mathrm{fed}}$ be obtained by column-normalizing $\hat{A}^*$, and let A be obtained by column-normalizing $A^* = A\,\mathrm{diag}(w)$. Then

$$\|\hat{a}_j - a_j\|_1 \le Ch_j\{\Delta_1(D, D_0) + \Delta_2 + \mathrm{Err}_{\mathrm{VH}}(O_\perp)\}.$$

Proof. Let $\hat{s}_k = \|\hat{A}_k^*\|_1$ and $s_k = \|A_k^*\|_1 = w_k$. Summing Lemma 44 over rows gives $|\hat{s}_k - s_k| \le C(\Delta_1 + \Delta_2 + \mathrm{Err}_{\mathrm{VH}})$. By Lemma 38, w_k is bounded away from zero and infinity, and so is $\hat{s}_k$ for small perturbations. Then

$$\hat{a}_j(k) - a_j(k) = \frac{\hat{a}_j^*(k) - a_j^*(k)}{\hat{s}_k} + \left(\frac{1}{\hat{s}_k} - \frac{1}{s_k}\right)a_j^*(k),$$

and summing over k proves the claim. □

Together, the preceding lemmas prove Lemma 13: first control ratios, then barycentric coordinates, then the pre-normalized rows, and finally the column-normalized topic matrix.

A.15 Additional bridge calculations

This appendix records the bridge estimates used to compare empirical whitening by M_d with fixed whitening by M_0. The correct perturbation object is the relative diagonal matrix

$$R_M := M_d^{-1/2} M_0^{1/2} - I_p,$$

not the raw difference $M_d^{-1/2} - M_0^{-1/2}$. The raw difference can have operator norm of order $\Delta_1(D, D_0) h_{\min}^{-1/2}$, whereas the relative perturbation is uniformly $O(\Delta_1(D, D_0))$.

Lemma 46 (Relative inverse-square-root perturbation). If $\Delta_1(D, D_0) \le c$ for a sufficiently small constant $c > 0$, then

$$\|R_M\| \le C\Delta_1(D, D_0), \qquad R_M := M_d^{-1/2} M_0^{1/2} - I_p.$$

Moreover,

$$\|u_d - \xi_1\|_2 \le C\Delta_1(D, D_0), \qquad |u_d(j) - \xi_1(j)| \le C\Delta_1(D, D_0)\sqrt{h_j}.$$

Proof. Because all matrices are diagonal, write $m_j = M_0(j, j)$ and $\hat{m}_j = M_d(j, j)$. The event $\Delta_1(D, D_0) \le c$ gives

$$|\hat{m}_j - m_j| \le \Delta_1(D, D_0) h_j, \qquad m_j \asymp h_j,$$

hence $|\hat{m}_j / m_j - 1| \le C\Delta_1(D, D_0)$. Applying the mean-value theorem to $x \mapsto x^{-1/2}$ on a neighborhood of one gives

$$\left| \left(\frac{\hat{m}_j}{m_j} \right)^{-1/2} - 1 \right| \le C\Delta_1(D, D_0),$$

which is exactly the operator bound for R_M. The bounds for $u_d - \xi_1$ follow from the mean-value theorem for $x^{1/2}$ and $\sum_j h_j = K$. □

Lemma 47 (Bridge bound for $\hat{G}_m - \tilde{G}_m$). Suppose $\Delta_1(D, D_0) \le c$. Let

$$T_m := M_0^{-1/2} \hat{S}_m M_0^{-1/2} = \tilde{G}_m + \frac{1}{N} I_p.$$

Then

$$\hat{G}_m - \tilde{G}_m = R_M T_m + T_m R_M + R_M T_m R_M.$$

Consequently,

$$\|\hat{G}_m - \tilde{G}_m\| \le C\Delta_1(D, D_0) \|T_m\|.$$

Furthermore, whenever $\|T_m\| \le C$,

$$\|(\hat{G}_m - \tilde{G}_m) e_j\|_2 \le C\sqrt{h_j}\Delta_1(D, D_0) + C\Delta_1(D, D_0) \|(\tilde{G}_m - G_{0,m}) e_j\|_2 + C\frac{\Delta_1(D, D_0)}{N}.$$

Proof. Since $M_d^{-1/2} = (I + R_M) M_0^{-1/2}$,

$$\hat{G}_m + \frac{1}{N} I_p = (I + R_M) T_m (I + R_M), \qquad \tilde{G}_m + \frac{1}{N} I_p = T_m.$$

Subtracting the two displays gives the stated bridge identity. The operator bound follows immediately from Lemma 46:

$$\|\hat{G}_m - \tilde{G}_m\| \le (2\|R_M\| + \|R_M\|^2) \|T_m\| \le C\Delta_1(D, D_0) \|T_m\|.$$

For the column bound, expand

$$(\hat{G}_m - \tilde{G}_m) e_j = R_M T_m e_j + T_m R_M e_j + R_M T_m R_M e_j.$$

The first and third terms are bounded by $C\Delta_1(D, D_0) \|T_m e_j\|_2$ and a smaller quadratic analogue. Since

$$T_m e_j = (\tilde{G}_m - G_{0,m}) e_j + G_{0,m} e_j + \frac{1}{N} e_j$$

and the oracle row-scale bound gives $\|G_{0,m} e_j\|_2 \le C\sqrt{h_j}$, these terms contribute the stochastic column term, the deterministic $\sqrt{h_j}\Delta_1$ term, and the identity-shift residue $C\Delta_1/N$. For the middle term, $R_M e_j = r_j^{(M)} e_j$ with $|r_j^{(M)}| \le C\Delta_1(D, D_0)$, so it is controlled by the same bound. This proves the lemma. □

A.15.1 Why shared empirical whitening does not destroy averaging

The averaged empirical projector admits the decomposition

$$\bar{P} - P = \frac{1}{M}\sum_{m=1}^{M}(\tilde{P}_m - P) + \frac{1}{M}\sum_{m=1}^{M}(\hat{P}_m - \tilde{P}_m).$$

The first average consists of independent fixed-whitening ordinary-SVD projectors and is handled by matrix and vector Bernstein inequalities; its radius already includes the deterministic diagonal-bias term $D_{0,m}^{\mathrm{svd}}$. The second average contains the dependence induced by the shared M_d, but each summand is perturbative because it carries the additional factor $\Delta_1(D, D_0)$.

A.16 Additional technical details for row-wise canonical recovery

The proof of Theorem 5 proceeds through projectors rather than through an arbitrary top-K eigenbasis. The synchronized chain is

$$\tilde{G}_m - G_{0,m} \to \tilde{P}_m - P \to \bar{P}^{(0)} - P \to \bar{P} - P \to \widehat{P} - P \to (\hat{\xi}_1, \hat{\Xi}_\perp) \to \hat{R}.$$

The key point is that the row-wise basis bound is not a statement about an arbitrary fully rotated top-K basis. The first column is fixed by the projected pooled marginal,

$$\hat{\xi}_1 = \frac{\widehat{P} M_d^{1/2}\mathbf{1}}{\|\widehat{P} M_d^{1/2}\mathbf{1}\|_2},$$

and only the residual basis is aligned by $O_\perp \in \mathbb{R}^{(K-1)\times(K-1)}$.

Lemma 48 (Combined averaged-projector row control). Under the assumptions of Theorem 5, with probability at least $1-\delta$,

$$\|\bar{P} - P\| + \max_j h_j^{-1/2}\|(\bar{P} - P)e_j\|_2 \le C\varepsilon_{n,M}(\delta).$$

Proof. Decompose $\bar{P} - P = (\bar{P}^{(0)} - P) + (\bar{P} - \bar{P}^{(0)})$. The fixed-whitening term is controlled by Theorem 4, Lemma 6, and Lemma 8; this term already contains the deterministic ordinary-SVD diagonal bias. The empirical-whitening projector bridge term is controlled by Lemma 8, the clientwise column event from Lemma 25, and the pooled marginal concentration bound. This is exactly the decomposition carried out in the proof of Theorem 5. □

Lemma 49 (From averaged-projector rows to the canonical pair). If

$$\|\bar{P} - P\| + \max_j h_j^{-1/2}\|(\bar{P} - P)e_j\|_2 \le \eta$$

with $\eta + \Delta_1(D, D_0) \le c$ for a sufficiently small constant $c > 0$, then there exists an orthogonal matrix $O_\perp \in \mathbb{R}^{(K-1)\times(K-1)}$ such that

$$\max_j h_j^{-1/2}\left\|e_j^\top\left([\hat{\xi}_1, \hat{\Xi}_\perp O_\perp] - [\xi_1, \Xi_\perp]\right)\right\|_2 \le C\{\eta + \Delta_1(D, D_0)\}.$$

Proof. Apply Lemma 9 to pass from $\bar{P} - P$ to $\widehat{P} - P$. Then apply Lemma 10 to recover the first direction and Lemma 11 to recover the residual basis. This is exactly the canonical-basis step in Theorem 5. □

A.17 Overspecified local SVD rank

This appendix gives a deterministic sufficient condition under which increasing the local SVD upload rank reduces the subspace-error upper bound in the multi-site setting. This discussion is optional and does not change Algorithm 2 or the main theory, which use rank K spectral uploads. The real-data implementation uses a modest overspecified working rank before the server extracts the final K-dimensional spectral subspace, motivated by the calculation below.

For each client, fix one ordered local SVD with a deterministic measurable rule for ties, and let $Q_m^{(L)}$ project onto its first L left singular vectors, where $K \le L \le p$. Thus, for a fixed client, the SVD subspaces are nested,

$$Q_m^{(K)} \preceq Q_m^{(K+1)} \preceq \cdots .$$

Let P be the rank-K oracle projector onto the common topic space. For any fixed client and any $L \ge K$, define the signal leakage

$$\lambda_m(L) := \|(I - Q_m^{(L)})P\|.$$

Because the subspaces are nested,

$$\lambda_m(L+1) \le \lambda_m(L), \qquad L \ge K. \tag{49}$$

This monotonicity is deterministic: projecting onto a larger local SVD subspace cannot increase the residual norm of any vector in the true signal space. The caveat is that a larger local rank can also include more off-signal directions. The next proposition gives a sufficient condition under which the improved signal retention dominates the extra contamination.

Proposition 4 (A sufficient condition for improvement with an overspecified local rank)**.** *For $L \ge K$, set*

$$\bar{Q}^{(L)} := \sum_{m=1}^{M} \alpha_m Q_m^{(L)}, \qquad \widehat{P}^{(L)} := \textit{the top-}K\textit{ projector of } \bar{Q}^{(L)}.$$

Define

$$a_L := \|(I-P)\bar{Q}^{(L)}P\| + \|P - P\bar{Q}^{(L)}P\|, \qquad b_L := \|(I-P)\bar{Q}^{(L)}(I-P)\|.$$

If $1 - b_L \ge c_* > 0$ and $a_L \le c_*/4$, then

$$\|\widehat{P}^{(L)} - P\| \le C_{c_*} a_L. \tag{50}$$

Consequently, if $L_2 > L_1 \ge K$, $a_{L_2} < a_{L_1} \le c_*/4$, and the off-signal bound $1 - b_L \ge c_*$ holds for both ranks, then the upper bound in (50) *decreases when the local upload rank is increased from L_1 to L_2.*

Proof. Write $\bar{Q}^{(L)}$ in the block decomposition induced by $P \oplus (I-P)$:

$$\bar{Q}^{(L)} = \begin{pmatrix} P\bar{Q}^{(L)}P & P\bar{Q}^{(L)}(I-P) \\ (I-P)\bar{Q}^{(L)}P & (I-P)\bar{Q}^{(L)}(I-P) \end{pmatrix}.$$

The definition of a_L implies

$$P\bar{Q}^{(L)}P \succeq (1-a_L)P, \qquad \|(I-P)\bar{Q}^{(L)}P\| \le a_L.$$

The assumption on b_L gives that the operator restricted to $P^\perp$ has norm at most $1 - c_*$. Hence the K signal eigenvalues of $\bar{Q}^{(L)}$ are separated from the off-signal spectrum by a gap at least $c_*/2$ after an $O(a_L)$ perturbation. Davis–Kahan applied to this block matrix yields (50). The final statement follows immediately. □

This argument concerns the multi-site regime. For $M = 1$ and $L > K$,

$$\dim\{\mathrm{range}(Q^{(L)}) \cap \mathrm{range}(P)^\perp\} \ge L - K > 0,$$

so $b_L = 1$ and the proposition's off-signal condition cannot hold. Moreover, the top-K subspace of a rank-L projector is not unique. A single-site overspecified-rank procedure therefore requires additional spectral information or a different extraction rule. In the multi-site case, a topic-loading guarantee additionally requires the row-normalized control and smallness conditions used by the deterministic Topic-SCORE propagation lemma.

A.18 Overview of the proof structure

This appendix summarizes the dependence structure of the main arguments. The proof has three layers.

Layer 1: client-level bridge decomposition. For each client,

$$\hat{G}_m - G_{0,m} = \{\tilde{G}_m - \mathbb{E}(\tilde{G}_m \mid W^{(1:M)})\} + D_{0,m}^{\mathrm{svd}} + (\hat{G}_m - \tilde{G}_m).$$

The first summand is the centered multinomial fluctuation under deterministic whitening M_0. The second is the deterministic diagonal bias from using the practical ordinary-SVD identity shift. The third summand is the empirical-whitening correction created by replacing M_0 with the shared M_d. In the bridge analysis the empirical-whitening correction is controlled through the relative diagonal perturbation

$$R_M = M_d^{-1/2} M_0^{1/2} - I_p,$$

which avoids the spurious $h_{\min}^{-1/2}$ factor that would appear in the raw difference $M_d^{-1/2} - M_0^{-1/2}$.

Layer 2: projector averaging. The server averages the practical SVD projectors $\hat{P}_m$. To recover independence, we insert only the fixed-whitening ordinary-SVD projectors $\tilde{P}_m$:

$$\bar{P} - P = (\bar{P} - \bar{P}^{(0)}) + (\bar{P}^{(0)} - P), \qquad \bar{P} = \frac{1}{M}\sum_{m=1}^{M} \hat{P}_m, \quad \bar{P}^{(0)} = \frac{1}{M}\sum_{m=1}^{M} \tilde{P}_m.$$

The fixed-whitening average is controlled by the variance-bias projector averaging theorem, with the deterministic diagonal bias $D_{0,m}^{\mathrm{svd}}$ included in its radius. The shared-whitening contribution is controlled separately through $\Delta_1(D, D_0)$ and is retained in the deterministic radius $\varepsilon_{n,M}(\delta)$.

Layer 3: canonical Topic-SCORE propagation. The averaged projector first gives $\widehat{P} - P$. The canonical first direction is then recovered as

$$\hat{\xi}_1 = \frac{\widehat{P} M_d^{1/2} \mathbf{1}}{\|\widehat{P} M_d^{1/2} \mathbf{1}\|_2},$$

and the residual basis is aligned only by $O_\perp \in \mathbb{R}^{(K-1)\times(K-1)}$. The deterministic chain is

$$\widehat{P} - P \to [\hat{\xi}_1, \hat{\Xi}_\perp O_\perp] \to \hat{R} \to \hat{\Pi} \to \hat{A}^* \to \hat{A}_{\mathrm{fed}}.$$

The three errors tracked in the final propagation are

$$\Delta_1(D, D_0), \qquad \Delta_2 = \max_j h_j^{-1/2} \|e_j^\top([\hat{\xi}_1, \hat{\Xi}_\perp O_\perp] - [\xi_1, \Xi_\perp])\|_2, \qquad \mathrm{Err}_{\mathrm{VH}}(O_\perp).$$

They enter as

$$\begin{aligned}
\Delta_2 &\Longrightarrow \max_j \|O_\perp^\top \hat{r}_j - r_j\|_2, \\
\max_j \|O_\perp^\top \hat{r}_j - r_j\|_2 + \mathrm{Err}_{\mathrm{VH}}(O_\perp) &\Longrightarrow \max_j \|\hat{\pi}_j - \pi_j\|_1, \\
\Delta_1 + \Delta_2 + \mathrm{Err}_{\mathrm{VH}}(O_\perp) &\Longrightarrow \max_j h_j^{-1} \|\hat{a}_j^* - a_j^*\|_1, \\
\max_j h_j^{-1} \|\hat{a}_j^* - a_j^*\|_1 &\Longrightarrow \max_j h_j^{-1} \|\hat{a}_j - a_j\|_1.
\end{aligned}$$

Where each theorem is used. Lemma 47 controls the empirical-whitening bridge. Lemma 8 supplies fixed-whitening projector moments and bias. Theorem 4 and Lemma 6 control the averaged projector. Lemmas 9, 10, and 11 convert the averaged projector into the canonical basis. Lemma 13 then converts the canonical-basis error into the topic-matrix error.

B Additional simulation details

B.1 Simulation methods

The simulations follow the shared-topic federated model in Section 2.1. In the base setting, $M = 5$ sites each contribute $n_m = 200$ patients, giving $n_{\mathrm{tot}} = 1000$ patients in total. The base vocabulary size is $p = 1000$, the number of latent components is $K = 5$, and each patient has count length

$N = 2000$. A single topic loading matrix $A \in \mathbb{R}^{p \times K}$ is shared across all sites, while the membership weights $w_i^{(m)}$ are generated according to either the homogeneous or heterogeneous membership design below.

The topic loading matrix is generated with an anchor-concept structure. For each component k, a_p concepts are designated as anchors for that component; in the base setting $a_p = p/100$. An anchor concept for component k has loading δ_p on component k, zero loading on the other components, and $\delta_p = 1/p$ in the base setting. The remaining non-anchor rows are generated from independent positive random entries and rescaled so that each column of A is a probability distribution over concepts. Thus a_p controls the number of rows available to identify each simplex vertex, while δ_p controls the signal strength of each anchor concept. The simulation code uses a fixed base seed and deterministic setting-specific seed offsets so that each grid value and replicate corresponds to an independent generated dataset.

Given A and a patient membership vector $w_i^{(m)}$, we generate raw concept counts by

$$X_i^{(m)} \sim \text{Multinomial}\{N, Aw_i^{(m)}\}, \qquad d_i^{(m)} = X_i^{(m)}/N.$$

The local site matrix is $D^{(m)} = [d_1^{(m)}, \ldots, d_{n_m}^{(m)}]$. Topic-SCORE methods use these frequency matrices, while the LDA benchmark uses the corresponding count matrices $X^{(m)} = ND^{(m)}$. We compare Federated Topic-SCORE with pooled Topic-SCORE, which applies Topic-SCORE to the centralized $[D^{(1)}, \ldots, D^{(M)}]$, single-site Topic-SCORE, which applies Topic-SCORE to one local site, and a federated seeded-LDA benchmark, labeled Federated LDA in the figures. Topic-SCORE methods use the same vertex-hunting routine and tuning choices within a given simulation configuration. Estimation accuracy is measured by the permutation-aligned ℓ_1 distance defined in Section 3. Each simulation setting reports medians and interquartile ranges across 30 independent replicates.

Federated LDA benchmark. The federated LDA benchmark is a practical federated LDA-style comparator motivated by recent work on federated LDA-based topic modeling [12]. It matches the non-repeated-cross-site communication regime more closely than iterative federated variational or neural topic models, while still producing an interpretable topic loading matrix for direct comparison with A.

Another study on federated LDA targets local differential privacy Wang et al. [19], and using it as a benchmark would conflate topic estimation accuracy with the effects of privacy noise. The federated LDA method used in our simulations thus provides a cleaner non-private baseline for evaluating topic estimation accuracy.

The benchmark method is applied on the same multinomial draws as the Topic-SCORE methods. For LDA fitting, the frequency matrix $D^{(m)}$ is converted back to the count matrix $X^{(m)} = ND^{(m)}$, with rows interpreted as concepts and columns as patients. Each site first fits an unsupervised LDA model with known topic number K. The server extracts the local topic-word probability matrices, stacks the local topics from all sites, normalizes them for cosine comparison, and clusters the local topics into K groups. For each group, the server selects the terms with the highest average probabilities within the group to form a seed dictionary containing one list of representative terms per topic. Sites then refit seeded LDA locally using this dictionary to guide topic estimation, and the server aggregates the resulting topic-word matrices with sample-size weights n_m/n_{tot}. The final matrix is column-normalized and evaluated with the same permutation-aligned ℓ_1 loss as the Topic-SCORE estimators.

The implementation fit the federated LDA with 20 seed terms per topic, 500 local unsupervised-LDA iterations, 500 local seeded-LDA iterations, seed weight 0.05, and sample-size weighted aggregation. Thus the Federated LDA protocol avoids patient-level pooling and repeated cross-site model-update communication, but the site-level LDA fits are iterative and computationally heavier than the spectral method used by Federated Topic-SCORE.

Compute resources. The Topic-SCORE-based methods, including pooled, federated, and single-site Topic-SCORE, are lightweight spectral procedures and each simulation replicate ran in seconds on a local CPU machine. The Federated LDA baseline was substantially more computationally intensive because each replicate required local unsupervised LDA fits followed by local seeded-LDA

refits. We therefore ran the full simulation array with the federated LDA baseline on a CPU cluster. The full grid consisted of 2,040 independent simulation jobs, run with up to 20 concurrent jobs on a CPU partition. The full array completed in about 5 hours of wall-clock time, corresponding to approximately 100 aggregate task-hours.

B.2 Homogeneous membership setting

In the homogeneous setting, all sites draw patient membership weights from the same distribution. The base design uses normalized positive random weights and includes a small number b_n of pure-topic patients for each component. The pure-topic patients have $w_i = e_k$ for one component k, so b_n controls how strongly the simulated patient memberships reach the corners of the membership simplex.

Beyond the main-text sensitivity analyses for p, n_{tot}, N, and site-size imbalance, we vary additional parameters that affect the spectral-simplex geometry. These include the number of components K, the number of anchor concepts a_p, the anchor strength δ_p, and the number of pure-topic patients b_n. Each sensitivity analysis varies one parameter at a time while holding the other parameters at their base values, except for the site-size imbalance experiment, which keeps $n_{\text{tot}} = 1000$ fixed and changes the fraction of patients contributed by the site used for the single-site benchmark.

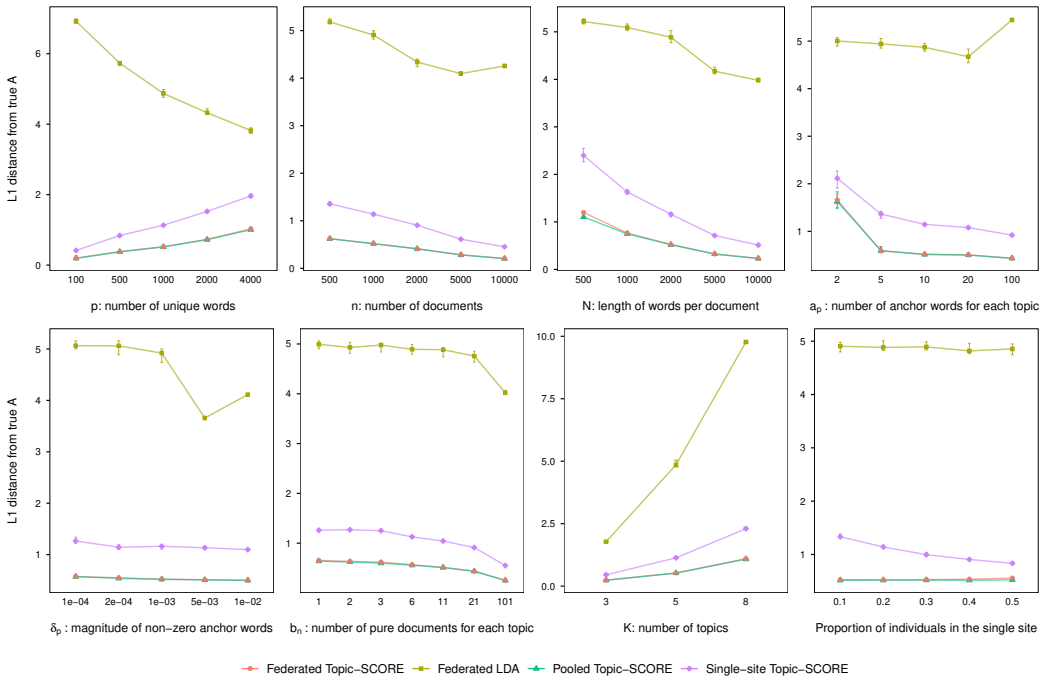


Figure 5: Additional homogeneous-membership simulation sensitivity analyses. Points show median permutation-aligned ℓ_1 error for estimating the true topic loading matrix A. Distinct point shapes and colors identify Federated Topic-SCORE, Federated LDA, pooled Topic-SCORE, and single-site Topic-SCORE. Error bars show interquartile ranges across 30 independent simulation replicates. The Federated Topic-SCORE and pooled Topic-SCORE curves nearly overlap across the displayed settings.

Across the additional homogeneous settings, Federated Topic-SCORE closely matches the pooled estimator and remains more accurate than single-site estimation (Figure 5). Larger vocabulary size and larger K make the estimation harder, whereas larger total sample size and longer patient records reduce error. Increasing the number of anchor concepts or the number of pure-topic patients improves the finite-sample geometry used by vertex hunting. The federated estimator is also stable under moderate site-size imbalance, indicating that the non-iterative summary aggregation is not overly sensitive to one site contributing a larger share of the total sample. Federated LDA has substantially

larger error throughout these homogeneous settings. It improves in easier regimes such as smaller vocabularies, longer records, or stronger anchor structure, but remains less accurate than the Topic-SCORE estimators, consistent with the original Topic-SCORE simulation comparisons in which federated LDA underperformed spectral Topic-SCORE recovery.

B.3 Heterogeneous membership setting

In the heterogeneous simulations, all sites share a common topic loading matrix A. We fix p, K, and N at their base values and use $M = 5$ sites with $n_m = 200$ patients per site. The Dirichlet concentration parameter is fixed at 10 in the base heterogeneous grid. For skewed patients at site m, let $q^{(m)}$ denote the mean membership weight vector, with a specified mean weight assigned to one topic and the remaining weight divided equally among the other $K - 1$ topics. We sample each patient's membership weight vector as $w_i^{(m)} \sim \mathrm{Dirichlet}(10q^{(m)})$. For balanced patients, we sample $w_i^{(m)} \sim \mathrm{Dirichlet}(10/K, \ldots, 10/K)$, giving a uniform mean membership weight vector $(1/K, \ldots, 1/K)^\top$. The scenarios below introduce differences in membership weight distributions across sites by varying the mean weights and the proportions of balanced and skewed patients.

We consider four heterogeneous scenarios. In Scenario 1, all patients at one site is skewed toward one topic, and the weight assigned to that topic is varied. In Scenario 2, one site contains a mixture of balanced and skewed patients; the skewed patients place 0.9 weight on one topic, and the proportion of skewed patients is varied. In Scenario 3, all patients at all sites are skewed, with different sites skewed toward different topic, and the skewed topic weight is varied. In Scenario 4, all sites contain mixtures of balanced and skewed patients, with different sites skewed toward different topic; the skewed component weight is fixed at 0.9, and the skewed-patient proportion is varied.

When only one site has an imbalanced membership weight distribution and the other sites remain balanced, Federated Topic-SCORE achieves estimation accuracy similar to pooled Topic-SCORE, while single-site Topic-SCORE fitted at the imbalanced site becomes less accurate (Figure 6, Scenarios 1 and 2). Federated LDA changes relatively little across these settings, but its estimation error remains higher than that of Federated Topic-SCORE.

In Scenarios 3 and 4, topic imbalance affects all sites. Under severe imbalance, concentration on a single topic reduces the information available at each site about the other topics, making the full topic loading matrix harder to estimate from local data. The estimation error of Federated Topic-SCORE increases as the imbalance becomes more severe, consistent with the local richness condition in Assumption 1, but generally remains lower than that of single-site Topic-SCORE. Federated LDA also has higher estimation error than Federated Topic-SCORE, although the difference becomes smaller under the most severe imbalance.

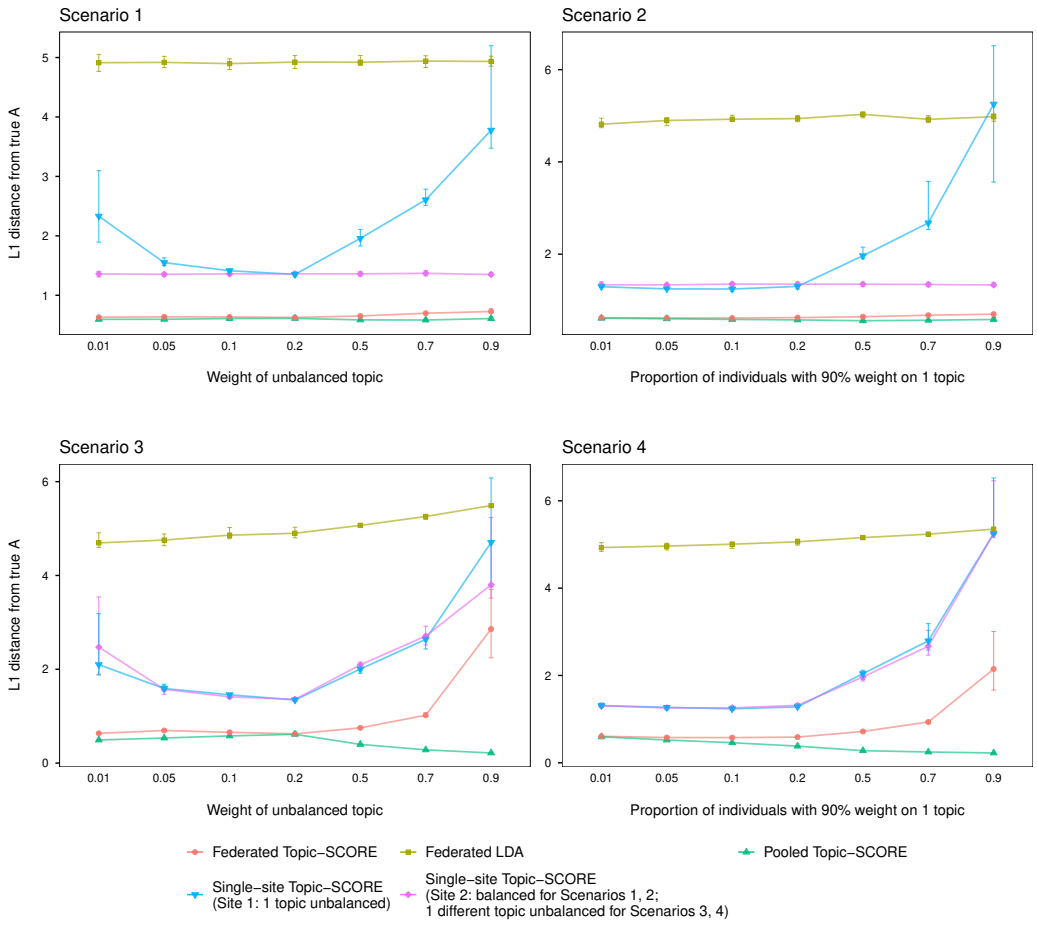


Figure 6: Additional heterogeneous-membership simulation scenarios. Points show median permutation-aligned ℓ_1 error for estimating the true topic loading matrix A. Distinct point shapes and colors identify methods in the shared legend, which distinguishes the unbalanced Site 1 single-site fit from the Site 2, which is balanced in Scenarios 1 and 2 and differently skewed in Scenarios 3 and 4. Error bars show interquartile ranges across 30 independent simulation replicates.

C Real-world EHR application methods

C.1 Cohort construction

The real-world application used EHR data from the Mass General Brigham healthcare system, including Massachusetts General Hospital (MGH), Newton–Wellesley Hospital (NWH), Brigham and Women's Hospital (BWH), and Spaulding Rehabilitation Hospital (SRH). Eligible patients had an inpatient acute ischemic stroke admission during January 1, 2005 through January 31, 2025. The index event was defined as the first qualifying inpatient acute ischemic stroke admission for each patient. Acute ischemic stroke was identified from the principal diagnosis code of the inpatient encounter, using ICD-10 codes beginning with I63 and, for the ICD-9 era, codes of the form 433.x1 or 434.x1. Each patient was assigned to the facility associated with the index admission. This site assignment defines the local EHR matrix used by each hospital in the federated analysis.

C.2 Concept construction

Observed ICD-9 and ICD-10 diagnosis codes were mapped to phecodes. When multiple granularities were available for a code, the pipeline used a stable coarser representation: a one-decimal phecode

when available, otherwise an integer-level phecode, and otherwise a two-decimal phecode folded to one decimal.

We used the diagnosis group represented by the phecodes as the EHR concepts. We then constructed a concept-by-patient matrix by counting the occurrence of each diagnosis group or phecode for each patients. The same construction could be generalized to concepts related to medication, procedure, laboratory values, or other standardized EHR concepts once they are included in the vocabulary.

In this analysis, we only included incident post-stroke concepts in the vocabulary to adjust for patients' baseline conditions. A concept was treated as incident for a patient if it was observed during the 30 days after the index admission and was not observed in that patient's pre-index prevalent window. The prevalent window covered the five years before the index admission and ended seven days before the index admission.

C.3 Concept-by-patient matrix and site contributions

For each patient, we counted retained incident concepts and formed a concept-by-patient matrix. To improve the stability and interpretability of the recovered topics, we retained concepts appearing in at least 20 patients and excluded concepts appearing in more than 80% of patients. We also retained only patients with at least 20 total counts across the retained concepts. Each patient row was then normalized by its total retained count to form concept frequencies. In the notation of the main method, each site stores the transpose of its local concept-by-patient matrix, $D^{(m)} \in \mathbb{R}^{p \times n_m}$.

Patients were assigned to sites where their index stroke admission happened. Patients with ambiguous multi-site assignment at the index admission were excluded from the analysis. Sites contributing more than 100 eligible patients were retained, yielding counts of 490 concepts for 7,057 patients across four hospitals: MGH ($n = 4{,}358$), NWH ($n = 154$), BWH ($n = 2{,}254$), and SRH ($n = 291$).

The table below summarizes demographic characteristics overall and stratified by site. The average pairwise standardized mean differences (SMDs) for age at admission and sex at birth exceeded 0.2, indicating descriptive differences in demographic composition and potential heterogeneity in population across sites. Note that SMDs are calculated for a variable only among individuals without missing value for that variable.

Table 2: Demographic characteristics overall and across hospitals. Hospitals: Massachusetts General Hospital (MGH), Newton–Wellesley Hospital (NWH), Brigham and Women's Hospital (BWH), and Spaulding Rehabilitation Hospital (SRH). SD: standard deviation; SMD: standardized mean difference, averaged across hospital pairs.

Variable	Overall	MGH	NWH	BWH	SRH	SMD
number of patients (n)	7,057	4,358	154	2,254	291	
Age at admission, years, mean (SD)	66.8 (15.6)	66.2 (15.8)	74.6 (14.1)	67.6 (15.0)	65.8 (17.3)	0.306
Missing	4 (0.1%)	4 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Sex at birth, n (%)						0.227
Female	855 (12.1%)	487 (11.2%)	35 (22.7%)	294 (13.0%)	39 (13.4%)	
Male	1,000 (14.2%)	564 (12.9%)	24 (15.6%)	346 (15.4%)	66 (22.7%)	
Missing	5,202 (73.7%)	3,307 (75.9%)	95 (61.7%)	1,614 (71.6%)	186 (63.9%)	
Race, n (%)						0.135
White	5,246 (74.3%)	3,251 (74.6%)	126 (81.8%)	1,675 (74.3%)	194 (66.7%)	
Non-White	1,384 (19.6%)	818 (18.8%)	25 (16.2%)	469 (20.8%)	72 (24.7%)	
Missing	427 (6.1%)	289 (6.6%)	3 (1.9%)	110 (4.9%)	25 (8.6%)	
Ethnic group, n (%)						0.145
Hispanic	190 (2.7%)	123 (2.8%)	0 (0.0%)	62 (2.8%)	5 (1.7%)	
Non-Hispanic	5,632 (79.8%)	3,416 (78.4%)	137 (89.0%)	1,880 (83.4%)	199 (68.4%)	
Missing	1,235 (17.5%)	819 (18.8%)	17 (11.0%)	312 (13.8%)	87 (29.9%)	

C.4 Model implementation and comparisons

We used $K = 5$ latent components for interpretability. Pooled Topic-SCORE was fit to the combined concept-patient matrix as a centralized reference. While the pooling procedure is not the deployable in general setting when sites cannot share individual data with one another, this approach is useful for benchmarking in our demonstrating example. Federated Topic-SCORE was fit to the four site matrices without sharing patient-level records. The federated implementation used global rank

$K = 5$ and local working rank $L = 10$, consistent with the overspecified-rank principle discussed in Section 2.3 and Appendix A.17.

Single-site Topic-SCORE was fit separately to each site. Federated and single-site estimates were compared with the pooled reference using the permutation-aligned ℓ_1 distance between topic loading matrices. Since vertex hunting involves random starts, benchmark comparisons were repeated over 20 replicates.

C.5 Patient subgrouping and survival analysis

After the server estimates the shared component or topic loading matrix $\widehat{A}_{\mathrm{fed}}$, it returns $\widehat{A}_{\mathrm{fed}}$ to the participating sites. Each site then estimates local patient membership weights $\widehat{W}^{(m)}$ from its own $D^{(m)}$, using the weighted least-squares step in (4). For our demonstrating subgroup analysis, each patient was assigned to a a subgroup corresponding to the topic with the largest estimated membership weight. We then fit the Kaplan-Meier survival curve for each subgroup as a post-hoc description Patients were followed from the date of index stroke admission until death, with censoring at last encounter or the end of available follow-up.

C.6 Outcome-agnostic patient profile and subgroup discovery utility of representation learning, topic modeling, and Federated Topic-SCORE applied to EHR analysis

Representation learning applied to EHR analysis can decompose complex EHR data structure and generate compact patient representations or profiles. The learned representations could potentially be used on various complementary and distinct machine learning tasks such as unsupervised patient subgrouping, supervised patient outcome prediction and generative modeling. While supervised learning and generative modeling could be achieved by other dedicated models, the unsupervised data structure discovery and outcome-agnostic patient profile and subgroup learning is a unique utility of representation learning. Specifically, a representation learning approach can summarize a multi-institutional population through compact and shared clinical profiles that potentially reveal intrinsic biological mechanisms of a disease process and quantify how strongly each patient expresses each profile, without optimizing for a particular outcome. Such outcome-agnostic representations can support patient subgroup identification, characterize clinical heterogeneity across subgroups, and be examined against multiple subsequent outcomes simultaneously and guide comprehensive precision care.

Among representation learning approaches, topic modeling remains actively developed and applied in both methodological studies and scientific applications due to its unique advantage in computation and interpretability. Recent work has used or extended topic models to identify EHR comorbidity subtypes with differential genetic risk [9], derive reproducible post-acute COVID-19 subphenotypes across healthcare systems [21], model heterogeneous longitudinal clinical notes [20], and integrate pretrained language-model embeddings into topic estimation [1]. Topic modeling provides direct interpretation at both the representation component and patient levels. Each topic or representation is represented by nonnegative loadings over observed clinical concepts that sum to one, and each patient is represented by nonnegative membership weights over topics that also sum to one. As such, topic loadings naturally characterize the concept composition of each representation or clinical profile, and topic weights naturally characterize the profile structure of a patient. Topic weights can then be used to construct outcome-agnostic patient subgroups that may inform subgroup-specific care strategies given their prognosis for various outcomes, as shown by our real data application.

Federated Topic-SCORE naturally extends the outcome-agnostic patient profile and subgroup discovery utility of representation learning and topic modeling in multi-institutional EHR settings, with its advantage of a non-iterative scheme that significantly mitigates communication and computation costs. To evaluate Federated Topic-SCORE relative to pooled and federated baselines, we applied the same analysis in Section 4.2 to pooled Topic-SCORE and Federated LDA in comparison against Federated Topic-SCORE and compared the degree of separation among the survival curves obtained from the three approaches, each of which learned topic loadings and weights from all sites and is comparable to the others. We evaluated the differentiation in survival patterns across subgroups using the median of pairwise differences (comparing two subgroups at a time) in subgroup-specific restricted mean survival time (RMST) and obtained patient-level bootstrap 95% confidence intervals (with 1,000

bootstrap samples for each estimator) to quantify the uncertainty, given the widely acknowledged clinical interpretability of RMST.

In addition to the mortality outcome that we examined in the main text, we also examined two more outcomes whose differentiated risk may imply targeted stroke care strategies: the post-index stroke recurrent ischemic stroke admission and intracerebral hemorrhage (ICH) admission, to demonstrate the outcome-agnostic feature of the use case. Given the competing risks imposed by mortality on these two outcomes, we examined the pairwise difference in subgroup-specific restricted mean time lost (RMTL) to evaluate the differentiation in time-to-event patterns across subgroups. Notably, the time-to-event outcome framework is applied to all three outcomes due to commonly observed right-censoring due to loss-to-follow-up in EHR data, but our approach can be generalized to categorical or continuous outcomes.

Table 3: Median pairwise RMST/RMTL difference for 365-day mortality, recurrent ischemic stroke readmission, or intracerebral hemorrhage (ICH) readmission. Values are the median absolute pairwise RMST (mortality) or RMTL (recurrent stroke and ICH) differences (in days) across the 10 comparisons among five dominant-topic patient groups, reported in days with patient-level bootstrap 95% confidence intervals in parentheses.

Outcome	Pooled Topic-SCORE	Federated Topic-SCORE	Federated LDA
Mortality (RMST)	41.5 (29.6, 53.2)	42.2 (28.7, 53.1)	13.8 (8.1, 22.0)
Recurrent stroke (RMTL)	9.9 (5.2, 21.2)	9.3 (3.4, 15.2)	9.0 (5.9, 15.2)
ICH (RMTL)	6.2 (4.8, 8.1)	5.8 (3.6, 10.8)	3.8 (2.3, 7.0)

As shown in Table 3, the mortality differentiation of Federated Topic-SCORE closely overlaps that of the gold standard pooled estimate. Both pooled and Federated Topic-SCORE reached a higher level of RMST difference than Federated LDA, with non-overlapping confidence intervals. The difference in RMTL of recurrent stroke and ICH readmission also shows that the Federated Topic-SCORE approaches pooled Topic-SCORE and has a higher level of differentiation than the federated baseline in point estimate, although with overlapping confidence intervals.

D Code and data availability

The R and Python implementations and simulation scripts, together with dependency specifications and instructions for running the simulations, are provided at `https://github.com/zroyan/Federated-Topic-SCORE`. Synthetic simulation inputs are generated by the scripts. Patient-level EHR data used in the real-world application are not publicly available because they contain protected health information and are governed by institutional data-use restrictions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The claims in the abstract and Section 1 match the paper's scope, which covers a non-iterative federated Topic-SCORE method using marginal and low-rank spectral summaries, theory for row-wise and topic-loading recovery, simulations with comparisons against benchmarks, and an real-world multi-facility EHR application.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Section 5 discusses modeling assumptions, local richness requirements, the non-private nature of the summaries, the limited scope of the benchmarks, single-network EHR generalizability, and the hypothesis-generating interpretation of the real-data findings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: Section 3 presents the main assumptions and theorems, while Appendix A provides the full finite-sample upper and lower bounds, perturbative conditions, auxiliary lemmas, and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes].

Justification: Section 4.1 and Appendix B.1 describe the simulation design, base setting, evaluation metric, parameter tuning choices, replicate summaries, and compute resources; R and Python simulation code and instructions are provided in the repository linked in Appendix D. Section 4.2 and Appendix C describe the EHR cohort, concept construction, model settings, and subgroup analysis to the extent possible without releasing protected health information.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No].

   Justification: R and Python simulation code and instructions are provided in the repository linked in Appendix D, and synthetic simulation data can be regenerated from those scripts. Raw EHR data cannot be publicly released because they contain protected health information and are governed by institutional data-use restrictions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Sections 4.1 and 4.2 provide the main experimental settings, and Appendices B.1, B.2, B.3, and C provide additional simulation settings, simulation implementation details, real-world cohort construction, real-world data preprocessing, and subgroup specifications.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

Justification: Figures 1 and 2 and the simulation appendix report medians and interquartile ranges across 30 independent simulation replicates, and Appendix B.1 states the approach to report summary statistics across replicates. For the real-data analysis, we report medians and interquartile ranges across 20 random initializations of vertex hunting for the comparisons with pooled Topic-SCORE. We also report patient-level bootstrap 95% confidence intervals for the RMST and RMTL subgroup-separation summaries, as detailed in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Appendix B.1 reports the computing environment and simulation runtimes. The Topic-SCORE-based methods completed each simulation replicate in seconds on a local CPU machine. The Federated LDA simulations comprised 2,040 jobs run on a CPU cluster, with at most 20 jobs running simultaneously. The full run using the R implementation took about 5 hours.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The EHR application uses governed, non-public patient-derived data under institutional review and data-use restrictions, and no raw patient-level data are released.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

    Justification: Section 5 discusses potential benefits for multi-site patient profiling and hypothesis generation, as well as risks from coding bias and limited generalizability.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes].

    Justification: Sections 4.2 and 5 state that patient-level EHR data are not publicly released, results are reported in aggregate, and learned clinical profiles require clinical review and external validation before use.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes].

    Justification: Existing methods and methodological components are cited in the main text and appendix, including Topic-SCORE, federated LDA modeling, federated neural topic modeling, distributed PCA, and related EHR representation learning work. Software dependencies for the simulation code are documented in the repository linked in Appendix D.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The paper provides code and documentation for our proposed new method through the repository linked in Appendix D. The EHR data are not released because they contain protected health information.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The work does not involve crowdsourcing or participant-facing experiments. The EHR study uses retrospective patient-derived records and is addressed under the institutional review item below.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes].

    Justification: Section 4.2 state that the real-world study analysis was approved by Mass General Brigham Institutional Review Board and that patient-level records are not publicly released.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs were used only for writing assistance such as editing and formatting. They were not used as part of the core methodological conceptualization or design processes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.