TRANSPORT TOPOLOGY FOR FIXED-SUPPORT SINKHORN LAYERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fixed-support Sinkhorn scaling exposes an exact quotient Markov operator. For finite active scores and compatible positive marginals, the homogeneous column-potential derivative of a finite row-column cycle is $M_t = C_t^\top R_t$, formed from the two actual half-step plans. At a balanced fixed point it becomes $M = \operatorname{diag}(b)^{-1} P^\top \operatorname{diag}(a)^{-1} P$. These row-stochastic operators descend to potentials modulo constants, and their transposes transport zero-mass quotient cotangents. We give a self-contained calculus including score and marginal source terms and derive Dobrushin bounds for homogeneous and projected-source tails. The main structural result is a face-lattice dichotomy: a fixed support and compatible marginals admit a score-uniform one-step global quotient-mixing certificate if and only if every nonempty feasible face of the transportation polytope has pairwise two-hop column overlap. Otherwise finite scores can make the coefficient arbitrarily close to one. A windowed characterization extends this criterion to ordered products. Partition heat-bath layers and product-coordinate sweeps provide exact sparse examples in which every one-step coefficient can equal one while the full-window coefficient is zero. Forced bus mass, plan-distortion bounds, and register examples give additional sufficient certificates. The results concern fixed-support quotient transport; they do not establish complete neural-gradient bounds, downstream accuracy, or system speedups.

1 INTRODUCTION

A sparse attention support determines which token interactions are allowed. For Sinkhorn-style transport layers, it also constrains a second object: the Markov operator induced by differentiating a row-column scaling cycle. This operator describes which quotient cotangent modes can mix and which can remain isolated. The relevant design question is therefore not only how many edges exist or whether tokens are reachable, but whether the induced reverse transport contracts on the appropriate quotient.

The distinction matters because connectivity does not by itself control mixing magnitude. In an equal-score biregular baseline, the derivative is an explicit two-step walk on the bipartite support. Under structured scores, however, the scaled plan can concentrate on a boundary face with much worse overlap. Conversely, a sparse schedule can mix over a window even when every individual layer has a one-step coefficient of one. We formalize these possibilities using Dobrushin coefficients, feasible faces, and products of quotient kernels.

Contribution and scope. The central result is an exact characterization of score-uniform one-step mixing for a fixed support–marginal pair, together with its windowed analogue. The supporting calculus distinguishes finite cycles from balanced fixed points and retains score, marginal, and projected source terms. Constructive consequences include capacity-forced bus flow and product-coordinate schedules. The certificates apply to the fixed-support quotient-transport component, not a full parameter-to-loss derivative without additional bounds on surrounding maps. We make no claim about support-changing differentiation, task quality, or production-kernel speed. Detailed statements and proofs accompany the main results in the appendices.

Positioning. Matrix scaling originates in Sinkhorn & Knopp (1967); Idel (2016) surveys existence, support, and convergence. Entropic optimal transport supplies the computational setting (Cuturi,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2013; Peyré & Cuturi, 2019). Implicit differentiation with respect to costs and capacities is treated by Eisenberger et al. (2022), and convergence of differentiated Sinkhorn iterates by Pauwels & Vaiter (2023). Block-Wise Differentiable Sinkhorn Attention (Forde, 2026) studies exact differentiation of a short refinement tail after a stopped base solve. The question here is complementary: which fixed supports and support schedules make the resulting quotient reverse dynamics mix, trap, or require windowed products? The contraction tools are classical (Gaubert & Qu, 2013; Levin & Peres, 2017); our structural statements couple them to the faces of transportation polytopes and to the actual Sinkhorn derivative.

2 Fixed-Support Calculus and the Quotient Kernel

Let $\Omega \subseteq [m] \times [n]$ be the active support and let $a \in \mathbb{R}^m_{>0}$ and $b \in \mathbb{R}^n_{>0}$ have common total mass T. Write

$$\mathcal{P} = \Pi(a, b; \Omega) = \{P \geq 0 : P\mathbf{1} = a,\ P^\top \mathbf{1} = b,\ P_{ij} = 0 \text{ off } \Omega\}.$$

The support is fixed during differentiation. We assume feasibility and componentwise compatibility; wherever strict positivity is used, including all face-lattice results, $\mathcal{P}$ contains a point positive on every active edge. Active scores s_{ij} are finite and dimensionless. For raw logits θ at temperature $\varepsilon > 0$, $s = \theta/\varepsilon$. The log-potential representation is $P_{ij}(u, v; s) = \exp(s_{ij} + u_i + v_j)$ on Ω and zero elsewhere.

The row and column normalizations are

$$u_i^+(v) = \log a_i - \log \sum_{j:(i,j)\in\Omega} e^{s_{ij}+v_j}, \tag{1}$$

$$v_j^+(u) = \log b_j - \log \sum_{i:(i,j)\in\Omega} e^{s_{ij}+u_i}. \tag{2}$$

In a finite cycle, let P_t^r denote the row-normalized intermediate plan and P_t^c the subsequent column-normalized plan. Define

$$R_t = \operatorname{diag}(a)^{-1} P_t^r, \qquad C_t = P_t^c \operatorname{diag}(b)^{-1}, \qquad M_t = C_t^\top R_t. \tag{3}$$

Each row of R_t and column of C_t sums to one. The two plans generally *differ*; substituting either plan for both half-steps gives the wrong finite-cycle derivative.

Proposition 1 (One-cycle derivative and source ledger)**.** *With $d_a = \dot{a} \oslash a$, $d_b = \dot{b} \oslash b$, $r_t(\dot{s})_i = \sum_j (R_t)_{ij}\dot{s}_{ij}$, and $c_t(\dot{s})_j = \sum_i (C_t)_{ij}\dot{s}_{ij}$, one finite cycle satisfies*

$$\dot{v}^+ = M_t \dot{v} + d_b - C_t^\top d_a + C_t^\top r_t(\dot{s}) - c_t(\dot{s}). \tag{4}$$

For an output cotangent η^+, the homogeneous input cotangent and the log-marginal and score covectors are

$$\eta^- = M_t^\top \eta^+, \qquad \lambda_b = \eta^+, \qquad \lambda_a = -C_t \eta^+, \tag{5}$$

$$g^s_{ij} = (R_t)_{ij}(C_t\eta^+)_i - (C_t)_{ij}\eta^+_j, \qquad (i,j) \in \Omega. \tag{6}$$

Raw marginal covectors are $g^a = \lambda_a \oslash a$ and $g^b = \lambda_b \oslash b$; raw-logit covectors are $g^\theta = g^s/\varepsilon$.

Derivation. Differentiating each log-sum-exp in (1)–(2) produces the conditional probabilities of its own half-step. Thus $\dot{u}^+ = d_a - R_t\dot{v} - r_t(\dot{s})$ and $\dot{v}^+ = d_b - C_t^\top \dot{u}^+ - c_t(\dot{s})$. Substitution gives (4); pairing with η^+ and collecting source coefficients gives the adjoints. The complete calculation and marginal-constraint projections are in Propositions A.1 and A.2. □

Quotient and compatibility. For a connected support, the column-potential space is $Q = \mathbb{R}^n/\operatorname{span}\{\mathbf{1}\}$ and $Q^* = \{\eta : \mathbf{1}^\top\eta = 0\}$, with $\|\eta\|_{\mathrm{TV}} = \frac{1}{2}\|\eta\|_1$. Since $M_t\mathbf{1} = \mathbf{1}$, $M_t^\top$ preserves Q^*. For a disconnected support with column components J_Γ, the exact gauge space is $G_\Omega = \operatorname{span}\{\mathbf{1}_{J_\Gamma}\}_\Gamma$ and the exact dual imposes zero mass *on each component*. The global coefficient may equal one because it retains intercomponent modes; it must not be reported as a sharp rate on that smaller component quotient. Schedules with changing component partitions use the common global quotient Q.

Marginal covectors also have constraints. If common component mass can vary, (λ_a, λ_b) is defined modulo $(a_{I_\Gamma}, -b_{J_\Gamma})$. If row and column component masses are fixed independently, there are two independent normals, $(a_{I_\Gamma}, 0)$ and $(0, b_{J_\Gamma})$. The actual marginal parameterization determines the projection; an unconstrained source covector must not be confused with a quotient input cotangent.

Balanced plans and support overlap. At a balanced fixed point, both half-step plans equal P, and

$$M(P) = \operatorname{diag}(b)^{-1} P^\top \operatorname{diag}(a)^{-1} P, \qquad M(P)_{\alpha\beta} = \sum_i \frac{P_{i\alpha} P_{i\beta}}{b_\alpha a_i}. \tag{7}$$

Here $b^\top M = b^\top$, so b is stationary mass. The balanced kernel is reversible with respect to b/T; a finite-cycle kernel need not be reversible or have this stationary mass. The support identity in (7) holds for every feasible plan, including boundary plans: $M_{\alpha\beta} > 0$ exactly when α and β share a row with positive plan mass on both edges (Theorem A.1).

For any row-stochastic A, its Dobrushin coefficient is

$$\tau(A) = \frac{1}{2} \max_{\alpha,\beta} \|A_{\alpha\cdot} - A_{\beta\cdot}\|_1 = 1 - \min_{\alpha,\beta} \sum_\gamma \min\{A_{\alpha\gamma}, A_{\beta\gamma}\}. \tag{8}$$

It bounds $\|A^\top \eta\|_{\mathrm{TV}} / \|\eta\|_{\mathrm{TV}}$ on Q^*. Thus $\tau(M) < 1$ precisely when all row pairs have common positive mass. This is a derivative certificate, not a generic graph statistic.

3 When Topology Forces Score-Uniform Mixing

For a nonempty face F of $\mathcal{P}$, let

$$\Omega_F = \{(i, \alpha) \in \Omega : \text{some } P \in F \text{ has } P_{i\alpha} > 0\}.$$

Call F *pairwise two-hop overlapping* if for every pair of columns α, β there is a column γ and rows i, k with

$$(i, \alpha), (i, \gamma), (k, \beta), (k, \gamma) \in \Omega_F. \tag{9}$$

The intermediate column may depend on the tested pair. The definition concerns row overlap of the induced column kernel, rather than merely connectivity of the original bipartite graph.

Theorem 1 (Face-lattice characterization)**.** *Let $\mathcal{P} = \Pi(a, b; \Omega)$ be nonempty with positive compatible marginals and a point positive on every active edge. A constant $\rho < 1$ satisfying*

$$\tau(M(P_\Omega(S; a, b))) \le \rho \quad \textit{for every finite active score matrix } S$$

exists if and only if every nonempty face of $\mathcal{P}$ satisfies (9)*. If a face fails this condition, then for every $\epsilon > 0$ some finite score matrix on the same support satisfies $\tau(M) > 1 - \epsilon$. The claim uses the global column quotient.*

Proof. Set $\Gamma(P) = 1 - \tau(M(P))$. Positive marginals make $M(P)$ and $\Gamma(P)$ continuous on the compact polytope. A transportation polytope is a standard-form polytope $\{x \ge 0 : Ax = c\}$; each face is obtained by setting edge variables to zero (De Loera & Kim, 2013). A plan in the relative interior of a face is positive exactly on its face support. By (7), $\Gamma(P) > 0$ if and only if the minimal face containing P satisfies (9). If every nonempty face is good, this positivity holds even on the boundary, and compactness gives $\delta_\star = \min_{P \in \mathcal{P}} \Gamma(P) > 0$. All finite-score plans therefore satisfy $\tau(M) \le 1 - \delta_\star$.

Conversely, let F be a bad face. Every face of a finite polytope is exposed, so choose B whose linear maximizers are exactly F. Let P_L maximize $L\langle B, P\rangle + H(P)$ over $\mathcal{P}$, where $H(P) = -\sum_{(i,j)\in\Omega} P_{ij} \log P_{ij}$ and $0 \log 0 = 0$. Write $m_B = \max_{P \in \mathcal{P}} \langle B, P\rangle$ and let P_F^{ent} be the unique entropy maximizer on F. Boundedness of entropy and optimality imply $m_B - \langle B, P_L\rangle = O(1/L)$, so every accumulation point is in F. Comparing with P_F^{ent} gives

$$H(P_L) \ge H(P_F^{\mathrm{ent}}) + L\big(m_B - \langle B, P_L\rangle\big) \ge H(P_F^{\mathrm{ent}}).$$

Continuity and strict concavity of entropy force every accumulation point to be P_F^{ent}; hence $P_L \to P_F^{\mathrm{ent}}$. The entropy maximizer is in $\operatorname{relint}(F)$: if a potentially positive face edge had zero mass,

moving toward a relative-interior point would increase entropy because $-\partial_x(x \log x) \to +\infty$ as $x \downarrow 0$. Its support is therefore Ω_F. Since F is bad, the limiting kernel has two rows with disjoint support. Thus $\Gamma(P_F^{\mathrm{ent}}) = 0$ and continuity yields $\tau(M(P_L)) \to 1$. Each P_L is induced by finite scores LB, establishing the obstruction. The detailed auxiliary arguments are also recorded in Theorems A.4 and A.5. □

Why scores can defeat a nominally good support. The zero-temperature limit maximizes $L\langle B, P\rangle + H(P)$, where $H(P) = -\sum_{(i,j)\in\Omega} P_{ij} \log P_{ij}$ and $0 \log 0 = 0$. Edges outside the optimal face vanish in the limit. This is the entropic penalty central-path mechanism (Cominetti & San Martín, 1994; Weed, 2018), specialized here to the support of the derivative kernel. Large *additive* scores are different: if $B_{ij} = \alpha_i + \beta_j + c_\Gamma$ within each component, then $\langle B, P\rangle$ is constant over the feasible polytope and the plan is unchanged for every scale. The score-sensitive quotient has dimension $|\Omega| - |I| - |J| + c$, the support graph's cycle rank. In particular, a strictly feasible forest has a singleton transport polytope and a score-independent plan (Proposition A.3 and Corollary A.2).

A graph baseline and its quantitative limitation. For a biregular support with degrees d_r, d_c, equal active scores, and uniform marginals, the plan is $1/E$ on its $E = md_r = nd_c$ edges. Hence

$$(M_0)_{\alpha\beta} = \frac{N_{\alpha\beta}}{d_c d_r}, \qquad N_{\alpha\beta} = \#\{i : (i,\alpha), (i,\beta) \in \Omega\}.$$

This is exactly the two-step bipartite walk. If another feasible plan satisfies $\kappa^{-1} \le EP_{i\alpha} \le \kappa$ on every active edge, then

$$M \ge \kappa^{-2} M_0, \qquad \tau(M) \le 1 - \kappa^{-2}(1 - \tau(M_0)). \tag{10}$$

The guarantee requires an observed or independently certified distortion factor, not support alone. Theorem A.3 also transfers elementwise lower bounds to window products. Equal-score spectral and circular local-band consequences are in Appendix D.

A checkable sufficient condition: forced bus mass. Suppose a row h is connected to every column and $\delta_j = a_h + b_j - T > 0$ for all j. Every feasible plan then obeys $P_{hj} \ge \delta_j$, since all non-bus rows together carry only $T - a_h$ mass. Consequently all rows of M share the same positive lower component, giving

$$\tau(M) \le 1 - \gamma_h, \qquad \gamma_h = \left(\min_j \frac{\delta_j}{b_j}\right) \frac{\sum_j \delta_j}{a_h} > 0. \tag{11}$$

This is score-uniform because the floor holds on *every* feasible face. It is sufficient, not necessary. More local capacity deficits can force bus edges without a single heavy universal row (Corollaries A.5 and A.6). By contrast, a register merely being adjacent does not establish a score-uniform mass floor. Equal-score and realized-plan register certificates are separated in Appendix E.

4 Projected-Source Tail Certificates

The homogeneous derivative is only one branch of a neural computation. To specify the remaining cotangents without silently dropping source terms, choose the centered lift $\iota([v]) = \Pi_0 v$, where $\Pi_0 = I - \mathbf{1}\mathbf{1}^\top/n$. Let $\Phi_t(v, z_t)$ produce the source variables of a local cycle from the potential and other independent coordinates, and let ν_t aggregate local source-space cotangents, including the score and marginal covectors in Proposition 1. The cotangent returned to the input quotient is

$$\xi_t = D\big([v] \mapsto \Phi_t(\iota([v]), z_t)\big)^* \nu_t = \Pi_0 D_1 \Phi_t(\iota([v_t]), z_t)^* \nu_t \in Q^*. \tag{12}$$

Here the derivative in the first expression is evaluated at $[v_t]$. Projection is the adjoint of the chosen lift, not an assumed contraction or another Sinkhorn step. A source coordinate independent of v contributes zero to $D_1\Phi_t$ and continues along its separate upstream path. The aggregate ν_t excludes the homogeneous branch, so no branch is counted twice. Definition B.1 gives the full interface, including component quotients and section-dependent maps.

Theorem 2 (Projected-source tail bound)**.** *Let $M_1, \dots, M_K$ be actual row-stochastic cycle kernels on a common column space, and let $\xi_t, \eta_{K+1} \in Q^*$. For the reverse recurrence $\eta_t = M_t^\top \eta_{t+1} + \xi_t$,*

$$\eta_1 = (M_K \cdots M_1)^\top \eta_{K+1} + \sum_{t=1}^{K} (M_{t-1} \cdots M_1)^\top \xi_t, \tag{13}$$

$$\|\eta_1\|_{\mathrm{TV}} \le \tau(M_K \cdots M_1)\|\eta_{K+1}\|_{\mathrm{TV}} + \sum_{t=1}^{K} \tau(M_{t-1} \cdots M_1)\|\xi_t\|_{\mathrm{TV}}. \tag{14}$$

The empty product is the identity. Each product coefficient may be replaced by the product of its one-step coefficients to obtain a looser bound.

Proof. Unroll the recurrence. Every term remains in Q^* because the kernels are row-stochastic and the sources are quotient VJPs. Apply Dobrushin contraction to each product and use the triangle inequality. Submultiplicativity yields the optional product-of-coefficients bound. See Theorems B.1 and B.2 for the complete argument. □

For a stationary quotient equation $\eta = M^\top \eta + \xi$ with $\tau(M) < 1$, the solution on Q^* is the convergent Neumann series $\eta = \sum_{k \ge 0} (M^\top)^k \xi$. Its first K terms satisfy

$$\|\eta - \eta^{(K)}\|_{\mathrm{TV}} \le \frac{\tau(M^K)}{1 - \tau(M)} \|\xi\|_{\mathrm{TV}}, \qquad \|\eta\|_{\mathrm{TV}} \le \frac{\|\xi\|_{\mathrm{TV}}}{1 - \tau(M)}. \tag{15}$$

Corollary B.1 proves existence, uniqueness, and the residual bound. These results do not bound a complete neural gradient unless $\|\xi_t\|_{\mathrm{TV}}$ and the norms of non-transport maps are also controlled.

5 Windowed Topology and Sparse Schedules

A one-step coefficient of one need not indicate failure of a schedule. The relevant quantity can be $\tau(M_K \cdots M_1)$, rather than $\prod_t \tau(M_t)$. Ordered stochastic-matrix products have a classical convergence theory (Wolfowitz, 1963); here their supports are induced by faces of transport polytopes.

Theorem 3 (Windowed face-lattice characterization)**.** *For $t = 1, \dots, K$, let $\mathcal{P}_t = \Pi(a_t, b_t; \Omega_t)$ have positive compatible marginals and strict feasibility on every active edge, on a common column index set. For a nonempty face F_t, let $\mathcal{R}_t(F_t)$ be its two-hop column relation. A constant $\delta_K > 0$ such that*

$$\tau\big(M_K(P_K) \cdots M_1(P_1)\big) \le 1 - \delta_K$$

for every finite score sequence exists if and only if every nonempty face tuple $(F_1, \dots, F_K)$ has pairwise overlapping reachable sets under the relation product $\mathcal{R}_K(F_K); \mathcal{R}_{K-1}(F_{K-1}); \cdots; \mathcal{R}_1(F_1)$. The order follows the matrix product. If a tuple fails the condition, finite score sequences can make its window coefficient arbitrarily close to one.

Proof outline. The support of the product kernel is the Boolean product of the two-hop relations. Pairwise overlap is therefore equivalent to a strictly positive row-overlap coefficient at the minimal face tuple containing the plans. Compactness of $\mathcal{P}_1 \times \cdots \times \mathcal{P}_K$ gives a uniform lower bound when all tuples overlap. Otherwise expose each bad face by a score direction and take its entropy-center limit; the limiting product has two disjoint row supports. The complete argument is Theorem C.1. □

Partition layers identify an exact tractable family. Let $\pi : X \to Z$ partition the column set, take rows indexed by $z \in Z$, and allow (z, x) precisely when $\pi(x) = z$. For a positive probability marginal b, set $a_z = \sum_{x : \pi(x) = z} b_x$. Each column has only one incident row, so the feasible plan is unique:

$$P_{zx} = b_x \mathbf{1}\{\pi(x) = z\}, \qquad M_{xy} = \mathbf{1}\{\pi(x) = \pi(y)\} \frac{b_y}{a_{\pi(x)}}. \tag{16}$$

This is a heat-bath resampling kernel (Levin & Peres, 2017); it is unchanged by any finite active scores (Theorem C.3). For non-product b, a systematic sweep need not produce perfect mixing: strong dependencies can make its coefficient arbitrarily close to one.

Theorem 4 (Product-coordinate window)**.** *Let $X = X_1 \times \cdots \times X_K$ and $b(x) = \prod_{t=1}^{K} \mu_t(x_t)$ be a strictly positive product probability marginal. Layer t has one row for each fiber $z = x_{-t}$, connects it to exactly the columns in that fiber, and has row marginal $a_t(z) = \sum_{u \in X_t} b(z, u)$. For every finite score matrix,*

$$P_t(z, x) = b(x)\mathbf{1}\{x_{-t} = z\}, \qquad (M_t)_{xy} = \mathbf{1}\{x_{-t} = y_{-t}\}\mu_t(y_t).$$

For any permutation σ of the coordinate layers,

$$M_{\sigma(K)} \cdots M_{\sigma(1)} = \mathbf{1}b^\top, \qquad \tau(M_{\sigma(K)} \cdots M_{\sigma(1)}) = 0. \tag{17}$$

If $|X_{-t}| \geq 2$, then $\tau(M_t) = 1$.

Proof. Column constraints force the displayed singleton plan. Substitution into (7) gives the coordinate-resampling kernel. Under a product marginal these kernels commute and resample independent coordinates. After all coordinates are updated, the output law is b independently of the initial column. Distinct fibers have disjoint one-step row supports, giving $\tau(M_t) = 1$ when $|X_{-t}| \geq 2$. Theorem C.4 gives the full derivation. □

For example, on $X = \{0, 1\}^r$ with uniform marginal, a layer that averages bit-flip pairs has $M_t = (I + F_t)/2$, where F_t flips bit t. When $r \geq 2$, every one-step coefficient equals one, but the product of all r layers is $\mathbf{1}\mathbf{1}^\top/2^r$ and has coefficient zero (Corollary C.1). This is an exact score-uniform *quotient-kernel* certificate. The construction has a singleton feasible plan at each layer; it should not be interpreted as evidence of expressive score-dependent attention or downstream improvement.

6 Support-Schedule Design Consequences

The structural characterizations distinguish a nominal support from its realized plan and a single layer from its schedule. Three quantitative consequences make these distinctions concrete; none asserts a task-quality gain.

Locality can be diffusive even when the support is connected. For the equal-score circular band in which row i connects to $i - r, \ldots, i + r$ modulo n, let $d = 2r + 1$ with $0 \leq r < n/2$. The derivative kernel is circulant with Fourier eigenvalues

$$\lambda_0 = 1, \qquad \lambda_k = \left(\frac{\sin(\pi k d/n)}{d \sin(\pi k/n)}\right)^2, \quad k = 1, \ldots, n - 1. \tag{18}$$

Moreover, $\tau(M) = 1$ whenever $n \geq 8r + 2$: two source columns then have disjoint radius-$2r$ derivative neighborhoods. This does not say that all long-window coefficients are one. The sharper obstruction is a slow Fourier mode. For fixed r and growing n,

$$\lambda_1 = 1 - \frac{\pi^2(d^2 - 1)}{3n^2} + O(n^{-4}). \tag{19}$$

The first nonconstant Fourier cotangent contracts in Euclidean norm by exactly λ_1^K after K steps. Thus $K = o(n^2)$ leaves that mode asymptotically uncontracted in this equal-score spectral model (Corollary D.3).

For a general equal-score biregular support, the symmetric kernel has zero-mass spectral rate $\lambda_\star = \|M\Pi_{\mathbf{1}^\perp}\|_2$ and satisfies

$$\tau(M^K) \leq \min\{1, \sqrt{n/2}\, \lambda_\star^K\}. \tag{20}$$

When $\lambda_\star \leq \rho < 1$, the choice $K \geq \lceil \log(\sqrt{n/2}/\epsilon)/(-\log \rho) \rceil$ suffices for a coefficient at most ϵ if $0 < \rho < 1$; if $\rho = 0$, any positive depth suffices. This is a certificate-dependent depth rule, not a conclusion from sparsity alone (Corollary D.1).

Register adjacency must produce common mass. For any realized kernel, the shared mass $\gamma_{\text{shared}} = \sum_\beta \min_\alpha M_{\alpha\beta}$ gives $\tau(M) \le 1 - \gamma_{\text{shared}}$. A sufficient transport-level condition is a set H of hub columns with a mass floor $p_{\min} > 0$: for every source column α and hub h, some witness row has $P_{i\alpha}, P_{ih} \ge p_{\min}$. Then

$$\gamma_{\text{shared}} \ge \frac{|H| p_{\min}^2}{b_{\max} a_{\max}}, \qquad a_{\max} = \max_i a_i, \quad b_{\max} = \max_\alpha b_\alpha. \tag{21}$$

The support provides the witness pattern; the scaled plan provides the floor. Without a floor or assumptions forcing one, a register interpretation is not a quantitative contraction guarantee. The proof is a direct lower bound on a summand in (7); Appendix B also gives the minorization formulation.

An equal-score register construction admits a closed form. Start with a d-regular support on n token rows and columns, where $d \ge 1$, add $r \ge 1$ register rows and columns, connect tokens to every register column, and connect register rows to every column. Use uniform marginals on $N = n + r$ states. If $t > 0$ solves $rt^2 + (n - r)t - d = 0$, then the active token–token, token–register, and register–register plan entries are respectively A, tA, and t^2A, with $A = 1/[N(d + rt)]$. The common register component satisfies

$$\tau(M) \le 1 - \gamma_{\text{reg}}, \qquad \gamma_{\text{reg}} = \frac{rt(d + rt^2)}{(d + rt)^2}. \tag{22}$$

Theorem E.1 proves the plan and this bound. Structured scores require a realized-plan or distortion certificate instead of the equal-score formula. In contrast, the forced-bus certificate (11) survives arbitrary scores because the marginal capacities force its edge floors on every feasible face.

A matched-edge certificate workflow. Evaluate an equal-score baseline, compare the realized powered or windowed coefficient with a local-band control at the same edge budget, and stress it with nonuniform marginals and scores. This proposed workflow identifies candidates for downstream testing; it does not establish task quality.

7 Computational Scope, Limitations, and Conclusion

Certificate cost. Let $d_i = |\Omega_i|$, $E = |\Omega|$, and let s_α be the number of nonzeros in row α of M. Building the balanced kernel by the rank-one contribution of each support row costs $O(\sum_i d_i^2)$ arithmetic, with $\operatorname{nnz}(M) \le \sum_i d_i^2$. It can nevertheless require $O(n^2)$ storage. Exact dense Dobrushin evaluation compares all row pairs and costs $O(n^3)$ time and $O(n^2)$ memory. Sparse row merges cost $O(\sum_{\alpha<\beta}(s_\alpha + s_\beta)) \le O(n \operatorname{nnz}(M))$. Window products can densify even when their factors are sparse.

The face characterizations quantify over all nonempty faces, or face tuples. Exhaustive support-pattern checking is exponential in the number of active edges; the characterization is not a scalable general-purpose algorithm. Practical routes are family theorems, realized-plan certificates, independently certified distortion or mass floors, and searches for counterexample faces. Spectral bounds, minorization, and the product-coordinate identity avoid dense products when their specific hypotheses hold.

Limitations. The face-lattice statements concern balanced feasible plans; finite-cycle contraction uses the two actual half-step plans instead. Global and component quotients must be distinguished. Arbitrary scores can defeat nominal support overlap by approaching bad faces. Source and non-transport operator norms are not supplied by a Dobrushin coefficient. Branching residual graphs require separate path-sum bookkeeping (Appendix G). The manuscript presents theoretical certificates and their costs, not empirical performance results, implementation benchmarks, or downstream task experiments.

Conclusion. Fixed-support Sinkhorn scaling yields an explicit quotient operator and source ledger. Feasible faces characterize when support and marginals force score-uniform mixing; ordered face tuples extend the dichotomy to schedules. Product-coordinate layers show that individually non-mixing kernels can mix perfectly over a window. Sparse transport supports should therefore be evaluated by quotient-gradient mixing under explicit assumptions, not by density or reachability alone.

AI USE STATEMENT

AI tools assisted with exposition and mathematical consistency review during manuscript preparation. For this revision, AI tools assisted with condensing and restating existing exposition, mathematical statements, and proof presentations from the supplied manuscript; adapting the LaTeX to the ICLR 2027 format; and checking submission requirements. This revision does not introduce new experimental results. Responsibility for the mathematical claims, proofs, citations, and final submission remains with the author(s).

REPRODUCIBILITY STATEMENT

This is a theoretical paper. The fixed-support assumptions, quotient conventions, complete half-step and adjoint calculations, and detailed proofs are provided in Appendices A–E. The certificate-level design workflow and residual-DAG qualification appear in Appendices F and G. No empirical results or executable implementation are presented in this version.

REFERENCES

Roberto Cominetti and Jaime San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1–3):169–187, 1994. doi: 10.1007/BF01582220.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.nips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Jesús A. De Loera and Edward D. Kim. Combinatorics and geometry of transportation polytopes: An update. *arXiv preprint arXiv:1307.0124*, 2013. URL https://arxiv.org/abs/1307.0124.

Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, Florian Bernard, and Daniel Cremers. A unified framework for implicit Sinkhorn differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 509–518, 2022.

Dylan B. Forde. Block-Wise Differentiable Sinkhorn Attention: Tail-refinement gradients with a Single-Active-Dustbin Bridge. *arXiv preprint arXiv:2605.08123v2*, 2026. URL https://arxiv.org/html/2605.08123v2. Title as printed in the v2 manuscript; arXiv listing uses Gap-Aware Dustbin Bridge.

Stéphane Gaubert and Zheng Qu. Dobrushin ergodicity coefficient for Markov operators on cones, and beyond. *arXiv preprint arXiv:1302.5226*, 2013.

Martin Idel. A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. *arXiv preprint arXiv:1609.06349*, 2016.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, second edition, 2017. URL https://pages.uoregon.edu/dlevin/MARKOV/. With contributions by Elizabeth L. Wilmer.

Edouard Pauwels and Samuel Vaiter. The derivatives of Sinkhorn–Knopp converge. *SIAM Journal on Optimization*, 33(3):1494–1517, 2023. doi: 10.1137/22M1512703. URL https://epubs.siam.org/doi/10.1137/22M1512703.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Jonathan Weed. An explicit analysis of the entropic penalty in linear programming. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1841–1855, 2018.

Jacob Wolfowitz. Products of indecomposable, aperiodic, stochastic matrices. *Proceedings of the American Mathematical Society*, 14(5):733–737, 1963. doi: 10.1090/S0002-9939-1963-0154756-3.

A Fixed-Support Sinkhorn Calculus and Detailed Proofs

Let $\Omega \subseteq [m] \times [n]$ be a fixed active support with row neighborhoods $\Omega_i = \{j : (i,j) \in \Omega\}$ and column neighborhoods $\Omega^j = \{i : (i,j) \in \Omega\}$. Let $a \in \mathbb{R}^m_{>0}$ and $b \in \mathbb{R}^n_{>0}$ be compatible marginals with $\mathbf{1}^\top a = \mathbf{1}^\top b = T$. Scores are dimensionless active-edge scores $s_{ij} \in \mathbb{R}$ for $(i,j) \in \Omega$; inactive entries are omitted, or viewed as $-\infty$. If raw logits are θ_{ij} at temperature $\varepsilon > 0$, then $s_{ij} = \theta_{ij}/\varepsilon$, so raw-score source terms below acquire the factor $1/\varepsilon$.

A log-potential pair $(u, v) \in \mathbb{R}^m \times \mathbb{R}^n$ represents the active-edge plan

$$P_{ij}(u, v; s) = \exp(s_{ij} + u_i + v_j) \quad ((i,j) \in \Omega), \qquad P_{ij} = 0 \quad ((i,j) \notin \Omega).$$

The pair is a Sinkhorn solution when $P\mathbf{1} = a$ and $P^\top \mathbf{1} = b$. The potentials have gauge freedom: on each connected component of the bipartite support graph, adding a constant to all row potentials and subtracting the same constant from all column potentials leaves P unchanged. The plan and all kernels below are gauge-invariant.

Assumption A.1 (Fixed support, finite scores, and feasible marginals)**.** *The support Ω is fixed during differentiation, all active scores are finite, and every connected component of the active bipartite support has equal row and column marginal mass. When strict positivity on active edges is needed, we assume that $\Pi(a, b; \Omega)$ contains a relative-interior point positive on every active edge.*

For differentiation of transport layers, Eisenberger et al. (2022) develop implicit gradients with respect to costs and target capacities, while Pauwels & Vaiter (2023) establish convergence of differentiated Sinkhorn iterates under their hypotheses. The calculation below makes the individual half-step source terms explicit and distinguishes a finite cycle from a balanced fixed point.

A.1 Half-Step Equations and Linearization

Given a column potential v, the exact row-normalization half-step is

$$u_i^+(v; s, a) = \log a_i - \log \sum_{j \in \Omega_i} \exp(s_{ij} + v_j), \tag{A.1}$$

$$P^r_{ij} = \exp(s_{ij} + u_i^+(v; s, a) + v_j), \qquad \sum_j P^r_{ij} = a_i. \tag{A.2}$$

Given a row potential u, the exact column-normalization half-step is

$$v_j^+(u; s, b) = \log b_j - \log \sum_{i \in \Omega^j} \exp(s_{ij} + u_i), \tag{A.3}$$

$$P^c_{ij} = \exp(s_{ij} + u_i + v_j^+(u; s, b)), \qquad \sum_i P^c_{ij} = b_j. \tag{A.4}$$

A full row-column cycle maps v to $\mathcal{T}(v) = v^+(u^+(v; s, a); s, b)$. At a fixed point of this map, the plan P has both marginals.

For a balanced fixed-point plan P, define

$$R = \operatorname{diag}(a)^{-1} P \in \mathbb{R}^{m \times n}, \qquad C = P \operatorname{diag}(b)^{-1} \in \mathbb{R}^{m \times n}.$$

Rows of R sum to one and columns of C sum to one. For an active-edge score perturbation $\dot{s}$, define the row and column score averages

$$r_i(\dot{s}) = \sum_{j \in \Omega_i} R_{ij} \dot{s}_{ij}, \qquad c_j(\dot{s}) = \sum_{i \in \Omega^j} C_{ij} \dot{s}_{ij}.$$

For marginal perturbations write $d_a = \dot{a} \oslash a$ and $d_b = \dot{b} \oslash b$, componentwise.

Proposition A.1 (Full first-order half-step linearization)**.** *At a balanced fixed-point plan P with positive marginals, the row half-step and column half-step have differentials*

$$\dot{u}^+ = d_a - R\dot{v} - r(\dot{s}), \tag{A.5}$$

$$\dot{v}^+ = d_b - C^\top \dot{u}^+ - c(\dot{s}). \tag{A.6}$$

Consequently one full row-column cycle on column potentials has linearization

$$\dot{v}^+ = C^\top R\,\dot{v} + \underbrace{d_b - C^\top d_a + C^\top r(\dot{s}) - c(\dot{s})}_{\zeta(\dot{s},\dot{a},\dot{b})}. \tag{A.7}$$

For fixed marginals, the source term is $\zeta(\dot{s}) = C^\top r(\dot{s}) - c(\dot{s})$. For raw logits θ at temperature ε, substitute $\dot{s} = \dot{\theta}/\varepsilon$.

Proof. Differentiate (A.1). The derivative of the log-sum-exp normalizer in row i is the R_{ij}-weighted average of $\dot{s}_{ij} + \dot{v}_j$, because at the fixed point $R_{ij} = P_{ij}/a_i$. This gives $\dot{u}_i^+ = \dot{a}_i/a_i - \sum_j R_{ij}\dot{v}_j - \sum_j R_{ij}\dot{s}_{ij}$, which is (A.5). Differentiating (A.3) gives $\dot{v}_j^+ = \dot{b}_j/b_j - \sum_i C_{ij}\dot{u}_i^+ - \sum_i C_{ij}\dot{s}_{ij}$, which is (A.6). Substituting the first display into the second gives (A.7). The signs come from the fact that both half-steps subtract a log normalizer; the two minus signs in the homogeneous composition yield the positive map $C^\top R$. □

Finite-cycle extension: two distinct half-step plans. The preceding single-plan formulas are evaluated at a balanced fixed point. For an actual finite cycle, let P_t^r and P_t^c be the row-normalized and column-normalized plans from (A.1)–(A.3), and define

$$R_t = \mathrm{diag}(a)^{-1}P_t^r, \qquad C_t = P_t^c\,\mathrm{diag}(b)^{-1}, \qquad M_t = C_t^\top R_t.$$

With $r_t(\dot{s})_i = \sum_j (R_t)_{ij}\dot{s}_{ij}$ and $c_t(\dot{s})_j = \sum_i (C_t)_{ij}\dot{s}_{ij}$, direct differentiation gives

$$\dot{v}^+ = M_t\dot{v} + d_b - C_t^\top d_a + C_t^\top r_t(\dot{s}) - c_t(\dot{s}).$$

Indeed, each log-normalizer differentiates to the conditional probabilities of its own half-step, and composing the two minus signs gives M_t. Rows of R_t and columns of C_t are normalized, so $M_t\mathbf{1} = \mathbf{1}$. Transposing this identity gives the source ledger in Proposition A.2 with R_t, C_t in place of R, C. In general $P_t^r \neq P_t^c$; neither plan may be substituted for both. The finite kernel need not be reversible or have stationary mass b. The general stochastic-kernel tail bounds below apply to these actual M_t. By contrast, the single-plan support weights, face-lattice characterization, and stationary implicit system concern balanced feasible plans, not arbitrary unbalanced intermediate iterates.

Definition A.1 (Quotient convention)**.** *For a connected support, column potentials are functions modulo additive constants. We write $Q = \mathbb{R}^n/\mathrm{span}\{\mathbf{1}\}$ and identify its dual with*

$$Q^* = \{\eta \in \mathbb{R}^n : \mathbf{1}^\top\eta = 0\}.$$

The total-variation norm on Q^* is $\|\eta\|_{\mathrm{TV}} = \frac{1}{2}\|\eta\|_1$. Centering, for example $\Pi_0 x = x - (\mathbf{1}^\top x/n)\mathbf{1}$, chooses a representative only. Since the kernels below are row-stochastic, they satisfy $M\mathbf{1} = \mathbf{1}$ and therefore $M^\top Q^* \subseteq Q^*$. Dobrushin certificates are statements on this quotient.

If the bipartite support has components Γ with column sets J_Γ, the full potential gauge space is $G_\Omega = \mathrm{span}\{\mathbf{1}_{J_\Gamma} : \Gamma\}$ and the exact component quotient is

$$Q_\Omega = \mathbb{R}^n/G_\Omega, \qquad Q_\Omega^* = \left\{\eta : \mathbf{1}_{J_\Gamma}^\top\eta = 0 \text{ for every } \Gamma\right\}.$$

Every single-support statement below remains valid componentwise after replacing (Q, Q^*) by (Q_Ω, Q_Ω^*) and centering separately on each J_Γ. For schedules whose component partitions change, the global quotient Q is the common state space. Its Dobrushin coefficient may equal one because it retains intercomponent Markov modes; that value must not be reported as a sharp contraction rate on the smaller component quotient.

A.2 Derivative Kernel and Overlap

At a balanced fixed point, the column-to-column homogeneous full-step kernel is

$$M = C^\top R \in \mathbb{R}^{n\times n}.$$

It is row-stochastic, and $b^\top M = b^\top$, so b is a stationary mass vector (and b/T a stationary distribution). The map M is the fixed-support quotient operator for the homogeneous part of one row-column Sinkhorn cycle; source terms enter through ζ in (A.7) and through the projected cotangent sources in Section B.

Proposition A.2 (Adjoint source ledger for one fixed-support cycle)**.** *Let $\eta^+ \in \mathbb{R}^n$ be a cotangent on the output column potential v^+ of one row-column cycle at a balanced fixed point. For a finite cycle, use the two-half-step extension above. The homogeneous cotangent on the input column potential is*

$$\eta^- = M^\top \eta^+.$$

The cotangents carried by the non-homogeneous source variables in (A.7) *are*

$$\lambda_b = \eta^+, \qquad \lambda_a = -C\eta^+, \tag{A.8}$$

$$g^s_{ij} = R_{ij}(C\eta^+)_i - C_{ij}\eta^+_j, \qquad (i,j) \in \Omega. \tag{A.9}$$

Here λ_a and λ_b are cotangents to the logarithmic marginal perturbations $d_a = \dot{a} \oslash a$ and $d_b = \dot{b} \oslash b$. Equivalently, the raw marginal and raw-logit cotangents are

$$g^a = -(C\eta^+) \oslash a, \qquad g^b = \eta^+ \oslash b, \qquad g^\theta_{ij} = g^s_{ij}/\varepsilon$$

when $s_{ij} = \theta_{ij}/\varepsilon$.

The displayed marginal covectors are unconstrained source covectors. If the marginals are parameterized inside the compatible fixed-component marginal manifold, only their quotient classes are observable. For each connected component $\Gamma = (I_\Gamma, J_\Gamma)$ of the support graph, admissible log-marginal perturbations satisfy

$$\sum_{i \in I_\Gamma} a_i (d_a)_i = \sum_{j \in J_\Gamma} b_j (d_b)_j.$$

Hence the log-marginal cotangent pair (λ_a, λ_b) is defined modulo the component normal $(a_{I_\Gamma}, -b_{J_\Gamma})$. Equivalently, any representative

$$(\lambda_a, \lambda_b) - \alpha_\Gamma (a_{I_\Gamma}, -b_{J_\Gamma})$$

gives the same derivative on admissible perturbations. A concrete Euclidean orthogonal representative is obtained with

$$\alpha_\Gamma = \frac{\sum_{i \in I_\Gamma} a_i (\lambda_a)_i - \sum_{j \in J_\Gamma} b_j (\lambda_b)_j}{\sum_{i \in I_\Gamma} a_i^2 + \sum_{j \in J_\Gamma} b_j^2}.$$

In raw marginal coordinates the analogous quotient is modulo $(\mathbf{1}_{I_\Gamma}, -\mathbf{1}_{J_\Gamma})$. If only one marginal family is varied while the other is fixed, this reduces to the usual componentwise fixed-mass projection of that marginal covector modulo constants on the varied component. If each component mass is fixed separately in both marginal families, then admissible raw variations obey $\mathbf{1}_{I_\Gamma}^\top \dot{a} = 0$ and $\mathbf{1}_{J_\Gamma}^\top \dot{b} = 0$ independently. The raw cotangents are then defined modulo $(\mathbf{1}_{I_\Gamma}, 0)$ and $(0, \mathbf{1}_{J_\Gamma})$ independently; in logarithmic coordinates the corresponding normals are $(a_{I_\Gamma}, 0)$ and $(0, b_{J_\Gamma})$. Thus the projection is determined by the actual marginal parameterization: one joint normal when common component mass may vary, two normals when that mass is fixed.

If surrounding differentiable maps feed these source variables from the input quotient coordinates at step t, their VJPs compose with (A.8)–(A.9) *and are then projected to Q^*. Thus the tail recurrence used below has the form*

$$\eta_t = M_t^\top \eta_{t+1} + \Pi_{Q^*} \psi_t,$$

where ψ_t denotes the aggregate non-transport cotangent delivered to the input column-potential coordinates before quotient projection. The displayed source adjoints are the complete one-cycle ledger; they are not additional applications of $M^\top$.

Proof. Pair an arbitrary perturbation with (A.7):

$$\langle \eta^+, \dot{v}^+ \rangle = \langle \eta^+, M\dot{v} \rangle + \langle \eta^+, d_b \rangle - \langle \eta^+, C^\top d_a \rangle + \langle \eta^+, C^\top r(\dot{s}) \rangle - \langle \eta^+, c(\dot{s}) \rangle.$$

The first term is $\langle M^\top \eta^+, \dot{v} \rangle$, giving $\eta^- = M^\top \eta^+$. The next two terms give $\lambda_b = \eta^+$ and $\lambda_a = -C\eta^+$. For scores,

$$\langle \eta^+, C^\top r(\dot{s}) \rangle = \sum_i (C\eta^+)_i \sum_j R_{ij} \dot{s}_{ij}, \qquad \langle \eta^+, c(\dot{s}) \rangle = \sum_j \eta^+_j \sum_i C_{ij} \dot{s}_{ij}.$$

Collecting the coefficient of each active $\dot{s}_{ij}$ gives (A.9). The raw marginal formulas follow from $d_a = \dot{a} \oslash a$ and $d_b = \dot{b} \oslash b$, and the raw-logit formula from $s = \theta/\varepsilon$. The componentwise marginal projection statement is just duality for the linear compatibility constraint $a_{I_\Gamma}^\top d_{a,I_\Gamma} - b_{J_\Gamma}^\top d_{b,J_\Gamma} = 0$: adding any multiple of its normal to the covector pairs to zero with every admissible perturbation. Finally, quotient projection to Q^* is necessary for cotangents returned to column-potential quotient coordinates; source-space cotangents use the corresponding source constraints described above. □

Lemma A.1 (Bridge from fixed-support half-steps to quotient topology)**.** *Let P be a realized active-support Sinkhorn plan with compatible marginals (a, b). For homogeneous perturbations of the column log potential,*

$$\dot{v}^+ = C^\top R\, \dot{v} = M \dot{v}.$$

Hence the homogeneous reverse cotangent propagation on Q^* is governed by $M^\top$.

Proof. Set $\dot{s} = 0$, $\dot{a} = 0$, and $\dot{b} = 0$ in Proposition A.1. The resulting forward linear map is $M = C^\top R$. Equivalently, the homogeneous part of Proposition A.2 is $\eta^- = M^\top \eta^+$. By Definition A.1, $M^\top$ preserves zero-mass cotangents and is independent of the chosen centered representative. □

Definition A.2 (Transport topology certificate)**.** *For a fixed support Ω, score-induced plan P, and quotient operator $M = C^\top R$, define*

$$\tau(M) = \frac{1}{2} \max_{p,q} \sum_k |M_{pk} - M_{qk}|.$$

We call $\tau(M)$ the one-step Dobrushin transport-topology certificate. When $\tau(M) < 1$, the realized fixed-support quotient operator contracts all zero-mass cotangents. When $\tau(M) = 1$, this certificate withholds a global one-step contraction guarantee.

Theorem A.1 (Support-to-kernel topology)**.** *For $\alpha, \beta \in [n]$, the full-step derivative kernel satisfies*

$$M_{\alpha\beta} = \sum_i \frac{P_{i\alpha} P_{i\beta}}{b_\alpha a_i}.$$

If all active plan entries are strictly positive, then

$$M_{\alpha\beta} > 0 \quad \Longleftrightarrow \quad \exists i \text{ such that } (i,\alpha) \in \Omega \text{ and } (i,\beta) \in \Omega.$$

Thus the support of each row of M is exactly the two-hop column co-neighborhood induced by the bipartite support Ω.

Proof. Substitute $C = P\,\mathrm{diag}(b)^{-1}$ and $R = \mathrm{diag}(a)^{-1} P$ into $M = C^\top R$:

$$M_{\alpha\beta} = \sum_i C_{i\alpha} R_{i\beta} = \sum_i \frac{P_{i\alpha}}{b_\alpha} \frac{P_{i\beta}}{a_i}.$$

Every summand is nonnegative. Under strict positivity on active edges, a summand is positive exactly when row i is incident to both columns α and β. □

Corollary A.1 (Qualitative overlap criterion)**.** *Under the same strict-positivity assumption, $\tau(M) < 1$ if and only if every pair of rows of the two-hop kernel support overlap. Equivalently, $\tau(M) = 1$ if and only if some two columns have disjoint two-hop co-neighborhoods.*

Proof. Dobrushin's coefficient is strictly below one exactly when every pair of row distributions has positive common mass. Theorem A.1 identifies those row supports with the two-hop co-neighborhoods. □

Structure of the remaining proofs. The remaining development asks when this overlap is forced rather than merely observed. The equal-score biregular case gives a graph-walk baseline; plan-distortion and zero-temperature arguments explain how scores can expose bad faces; the face-lattice theorem gives the exact score-uniform one-step criterion; and the windowed face-lattice plus partition heat-bath results show how scheduled sparse layers can mix even when every individual layer has $\tau = 1$.

A.3 Equal-Score Topology Baseline and Plan Distortion

Theorem A.2 (Biregular equal-score topology kernel)**.** *Let $\Omega \subseteq [m] \times [n]$ be a fixed bipartite support with constant positive row degree d_r and constant positive column degree d_c. Use uniform compatible marginals $a_i = 1/m$ and $b_\alpha = 1/n$, and set all active effective scores equal. Let $E = |\Omega| = md_r = nd_c$. Then the Sinkhorn-scaled plan is*

$$P_{i\alpha} = \frac{1}{E}\mathbf{1}\{(i,\alpha) \in \Omega\}.$$

Consequently,

$$M_{\alpha\beta} = \frac{N_{\alpha\beta}}{d_c d_r}, \qquad N_{\alpha\beta} = \#\{i : \ (i,\alpha) \in \Omega,\ (i,\beta) \in \Omega\}.$$

Thus, in this baseline, M is exactly the two-step random walk on the bipartite support graph: from column α, choose an incident row uniformly, then choose an incident column of that row uniformly. In particular, $\tau(M)$ and $\tau(M^K)$ are pure topology quantities.

Proof. Since $|\Omega| = md_r = nd_c$, the displayed plan has row sums $d_r/|\Omega| = 1/m$ and column sums $d_c/|\Omega| = 1/n$. It is therefore a valid Sinkhorn scaling of the equal-score active matrix; uniqueness of the entropic matrix-scaling plan fixes this primal plan even if the support has multiple compatible connected components. The dual scalings may have separate component gauges, but P and M are unchanged. For active edges, $R_{i\alpha} = P_{i\alpha}/a_i = 1/d_r$ and $C_{i\alpha} = P_{i\alpha}/b_\alpha = 1/d_c$. Substituting into $M = C^\top R$ gives

$$M_{\alpha\beta} = \sum_i \frac{\mathbf{1}\{(i,\alpha) \in \Omega\}\mathbf{1}\{(i,\beta) \in \Omega\}}{d_c d_r} = \frac{N_{\alpha\beta}}{d_c d_r}.$$

The row sum is one because each of the d_c rows incident to α has d_r incident columns. □

Theorem A.2 is the cleanest fixed-support certificate baseline in this paper. For equal-score biregular supports, a sparse attention topology has an exact Markov-chain representation: its Sinkhorn linearization is the Markov chain on the column-overlap graph. Score-dependent and non-biregular cases then become perturbations or realized-plan certificates rather than graph-only guarantees. If the support has multiple closed column-overlap components, the same formula holds componentwise but the global Dobrushin coefficient remains one, so the theorem does not certify mixing across disconnected topology.

Theorem A.3 (Plan-distortion stability around a topology kernel)**.** *Let Ω be biregular with row degree d_r, column degree d_c, uniform marginals, and $E = |\Omega|$. Let M_0 be the equal-score topology kernel from Theorem A.2. Let P be any other feasible plan on the same support with the same uniform marginals, and write*

$$q_{i\alpha} = EP_{i\alpha} \qquad ((i,\alpha) \in \Omega).$$

If for some $\kappa \geq 1$,

$$\kappa^{-1} \leq q_{i\alpha} \leq \kappa \qquad \text{for every active edge } (i,\alpha) \in \Omega,$$

then the realized derivative kernel $M = C^\top R$ satisfies the elementwise comparison

$$M \geq \kappa^{-2} M_0.$$

Consequently,

$$\tau(M) \leq 1 - \kappa^{-2}\big(1 - \tau(M_0)\big).$$

More generally, if row-stochastic kernels M_t and reference kernels M_t^0 on the same finite state space satisfy $M_t \geq \delta_t M_t^0$ elementwise with $\delta_t \in [0,1]$ for $t = 1, \ldots, K$, then

$$M_K \cdots M_1 \geq \left(\prod_{t=1}^K \delta_t\right) M_K^0 \cdots M_1^0$$

and hence

$$\tau(M_K \cdots M_1) \leq 1 - \left(\prod_{t=1}^K \delta_t\right)\big(1 - \tau(M_K^0 \cdots M_1^0)\big).$$

Proof. With uniform marginals, $a_i = 1/m$, $b_\alpha = 1/n$, and $E = md_r = nd_c$. For the realized plan,

$$M_{\alpha\beta} = \sum_i \frac{P_{i\alpha}P_{i\beta}}{b_\alpha a_i} = \frac{1}{d_c d_r} \sum_{i:\,(i,\alpha),(i,\beta)\in\Omega} q_{i\alpha}q_{i\beta}.$$

The edgewise distortion assumption gives $q_{i\alpha}q_{i\beta} \geq \kappa^{-2}$ for every common incident row, while $(M_0)_{\alpha\beta}$ is the same sum with each common-row contribution equal to $1/(d_c d_r)$. This proves $M \geq \kappa^{-2} M_0$.

If two row-stochastic kernels A, B satisfy $A \geq \delta B$ elementwise, then $A = \delta B + (1-\delta)N$ for a row-stochastic kernel N when $\delta < 1$ (with the evident interpretation when $\delta = 1$). Thus

$$\tau(A) \leq \delta\tau(B) + (1-\delta) = 1 - \delta(1-\tau(B)).$$

This gives the one-step Dobrushin bound. The product comparison follows by iterating elementwise nonnegative matrix multiplication: if $M_t \geq \delta_t M_t^0$, then every term in $M_K \cdots M_1$ dominates the corresponding term in $(\prod_t \delta_t) M_K^0 \cdots M_1^0$. Applying the same mixture argument to the product gives the final display. □

Theorem A.3 is the first bridge from pure topology to score-dependent plans in this paper. It is not a claim that graph support alone controls the magnitude of $\tau(M)$. It says that once a plan distortion factor is observed or independently certified, the topology certificate degrades continuously rather than disappearing. This is a suitable form for score-robust certificate checks and for score-range corollaries.

A.4 Score Gauges and Cycle Sensitivity

Proposition A.3 (Additive scores are topology gauges)**.** *Fix a support Ω and compatible marginals a, b, and assume strict feasibility: there exists $Q \in \Pi(a, b; \Omega)$ with $Q_{ij} > 0$ on every fixed-support edge. Suppose a score pattern is row-plus-column separable on each connected component of the fixed support:*

$$B_{ij} = \alpha_i + \beta_j + c_\Gamma \qquad ((i,j) \in \Gamma),$$

where Γ ranges over connected components of the bipartite support. Let P_L be the entropic Sinkhorn plan for score $S = LB$, and let P_0 be the plan for the zero score on the same support and marginals. Then

$$P_L = P_0 \qquad \text{for every } L.$$

Consequently the derivative kernel $M = C^\top R$ is unchanged and the plan-distortion factor relative to P_0 is $\kappa_L = 1$. In the uniform-marginal biregular case, the topology kernel remains exactly the two-step support random walk from Theorem A.2.

Proof. For every feasible plan $P \in \Pi(a, b; \Omega)$,

$$\langle B, P\rangle = \sum_i \alpha_i \sum_j P_{ij} + \sum_j \beta_j \sum_i P_{ij} + \sum_\Gamma c_\Gamma m_\Gamma = \alpha^\top a + \beta^\top b + \sum_\Gamma c_\Gamma m_\Gamma,$$

where m_Γ is the fixed marginal mass in component Γ. This is constant over the transport polytope. Therefore maximizing $L\langle B, P\rangle + H(P)$ is the same optimization problem as maximizing $H(P)$ alone. The primal plan is thus P_0 for every L, so the derivative kernel and the distortion factor are unchanged. □

On a connected support, the additive condition is equivalent to vanishing alternating cycle sums for B. Thus row/column gauge scores and rank-one Gibbs kernels $\exp(LB_{ij}) = \exp(L\alpha_i)\exp(L\beta_j)$ cannot create plan distortion. Score sensitivity is carried by non-additive cycle imbalance, not by raw score range alone.

Corollary A.2 (Cycle space is the score-sensitive topology)**.** *Fix compatible positive marginals a, b and assume strict feasibility on the active-edge set Ω. Let the fixed support have c connected components, with active row and column vertex sets I and J. The quotient dimension of score patterns that can affect the entropic plan, after modding out row/column additive gauges on each component, is*

$$|\Omega| - |I| - |J| + c,$$

the cycle rank of the bipartite support graph. In particular, if every connected component is a tree and the compatible transport polytope is strictly feasible, then $\Pi(a, b; \Omega)$ is a singleton. Hence P_L, the derivative kernel $M = C^\top R$, and all quotient-cotangent certificates are independent of the score pattern B.

Proof. On a connected component, row-plus-column scores form a subspace of dimension $|I_\Gamma| + |J_\Gamma| - 1$, since adding a constant to all row prices and subtracting it from all column prices gives the same edge scores. Summing over components gives additive-score dimension $|I| + |J| - c$. Quotienting the $|\Omega|$ active-edge score coordinates by this gauge subspace leaves $|\Omega| - |I| - |J| + c$, the graph cycle rank. Proposition A.3 shows that these additive directions do not change the entropic optimization problem.

Conversely, additive gauges are the only score directions that leave the entropic plan unchanged. Let P_0 be the zero-score entropic plan and suppose the score B has the same entropic plan P_0. The KKT equations for the two strictly concave entropy-regularized transport problems give dual prices (u^0, v^0) and (u^B, v^B) such that, on every active edge,

$$\log(P_0)_{ij} = -u_i^0 - v_j^0 = B_{ij} - u_i^B - v_j^B.$$

Thus $B_{ij} = (u_i^B - u_i^0) + (v_j^B - v_j^0)$ on the support, up to the usual component gauges. Therefore a non-gauge score direction changes the entropic plan for some finite score scale, and the score-sensitive quotient dimension is exactly the cycle rank.

If the support is a forest, the cycle rank is zero. Equivalently, the transport constraints have a unique feasible solution on each tree component: remove a leaf and its incident edge recursively; the edge mass is forced by the leaf marginal, and strict feasibility keeps the forced masses positive. Thus every score pattern is a gauge direction on the support, and the plan and its derivative kernel are fixed by the marginals and support alone. □

A.5 Zero-Temperature Face Obstructions

Theorem A.4 (Zero-temperature entropy-center obstruction)**.** *Let $\Pi(a, b; \Omega)$ be the compact transport polytope on a fixed feasible support Ω, and assume it contains a relative-interior point that is positive on every active edge. Let B be a fixed score pattern on Ω. For $L > 0$, let P_L be the entropic Sinkhorn plan for score $S = LB$ and marginals a, b, equivalently the maximizer of*

$$L\langle B, P\rangle + H(P), \qquad H(P) = -\sum_{(i,j)\in\Omega} P_{ij} \log P_{ij},$$

over $P \in \Pi(a, b; \Omega)$, with the convention $0 \log 0 = 0$. Let

$$\mathcal{F}_B = \arg\max_{P\in\Pi(a,b;\Omega)} \langle B, P\rangle$$

be the optimal face of the zero-temperature linear transport problem. Let P_B^{ent} be the unique maximizer of H over $\mathcal{F}_B$. Then

$$P_L \to P_B^{\mathrm{ent}} \qquad \text{as } L \to \infty.$$

Consequently, if an active edge $e \in \Omega$ has $P_e = 0$ for every $P \in \mathcal{F}_B$, then $(P_L)_e \to 0$. In particular, any edgewise distortion factor relative to a positive reference plan on all of Ω diverges.

Proof. Let $M_B = \max_{P\in\Pi(a,b;\Omega)} \langle B, P\rangle$ and choose $P^\star \in \mathcal{F}_B$. Since H is strictly concave and continuous on the compact polytope, its restriction to the compact convex face $\mathcal{F}_B$ has a unique maximizer P_B^{ent}.

Since $\Pi(a, b; \Omega)$ is compact and H is bounded above and below on it, there is a finite constant C such that $|H(P) - H(P^\star)| \le C$ for every feasible P. Optimality of P_L gives

$$L\langle B, P_L\rangle + H(P_L) \ge LM_B + H(P^\star),$$

hence $\langle B, P_L\rangle \ge M_B - C/L$. Every accumulation point $\bar{P}$ of P_L is feasible and therefore satisfies $\langle B, \bar{P}\rangle = M_B$, so $\bar{P} \in \mathcal{F}_B$.

Now use P_B^{ent} as the comparison point. Optimality gives

$$L\langle B, P_L\rangle + H(P_L) \geq LM_B + H(P_B^{\mathrm{ent}}).$$

Writing $g_L = M_B - \langle B, P_L\rangle \geq 0$, this says $H(P_L) \geq H(P_B^{\mathrm{ent}}) + Lg_L \geq H(P_B^{\mathrm{ent}})$. Therefore any accumulation point $\bar{P} \in \mathcal{F}_B$ satisfies $H(\bar{P}) \geq H(P_B^{\mathrm{ent}})$. By uniqueness of the entropy maximizer on $\mathcal{F}_B$, $\bar{P} = P_B^{\mathrm{ent}}$. All subsequential limits are the same, hence the whole sequence converges.

If an edge e has zero mass on every point of $\mathcal{F}_B$ but some subsequence has $(P_{L_k})_e \geq \epsilon > 0$, compactness gives a further subsequence converging to an element of $\mathcal{F}_B$ with edge mass at least ϵ, a contradiction. Thus $(P_L)_e \to 0$. If the reference plan is positive on e, the ratio between the reference edge mass and $(P_L)_e$ diverges. □

Theorem A.4 gives the zero-temperature reading of the topology certificate. Equal-score topology describes the reference Markov operator. Large structured scores select an entropy-centered point of an optimal face of a linear transport problem. Edges outside that face disappear, and the realized derivative kernel can inherit the topology of the selected face rather than the topology of the nominal architecture. This is the finite-dimensional entropic-penalty central-path phenomenon specialized to transport polytopes; see Cominetti & San Martín (1994) and Weed (2018) for general LP analyses.

Corollary A.3 (Exposed bad-face obstruction to graph-only contraction)**.** *Assume the setting of Theorem A.4. Suppose there is a score direction B such that the entropy-centered point P_B^{ent} of the exposed zero-temperature face has derivative kernel*

$$M_B = \left(P_B^{\mathrm{ent}}\,\mathrm{diag}(b)^{-1}\right)^{\top}\left(\mathrm{diag}(a)^{-1}P_B^{\mathrm{ent}}\right)$$

with $\tau(M_B) = 1$. Then for every $\varepsilon > 0$ there exists L such that the entropic plan for score LB has derivative kernel M_L satisfying

$$\tau(M_L) > 1 - \varepsilon.$$

Thus this support/marginal instance admits no score-uniform global column-quotient upper bound $\tau(M) \leq \rho < 1$ over arbitrary score directions with those fixed compatible marginals.

Proof. Theorem A.4 gives $P_L \to P_B^{\mathrm{ent}}$. The map

$$P \mapsto \left(P\,\mathrm{diag}(b)^{-1}\right)^{\top}\left(\mathrm{diag}(a)^{-1}P\right)$$

is continuous because the positive marginals a, b are fixed. Dobrushin's coefficient is a maximum over finitely many continuous row-pair total-variation distances, hence it is continuous. Therefore $\tau(M_L) \to \tau(M_B) = 1$, which gives the claim. □

Theorem A.5 (Face-lattice characterization of score-uniform global mixing)**.** *Let*

$$\mathcal{P} = \Pi(a, b; \Omega)$$

be a nonempty compact transport polytope on a fixed finite support, with positive compatible marginals and a relative-interior point positive on every edge of Ω. This strict-feasibility hypothesis is essential: finite scores then produce plans in $\mathrm{relint}(\mathcal{P})$, while boundary faces appear only as zero-temperature limits. For any feasible plan $P \in \mathcal{P}$, define

$$M(P) = \left(P\,\mathrm{diag}(b)^{-1}\right)^{\top}\left(\mathrm{diag}(a)^{-1}P\right), \qquad \Gamma(P) = \min_{\alpha,\beta}\sum_{\gamma}\min\{M(P)_{\alpha\gamma}, M(P)_{\beta\gamma}\}.$$

Thus $\tau(M(P)) = 1 - \Gamma(P)$. For a nonempty face F of $\mathcal{P}$, write

$$\Omega_F = \{(i,\alpha) \in \Omega : \text{ some } P \in F \text{ has } P_{i\alpha} > 0\}.$$

Say that F has pairwise two-hop column overlap if, for every pair of columns α, β, there is a column γ such that each of α and β shares an active row with γ inside Ω_F:

$$\exists i, k \quad (i,\alpha), (i,\gamma), (k,\beta), (k,\gamma) \in \Omega_F.$$

Then the following are equivalent.

1. *There exists $\delta > 0$ such that every finite-score Sinkhorn plan on the fixed support satisfies $\Gamma(P_\Omega(S; a, b)) \geq \delta$, or equivalently $\tau(M(P_\Omega(S; a, b))) \leq 1 - \delta$.*

2. *Every nonempty face of $\mathcal{P}$ has pairwise two-hop column overlap.*

Equivalently, a fixed support and compatible marginals force score-uniform one-step global column-quotient mixing exactly when no feasible face can collapse the derivative kernel into two rows with disjoint support. Componentwise contraction would require restating the condition on the corresponding component quotient.

Proof. First, $M(P)$ and $\Gamma(P)$ are continuous functions of P on $\mathcal{P}$, because the marginals a, b are fixed and positive and Γ is a minimum of finitely many continuous row-overlap functions. For a plan P, the support-to-kernel identity gives

$$M(P)_{\alpha\gamma} > 0 \quad \Longleftrightarrow \quad \exists i \text{ with } P_{i\alpha} > 0 \text{ and } P_{i\gamma} > 0.$$

Transportation polytopes are standard-form polytopes $\{x \geq 0 : Ax = c\}$, so every face is obtained by setting a subset of edge variables to zero. Hence the minimal face containing P has support $\operatorname{supp}(P)$, and the relative interior of a face F is positive exactly on Ω_F. Therefore $\Gamma(P) > 0$ if and only if the minimal face containing P has pairwise two-hop column overlap.

Assume every nonempty face has pairwise two-hop column overlap. Then $\Gamma(P) > 0$ for every $P \in \mathcal{P}$, including boundary plans. By compactness and continuity,

$$\delta_* = \min_{P \in \mathcal{P}} \Gamma(P) > 0.$$

Every finite-score Sinkhorn plan lies in $\mathcal{P}$, so $\Gamma(P_\Omega(S; a, b)) \geq \delta_*$ for every finite score matrix.

Conversely, suppose a nonempty face F fails pairwise two-hop column overlap. Every face of a finite polytope is exposed, so there is a score direction B whose maximizers over $\mathcal{P}$ are exactly F. By Theorem A.4, the entropic Sinkhorn plan for score LB converges as $L \to \infty$ to the unique entropy maximizer P_F^{ent} over F. The entropy maximizer over a nonempty face lies in $\operatorname{relint}(F)$: if an edge in Ω_F had zero mass, moving from the maximizer a small distance toward any relative-interior point of F would increase entropy to first order because $-\partial_x(x \log x) \to +\infty$ as $x \downarrow 0$. Thus the positive support of P_F^{ent} is exactly Ω_F.

Since Ω_F fails pairwise two-hop column overlap, two rows of $M(P_F^{\mathrm{ent}})$ have disjoint support. Hence $\Gamma(P_F^{\mathrm{ent}}) = 0$ and $\tau(M(P_F^{\mathrm{ent}})) = 1$. Continuity gives $\Gamma(P_\Omega(LB; a, b)) \to 0$, so no positive score-uniform lower bound δ can hold over finite scores. □

Corollary A.4 (Score-uniform mixing dichotomy)**.** *Under the hypotheses of Theorem A.5, exactly one of the following alternatives holds.*

1. *There is a constant $\rho < 1$, depending only on (a, b, Ω), such that every finite-score Sinkhorn plan on the fixed support satisfies $\tau(M(P_\Omega(S; a, b))) \leq \rho$.*
2. *For every $\varepsilon > 0$, there is a finite score matrix S on the same support such that*

$$\tau(M(P_\Omega(S; a, b))) > 1 - \varepsilon.$$

The first alternative holds exactly when every nonempty feasible face of $\Pi(a, b; \Omega)$ has pairwise two-hop column overlap.

Proof. If every nonempty face has pairwise two-hop column overlap, the theorem gives a positive lower bound $\Gamma(P) \geq \delta$ for every finite-score plan, so setting $\rho = 1 - \delta$ proves the first alternative. If some face fails the overlap condition, the converse part of the theorem constructs an exposed score direction B for which $\Gamma(P_\Omega(LB; a, b)) \to 0$ as $L \to \infty$. Choosing L large enough gives $\tau(M(P_\Omega(LB; a, b))) > 1 - \varepsilon$, which proves the second alternative. The alternatives are mutually exclusive because $\tau(M) \leq \rho < 1$ for all finite scores is incompatible with finite scores whose coefficients exceed $1 - \varepsilon$ for every $\varepsilon > 0$. □

Theorem A.5 and Corollary A.4 are the sharp fixed-support/marginal results of this paper. They explain why equal-score topology can be predictive while arbitrary scores remain dangerous: a nominally well-connected support may still have a bad boundary face, and large scores can expose

that face. They also give a concrete design target. To make topology force mixing, one must rule out bad feasible faces, or else work with realized-plan, windowed, or plan-distortion certificates.

Exact checking should not be oversold. For a fixed tested column pair α, β, the bad-face search can be organized as a finite disjunction of ordinary transport-feasibility problems: after guessing the rows carrying positive mass into α and β and guessing, for each intermediate column, on which side the two-hop overlap is killed, each resulting subproblem has standard Hall/Gale or max-flow certificates. The difficult part is the exponential support-pattern choice. Thus face enumeration is suitable only as a tiny diagnostic, while Corollaries A.5 and A.6 are checkable sufficient architecture certificates.

Corollary A.5 (Capacity-deficit forced bus edges)**.** *Assume the hypotheses of Theorem A.5. For an active edge $(i, \alpha) \in \Omega$, define*

$$d_{i\alpha} = \left[b_\alpha - \sum_{k:\,(k,\alpha)\in\Omega,\ k\neq i} a_k \right]_+ ,$$

and set $d_{i\alpha} = 0$ for inactive edges. Then every feasible plan satisfies $P_{i\alpha} \geq d_{i\alpha}$. Let B be a set of bus rows. Suppose that for every pair of distinct columns α, β, there exists $i \in B$ with $d_{i\alpha} > 0$ and $d_{i\beta} > 0$. Then every feasible face has pairwise two-hop column overlap and the support–marginal pair has a score-uniform one-step global column-quotient certificate. More quantitatively,

$$\tau(M(P)) \leq 1 - \gamma_0 \qquad \textit{for every feasible } P,$$

where

$$\gamma_0 = \min_{\alpha\neq\beta} \max_{i\in B:\, d_{i\alpha}d_{i\beta}>0} \sum_{\gamma:\, d_{i\gamma}>0} \min\left\{ \frac{d_{i\alpha}d_{i\gamma}}{b_\alpha a_i}, \frac{d_{i\beta}d_{i\gamma}}{b_\beta a_i} \right\}.$$

Proof. For any feasible plan,

$$P_{i\alpha} = b_\alpha - \sum_{k\neq i} P_{k\alpha} \geq b_\alpha - \sum_{k:\,(k,\alpha)\in\Omega,\ k\neq i} a_k$$

on active edges, and the positive part gives the lower bound. If $d_{i\alpha}, d_{i\beta} > 0$, then for every γ with $d_{i\gamma} > 0$,

$$M_{\alpha\gamma} \geq \frac{d_{i\alpha}d_{i\gamma}}{b_\alpha a_i}, \qquad M_{\beta\gamma} \geq \frac{d_{i\beta}d_{i\gamma}}{b_\beta a_i}.$$

Summing coordinatewise minima over such γ gives a common mass lower bound for rows α and β of M. Taking the best bus row for each pair and then the worst pair gives $\Gamma(P) \geq \gamma_0 > 0$, hence $\tau(M(P)) \leq 1 - \gamma_0$. Since the lower bounds hold on every feasible plan and every feasible face, Theorem A.5 also gives the qualitative face-lattice conclusion. □

Corollary A.5 is the support-schedule version of the face-lattice theorem. It says that support adjacency alone is not enough: support and marginals must force positive edge flow on every feasible face. The following special case is easier to read and useful for sanity checks.

Corollary A.6 (Heavy universal bus row)**.** *Assume the hypotheses of Theorem A.5, and let $T = \mathbf{1}^\top a = \mathbf{1}^\top b$. Suppose there is a row h whose support contains every column, and suppose*

$$\delta_j := a_h + b_j - T > 0 \qquad \textit{for every column } j.$$

Then every feasible face of $\Pi(a, b; \Omega)$ has pairwise two-hop column overlap. In fact, every feasible plan satisfies

$$P_{hj} \geq \delta_j \qquad \textit{for every } j.$$

Consequently the derivative kernel has the uniform common-component bound

$$M_{\alpha\beta} \geq \left(\min_p \frac{\delta_p}{b_p} \right) \frac{\delta_\beta}{a_h} \qquad \textit{for every } \alpha, \beta,$$

and therefore

$$\tau(M) \leq 1 - \gamma_h, \qquad \gamma_h = \left(\min_p \frac{\delta_p}{b_p} \right) \frac{\sum_\beta \delta_\beta}{a_h}.$$

This is a score-uniform one-step global column-quotient certificate.

Proof. For any feasible plan P, the non-bus rows have total mass $T - a_h$. Thus column j can receive at most $T - a_h$ mass from non-bus rows, and hence

$$P_{hj} = b_j - \sum_{i \neq h} P_{ij} \geq b_j - (T - a_h) = \delta_j > 0.$$

Therefore the bus row is positive to every column in every feasible plan and, by the coordinate-face structure of transportation polytopes, in every relative face support. Every pair of columns then shares row h, so Theorem A.5 applies.

The explicit lower bound follows from the support-to-kernel formula:

$$M_{\alpha\beta} = \sum_i \frac{P_{i\alpha} P_{i\beta}}{b_\alpha a_i} \geq \frac{P_{h\alpha} P_{h\beta}}{b_\alpha a_h} \geq \frac{\delta_\alpha \delta_\beta}{b_\alpha a_h} \geq \left(\min_p \frac{\delta_p}{b_p} \right) \frac{\delta_\beta}{a_h}.$$

Thus every row of M dominates the same subprobability vector $\ell_\beta = (\min_p \delta_p / b_p) \delta_\beta / a_h$ of total mass γ_h. The sparse-support overlap certificate gives $\tau(M) \leq 1 - \gamma_h$. □

The condition in Corollary A.6 is deliberately strong but checkable. It is not necessary for topology-forced mixing; it is a simple architectural sufficient condition. If $a_h + b_j \leq T$ for some column, the bus row can avoid that column in principle, and the corollary withholds rather than guessing. Multiple lighter bus rows can be handled by Corollary A.5, but only when the forced edges cover every column pair through at least one common bus row.

Architecture corollaries in appendix. The core face-lattice theorem is independent of the concrete architecture families used as examples. Equal-score spectral rules, local-band spectra, and other graph-design corollaries are collected in Appendix D. They are useful design consequences, but they are not assumptions in the support-to-kernel or face-lattice proofs.

B Dobrushin and Tail-Cotangent Certificates

The following facts are standard Dobrushin/Doeblin tools (Gaubert & Qu, 2013; Levin & Peres, 2017), stated here with the quotient convention of Definition A.1. They are the bridge from the fixed-support half-step calculus to rigorous tail-gradient bounds.

Definition B.1 (Projected source-cotangent interface)**.** *Fix the common column quotient $Q = \mathbb{R}^n / \mathrm{span}\{\mathbf{1}\}$ for the tail and let*

$$\iota : Q \longrightarrow \mathbf{1}^\perp, \qquad \iota([v]) = \Pi_0 v, \qquad \Pi_0 = I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top$$

be its unique centered lift. At step t, let $\mathcal{Z}_t$ collect the other independent local coordinates (parameters, values, residual state, or features), and let

$$\Phi_t : \mathbb{R}^n \times \mathcal{Z}_t \longrightarrow \mathcal{S}_t$$

collect every differentiable surrounding map that produces a source variable for the Sinkhorn cycle, evaluated at the centered input potential and fixed local coordinate z_t. The source space may contain active-edge scores, logarithmic marginals, values, query/key/value features, residual inputs, or loss-side quantities. Let $\nu_t \in \mathcal{S}_t^*$ be the aggregate source-space cotangent after all local reverse branches have been summed. Its score and marginal coordinates are the covectors in Proposition A.2, represented in the marginal cotangent quotients specified there.

The exact cotangent returned to the input quotient is the partial VJP of the composite map $[v] \mapsto \Phi_t(\iota([v]), z_t)$:

$$\begin{aligned} \xi_t &= D([v] \mapsto \Phi_t(\iota([v]), z_t))\,([v_t])^* \nu_t \\ &= \iota^* D_1 \Phi_t(\iota([v_t]), z_t)^* \nu_t = \Pi_0 \psi_t \in Q^*, \\ \psi_t &:= D_1 \Phi_t(\iota([v_t]), z_t)^* \nu_t. \end{aligned}$$

The final identity uses Euclidean coordinates on the centered section. Thus projection is not an assumed contraction and not another Sinkhorn transpose step: it is the adjoint of the chosen quotient lift. If Φ_t is already gauge-invariant, then $\mathbf{1}^\top \psi_t = 0$ and the projection changes nothing. For a

fixed disconnected support, replace ι and Π_0 by the component-centered lift and projection from Definition A.1. A source branch that does not descend to the declared quotient must first be expressed on a specified section; its section VJP, rather than an arbitrary raw covector, defines ξ_t. Source coordinates independent of the potential have zero $D_1\Phi_t$ contribution to ξ_t; their $D_2\Phi_t^*\nu_t$ cotangents continue on their own upstream paths. The aggregate ν_t contains each direct local source branch once and excludes the homogeneous Sinkhorn branch $M_t^\top \eta_{t+1}$, so the recurrence below neither drops nor duplicates a branch.

Theorem B.1 (Zero-mass quotient contraction)**.** *Let A be any row-stochastic kernel on n states. For every cotangent $\eta \in Q^* = \{\eta : \mathbf{1}^\top \eta = 0\}$,*

$$\|A^\top \eta\|_{\mathrm{TV}} \le \tau(A)\|\eta\|_{\mathrm{TV}}.$$

For row-stochastic kernels $A_1, \dots, A_K$,

$$\|(A_K \cdots A_1)^\top \eta\|_{\mathrm{TV}} \le \tau(A_K \cdots A_1)\|\eta\|_{\mathrm{TV}} \le \prod_{t=1}^{K} \tau(A_t)\|\eta\|_{\mathrm{TV}}.$$

Proof. A zero-mass signed vector η has Jordan decomposition $\eta = \eta^+ - \eta^-$ with $\eta^\pm \ge 0$ and $\eta^+(\mathbf{1}) = \eta^-(\mathbf{1}) = \|\eta\|_{\mathrm{TV}}$. After normalizing the two positive parts to probability vectors, $A^\top \eta$ is the difference of the two probability measures pushed forward by the Markov kernel A, times $\|\eta\|_{\mathrm{TV}}$. The Dobrushin coefficient is exactly the largest possible total-variation distance between two rows, equivalently the largest contraction factor on differences of probability measures. This proves the one-step inequality. The product kernel is row-stochastic, so the first inequality applies to it. Submultiplicativity of Dobrushin coefficients gives $\tau(A_K \cdots A_1) \le \prod_t \tau(A_t)$. □

Theorem B.2 (Projected-source tail-cotangent bound)**.** *Let $M_1, \dots, M_K \in \mathbb{R}^{n\times n}$ be the fixed-support quotient operators produced by a sequence of Sinkhorn row-column cycles on a common column state space, and let projected quotient sources $\xi_t \in Q^*$ collect the zero-mass cotangents delivered to the input quotient coordinates at step t. Concretely, the Sinkhorn score and marginal components are the adjoints in Proposition A.2; all local source-space covectors are summed into ν_t, and the exact partial VJP of $[v] \mapsto \Phi_t(\iota([v]), z_t)$ in Definition B.1 constructs ξ_t. Suppose the reverse tail recurrence is*

$$\eta_t = M_t^\top \eta_{t+1} + \xi_t, \qquad t = 1, \dots, K, \tag{B.1}$$

with terminal quotient cotangent $\eta_{K+1} \in Q^*$. Then

$$\eta_1 = (M_K \cdots M_1)^\top \eta_{K+1} + \sum_{t=1}^{K} (M_{t-1} \cdots M_1)^\top \xi_t, \tag{B.2}$$

where the empty product is the identity. Consequently the homogeneous residual obeys

$$\|(M_K \cdots M_1)^\top \eta\|_{\mathrm{TV}} \le \tau(M_K \cdots M_1)\|\eta\|_{\mathrm{TV}}, \tag{B.3}$$

and the inhomogeneous tail obeys the path/product sum bound

$$\|\eta_1\|_{\mathrm{TV}} \le \tau(M_K \cdots M_1)\|\eta_{K+1}\|_{\mathrm{TV}} + \sum_{t=1}^{K} \tau(M_{t-1} \cdots M_1)\|\xi_t\|_{\mathrm{TV}}. \tag{B.4}$$

A looser but sometimes cheaper certificate replaces each product coefficient by $\prod_{\ell=1}^{t-1} \tau(M_\ell)$.

Proof. Unroll (B.1). For $K = 2$, $\eta_1 = M_1^\top M_2^\top \eta_3 + M_1^\top \xi_2 + \xi_1$; the general formula (B.2) follows by induction, and $M_1^\top \cdots M_K^\top = (M_K \cdots M_1)^\top$. All terms remain in Q^* because the M_t are row-stochastic and the sources are quotient VJPs. Apply Theorem B.1 to the homogeneous product and to each source product, then use the triangle inequality. Submultiplicativity gives the optional product-of-one-step-coefficients bound. □

Corollary B.1 (Implicit quotient VJP certificate)**.** *Let M be a row-stochastic fixed-support quotient operator with $\tau(M) < 1$, and let $\xi \in Q^*$ be a projected source cotangent. The implicit quotient equation*

$$\eta = M^\top \eta + \xi$$

has a unique solution in Q^*,

$$\eta = \sum_{k=0}^{\infty} (M^\top)^k \xi,$$

and

$$\|\eta\|_{\mathrm{TV}} \le \frac{\|\xi\|_{\mathrm{TV}}}{1-\tau(M)}.$$

If $\eta^{(K)} = \sum_{k=0}^{K-1} (M^\top)^k \xi$, then the truncation residual is bounded by

$$\|\eta - \eta^{(K)}\|_{\mathrm{TV}} \le \frac{\tau(M^K)}{1-\tau(M)} \|\xi\|_{\mathrm{TV}} \le \frac{\tau(M)^K}{1-\tau(M)} \|\xi\|_{\mathrm{TV}}.$$

Proof. On Q^*, Theorem B.1 gives $\|(M^\top)^k \xi\|_{\mathrm{TV}} \le \tau(M)^k \|\xi\|_{\mathrm{TV}}$. The Neumann series therefore converges absolutely in the finite-dimensional quotient space and solves $(I - M^\top)\eta = \xi$. If another solution existed, their difference $z \in Q^*$ would satisfy $z = M^\top z$, hence $\|z\|_{\mathrm{TV}} \le \tau(M)\|z\|_{\mathrm{TV}}$; since $\tau(M) < 1$, $z = 0$. The norm bound is the geometric series. For the truncation residual, $\eta - \eta^{(K)} = (M^K)^\top \eta$, so Theorem B.1 gives $\|\eta - \eta^{(K)}\|_{\mathrm{TV}} \le \tau(M^K)\|\eta\|_{\mathrm{TV}}$, and the previous bound on $\|\eta\|_{\mathrm{TV}}$ completes the proof. Submultiplicativity gives the final inequality. □

Scope of the tail theorem. Theorem B.2 proves the fixed-support quotient-transport component exactly. A full neural-gradient statement also has to bound $\|\xi_t\|_{\mathrm{TV}}$ and the operator norms of any non-transport maps between quotient spaces. Proposition A.2 supplies the one-cycle source VJPs; this theorem supplies the quotient transport of the resulting projected sources. Without the surrounding source-norm and map-norm bounds, $\tau(M_K \cdots M_1)$ is a rigorous transport-tail certificate, not a complete parameter-to-loss gradient guarantee.

Proposition B.1 (Sparse-support overlap certificate)**.** *Let M be the derivative kernel induced by a fixed support and plan. Suppose there are $\gamma \in [0,1]$ and a probability vector q such that*

$$M_{ij} \ge \gamma q_j \qquad \text{for all } i, j.$$

Then $\tau(M) \le 1 - \gamma$.

Proof. Write each row as $M_i = \gamma q + (1-\gamma)\widetilde{M}_i$, where $\widetilde{M}_i$ is a probability vector when $\gamma < 1$. The common component cancels in $M_i - M_{i'}$, so the total-variation distance between any two rows is at most $1 - \gamma$. □

This proposition is deliberately stated as a kernel-level condition. A support graph can suggest overlap, but scores and marginal masses determine the actual lower bound after scaling. One can therefore compute the realized shared component

$$\gamma_{\mathrm{cert}} = \sum_j \min_i M_{ij}$$

from the realized kernel rather than claiming a graph-only theorem.

Theorem B.3 (Mass-floored hub certificate)**.** *Let $H \subseteq [n]$ be a set of hub columns. Suppose there is a number $p_{\min} > 0$ such that for every source column $\alpha \in [n]$ and every hub $h \in H$, there exists a witness row $i = i(\alpha, h)$ with*

$$P_{i\alpha} \ge p_{\min}, \qquad P_{ih} \ge p_{\min}.$$

Let $a_{\max} = \max_i a_i$ and $b_{\max} = \max_\alpha b_\alpha$. Then

$$M_{\alpha h} \ge \frac{p_{\min}^2}{b_{\max} a_{\max}} \qquad \text{for every } \alpha \in [n],\ h \in H.$$

Consequently, writing

$$\gamma_{\mathrm{shared}} = \sum_\beta \min_\alpha M_{\alpha\beta},$$

we have

$$\gamma_{\mathrm{shared}} \ge |H| \frac{p_{\min}^2}{b_{\max} a_{\max}}.$$

Equivalently, with the defensive cap

$$\gamma_H = \min\left\{1, |H|\frac{p_{\min}^2}{b_{\max}a_{\max}}\right\},$$

the derivative kernel has a common component of mass at least γ_H and $\tau(M) \le 1 - \gamma_H$.

Proof. Using Theorem A.1, for each α and h,

$$M_{\alpha h} = \sum_i \frac{P_{i\alpha}P_{ih}}{b_\alpha a_i}.$$

All summands are nonnegative. The witness row $i(\alpha, h)$ contributes at least

$$\frac{p_{\min}^2}{b_\alpha a_{i(\alpha,h)}} \ge \frac{p_{\min}^2}{b_{\max}a_{\max}}.$$

Therefore each row of M assigns at least this much mass to every hub column in the fixed common set H. Summing the rowwise minimum over the distinct columns $h \in H$ gives the displayed lower bound on γ_{shared}; applying Proposition B.1 gives the Dobrushin bound. Under exact hypotheses the uncapped lower bound is automatically at most one because M is row-stochastic; the cap is included only to make numerical reporting robust. □

This theorem is the safe quantitative form of the topology story. The support graph supplies the hub-witness pattern; the scaled plan supplies the mass floor. It is therefore a realized-plan certificate, not a pure support theorem. Without a mass floor or explicit score/marginal assumptions implying one, dense support alone does not prevent $\tau(M)$ from approaching one.

C Negative and Windowed Topology Certificates

The certificate is most useful when it distinguishes topology failures from topology fixes. The following negative examples are deliberately simple.

Proposition C.1 (Closed disconnected components)**.** *If the row-stochastic kernel M has two closed components with disjoint transition supports, then $\tau(M) = 1$. Hence the one-step certificate gives no global quotient contraction.*

Proof. Choose one row from each closed component. Their row distributions have zero overlap, so the total-variation distance between the rows is one. This is the maximal value in the Dobrushin coefficient. □

Proposition C.2 (Permutation-like supports)**.** *If M is a permutation matrix on at least two states, or more generally if two rows of M are point masses on different states, then $\tau(M) = 1$. Identity-like transport therefore preserves some quotient cotangent modes exactly.*

Proof. Two distinct point masses have total-variation distance one. A permutation matrix on at least two states has two rows that are point masses on different states. For the identity matrix on at least two states, every zero-mass cotangent is propagated without decay. □

One-step failure is not the same as architectural failure. Alternating supports or repeated layers can mix over a window.

Theorem C.1 (Windowed face-lattice characterization)**.** *For $t = 1, \dots, K$, let $\mathcal{P}_t = \Pi(a_t, b_t; \Omega_t)$ be nonempty compact transport polytopes on the same column index set, with positive compatible marginals and strict feasibility on every active edge of Ω_t. For $P_t \in \mathcal{P}_t$, define*

$$M_t(P_t) = \big(P_t \operatorname{diag}(b_t)^{-1}\big)^\top \big(\operatorname{diag}(a_t)^{-1}P_t\big).$$

For a face tuple $F = (F_1, \dots, F_K)$ with every F_t nonempty, let $\mathcal{R}_t(F_t)$ be the two-hop column relation induced by the face support Ω_{F_t}:

$$\alpha\, \mathcal{R}_t(F_t)\, \gamma \quad \Longleftrightarrow \quad \exists i \text{ with } (i,\alpha), (i,\gamma) \in \Omega_{F_t}.$$

Say that the tuple has windowed pairwise product overlap if, for every pair of columns α, β, the K-step reachable sets from α and β through the relation product, applied in matrix-product order,

$$\mathcal{R}_K(F_K); \mathcal{R}_{K-1}(F_{K-1}); \cdots ; \mathcal{R}_1(F_1)$$

intersect. Then the following are equivalent.

1. *There exists $\delta_K > 0$ such that every finite score sequence on the fixed supports satisfies*
$$\tau(M_K(P_K) \cdots M_1(P_1)) \leq 1 - \delta_K.$$
2. *Every face tuple $(F_1, \ldots, F_K)$ with every F_t nonempty has windowed pairwise product overlap.*

Thus a scheduled sparse architecture can force global quotient mixing over a window even when each one-step topology withholds a contraction certificate.

Proof. For any feasible tuple $(P_1, \ldots, P_K)$, the support of $M_K(P_K) \cdots M_1(P_1)$ is the Boolean product obtained by first following the support relation of $M_K(P_K)$, then the support relation of $M_{K-1}(P_{K-1})$, and so on through $M_1(P_1)$. By the support-to-kernel theorem, the one-step support relation of $M_t(P_t)$ is the two-hop relation induced by $\mathrm{supp}(P_t)$. Hence the product kernel has Dobrushin coefficient strictly below one exactly when every pair of its rows has overlapping support, which is precisely the matrix-product-order windowed product-overlap condition for the minimal face tuple containing the plans.

If every face tuple has windowed product overlap, then

$$\Gamma_K(P_1, \ldots, P_K) = 1 - \tau(M_K(P_K) \cdots M_1(P_1))$$

is positive on the compact product polytope $\mathcal{P}_1 \times \cdots \times \mathcal{P}_K$. Continuity gives a positive minimum δ_K, and every finite-score sequence satisfies the stated bound.

Conversely, suppose a face tuple $(F_1, \ldots, F_K)$ fails windowed product overlap. Each face F_t is exposed by some score direction B_t. Applying Theorem A.4 independently to the score sequence LB_t gives plans $P_t(L)$ converging to entropy-center points in $\mathrm{relint}(F_t)$. The corresponding product kernels converge to a row-stochastic kernel with two disjoint product-row supports, so its Dobrushin coefficient is one. By continuity,

$$\tau(M_K(P_K(L)) \cdots M_1(P_1(L))) \to 1,$$

and no positive score-uniform window bound can hold. □

As in the one-step face-lattice theorem, this is a structural characterization rather than a scalable checker. Exact verification over arbitrary supports requires quantifying over face tuples of a product of transport polytopes. The point is the architectural dichotomy: a scheduled support family either rules out bad zero-temperature face sequences, or some finite score sequence can make the window certificate arbitrarily weak. The product-coordinate theorem below is a tractable family where the good side of the dichotomy can be proved without enumerating faces.

Theorem C.2 (Windowed Dobrushin certificate)**.** *Let $M_1, \ldots, M_K$ be row-stochastic kernels. For every zero-mass cotangent η,*

$$\|(M_K \cdots M_1)^\top \eta\|_{\mathrm{TV}} \leq \tau(M_K \cdots M_1)\|\eta\|_{\mathrm{TV}}.$$

It is possible that $\tau(M_t) = 1$ for every t while $\tau(M_K \cdots M_1) < 1$.

Proof. The first claim is the zero-mass quotient contraction applied to the product kernel, which is row-stochastic. For the second claim, take two partition kernels on four states: one averages within $\{1, 2\}$ and $\{3, 4\}$, and the next averages within $\{1, 3\}$ and $\{2, 4\}$. Each kernel has disjoint row pairs and $\tau = 1$, but their product sends every row to the uniform distribution, so the two-step coefficient is zero. □

Theorem C.3 (Partition heat-bath transport layers)**.** *Let X be a finite column set with a positive probability vector b. Let $\pi : X \to Z$ be a partition map, use rows indexed by $z \in Z$, and define the active support by*

$$(z, x) \in \Omega \quad \Longleftrightarrow \quad \pi(x) = z.$$

Set the row marginal to the block mass

$$a_z = \sum_{x:\, \pi(x)=z} b_x.$$

Then the transportation polytope $\Pi(a, b; \Omega)$ is a singleton. For every finite score matrix on this active support, the Sinkhorn plan is

$$P_{zx} = b_x \, \mathbf{1}\{\pi(x) = z\}.$$

The induced column-to-column derivative kernel is the heat-bath kernel

$$M_{xy} = \mathbf{1}\{\pi(x) = \pi(y)\} \, \frac{b_y}{a_{\pi(x)}}.$$

Consequently any sequence of partition-fiber transport layers has exactly the same quotient-kernel product as the corresponding sequence of heat-bath resampling kernels. Windowed Dobrushin certificates therefore apply directly to the resulting kernel products.

Proof. Each column x is incident to exactly one row, namely $\pi(x)$. The column constraint therefore forces $P_{\pi(x),x} = b_x$, and all other entries in that column are inactive. The row sums match by the definition of a_z, so this is the unique feasible plan; because the feasible set is a singleton, finite scores cannot change the plan. Substituting this plan into the support-to-kernel formula gives

$$M_{xy} = \sum_z \frac{P_{zx} P_{zy}}{b_x a_z} = \mathbf{1}\{\pi(x) = \pi(y)\} \frac{b_y}{a_{\pi(x)}}.$$

The statement about products follows by applying the same identity to every layer. □

Theorem C.3 is deliberately stated as a kernel identification theorem, not as a universal mixing theorem. For non-product b, a sweep of such layers is a systematic block-Gibbs kernel for b; it may contract, but the coefficient depends on additional minorization or reachability properties of the block conditionals and can be arbitrarily close to one under strong dependencies. The stronger identity $M_{\sigma(K)} \cdots M_{\sigma(1)} = \mathbf{1} b^\top$ in the next theorem uses the product form of b in an essential way.

Theorem C.4 (Product-coordinate window certificate)**.** *Let*

$$X = X_1 \times \cdots \times X_K$$

be a finite product column set. Let $b(x) = \prod_{t=1}^K \mu_t(x_t)$ be a strictly positive product marginal. For layer t, let rows be indexed by fibers $z \in X_{-t} := \prod_{s \neq t} X_s$, connect row z exactly to columns $x \in X$ with $x_{-t} = z$, and set

$$a_t(z) = \sum_{u \in X_t} b(z, u) = \prod_{s \neq t} \mu_s(z_s).$$

Then the feasible transport plan on layer t is unique, for every finite score matrix on the active support:

$$P_t(z, x) = b(x) \, \mathbf{1}\{x_{-t} = z\}.$$

Consequently its derivative kernel is the coordinate-resampling kernel

$$(M_t)_{xy} = \mathbf{1}\{x_{-t} = y_{-t}\} \, \mu_t(y_t).$$

For any permutation σ of $\{1, \ldots, K\}$, applying each coordinate layer once gives

$$M_{\sigma(K)} \cdots M_{\sigma(1)} = \mathbf{1} b^\top, \qquad \tau(M_{\sigma(K)} \cdots M_{\sigma(1)}) = 0.$$

If the product of the coordinates other than t has at least two elements, then $\tau(M_t) = 1$. Thus coordinate-factorized sparse layers may withhold every one-step contraction certificate while a full coordinate window gives perfect score-uniform quotient mixing.

Proof. Fix t. Each column x is incident to exactly one row, namely $z = x_{-t}$. The column marginal constraint therefore forces $P_t(z, x) = b(x)$ on that active edge. The row sums match by the definition of $a_t(z)$:

$$\sum_{x:\, x_{-t}=z} P_t(z, x) = \sum_{u \in X_t} b(z, u) = a_t(z).$$

Thus the transportation polytope for layer t is a singleton and finite scores cannot change the plan.

Using the support-to-kernel identity,

$$(M_t)_{xy} = \sum_z \frac{P_t(z, x) P_t(z, y)}{b(x) a_t(z)}.$$

There is a nonzero summand only when $x_{-t} = y_{-t} = z$. In that case

$$(M_t)_{xy} = \frac{b(x) b(y)}{b(x) a_t(x_{-t})} = \frac{b(y)}{a_t(x_{-t})} = \mu_t(y_t),$$

which proves the coordinate-resampling formula.

The kernel M_t leaves all coordinates except t fixed and resamples coordinate t from μ_t. These coordinate-resampling kernels commute because b is a product measure and each kernel acts on a different coordinate conditional. After every coordinate has been applied once, the output distribution is b independently of the input column, so the product kernel is $\mathbf{1} b^\top$ and its Dobrushin coefficient is zero.

If X_{-t} has at least two elements, choose columns x, y with $x_{-t} \neq y_{-t}$. The supports of rows x and y of M_t lie in disjoint fibers, so the two rows have total-variation distance one and $\tau(M_t) = 1$. □

Corollary C.1 (Hypercube butterfly window)**.** *Let $n = 2^r$ with $r \geq 2$, and index columns by bitstrings $x \in \{0,1\}^r$. For layer $t \in \{1, \ldots, r\}$, let the fixed support have one row for each unordered pair $\{x, x \oplus e_t\}$, where e_t flips bit t, and connect that row to exactly the two columns in the pair. Use equal active scores, row marginals $a_i = 2/n$ on the $n/2$ matching rows, and column marginals $b_x = 1/n$. Let M_t be the resulting column-to-column derivative kernel. Then*

$$M_t = \frac{1}{2}(I + F_t),$$

where F_t is the permutation matrix for bit flip $x \mapsto x \oplus e_t$. Consequently $\tau(M_t) = 1$ for every t, but

$$M_r M_{r-1} \cdots M_1 = \frac{1}{n} \mathbf{1}\mathbf{1}^\top \qquad \textit{and hence} \qquad \tau(M_r M_{r-1} \cdots M_1) = 0.$$

Proof. Apply Theorem C.4 with $X_t = \{0, 1\}$ and μ_t uniform. The fiber row indexed by x_{-t} is exactly the unordered pair $\{x, x \oplus e_t\}$, and the coordinate-resampling kernel is $M_t = (I + F_t)/2$. Since $r \geq 2$, each X_{-t} has at least two elements, so $\tau(M_t) = 1$. Applying all bit coordinates once gives the uniform kernel $n^{-1}\mathbf{1}\mathbf{1}^\top$, hence the window coefficient is zero. □

D Spectral Architecture Corollaries

Corollary D.1 (Spectral topology certificate)**.** *In the setting of Theorem A.2, M is symmetric and doubly stochastic. Let*

$$\lambda_\star = \|M \Pi_{\mathbf{1}^\perp}\|_2 = \max\{|\lambda| : \lambda \textit{ is an eigenvalue of } M \textit{ on } \mathbf{1}^\perp\}.$$

Then, for every zero-mass cotangent η and every integer $K \geq 1$,

$$\|(M^K)^\top \eta\|_2 \leq \lambda_\star^K \|\eta\|_2, \qquad \|(M^K)^\top \eta\|_{\mathrm{TV}} \leq \sqrt{n/2}\, \lambda_\star^K \|\eta\|_{\mathrm{TV}}.$$

Moreover,

$$\tau(M^K) \leq \min\{1, \sqrt{n/2}\, \lambda_\star^K\}.$$

Proof. The formula $M = N/(d_c d_r)$ has $N = N^\top$ and row sums $d_c d_r$, hence M is symmetric and doubly stochastic. The Euclidean bound follows by diagonalizing M on the invariant subspace $\mathbf{1}^\perp$. For total variation, use $\|x\|_{\mathrm{TV}} \le \sqrt{n}\|x\|_2/2$ and $\|\eta\|_2 \le \sqrt{2}\|\eta\|_{\mathrm{TV}}$ for zero-mass η. For Dobrushin's coefficient, apply the same Euclidean contraction to the zero-mass row difference $e_\alpha - e_\beta$, whose Euclidean norm is $\sqrt{2}$, then maximize over row pairs and cap by the trivial bound $\tau \le 1$. □

This corollary is the bridge from graph design to tail-depth design. In the equal-score biregular baseline, choosing a support is choosing a reversible Markov chain for quotient cotangents. Expander-like column-overlap walks make $\lambda_\star$ small, while butterfly-like supports may have weak one-step $\tau(M)$ but strong powered certificates for M^K.

Corollary D.2 (Expander tail-depth rule)**.** *In the setting of Theorem A.2, suppose the zero-mass spectral rate satisfies $\lambda_\star \le \rho < 1$. Then*

$$\tau(M^K) \le \min\{1, \sqrt{n/2}\rho^K\}.$$

If $\rho = 0$, any $K \ge 1$ suffices. If $0 < \rho < 1$, then for any $\varepsilon \in (0,1)$ it is sufficient to choose

$$K \ge \left\lceil \frac{\log(\sqrt{n/2}/\varepsilon)}{-\log\rho} \right\rceil$$

to obtain $\tau(M^K) \le \varepsilon$.

Proof. Substitute $\lambda_\star \le \rho$ in Corollary D.1 and solve $\sqrt{n/2}\rho^K \le \varepsilon$ for K. □

Corollary D.3 (Circular local-band spectrum)**.** *Consider the equal-score biregular baseline on n circular positions where row i is connected to columns $i-r, \dots, i+r$ modulo n. Assume $0 \le r < n/2$, so the local degree is $d = 2r+1$. Then the derivative kernel M is circulant and has Fourier eigenvalues*

$$\lambda_k = \frac{1}{d^2}\left|\sum_{s=-r}^{r} \exp(2\pi \mathrm{i} k s/n)\right|^2 = \begin{cases} 1, & k = 0, \\ \left(\dfrac{\sin(\pi k d/n)}{d\sin(\pi k/n)}\right)^2, & k = 1, \dots, n-1. \end{cases}$$

Consequently the zero-mass spectral rate is

$$\lambda_{\mathrm{band}} = \max_{1 \le k \le n-1} \left(\frac{\sin(\pi k d/n)}{d\sin(\pi k/n)}\right)^2,$$

and Corollary D.1 gives

$$\tau(M^K) \le \min\{1, \sqrt{n/2}\,\lambda_{\mathrm{band}}^K\}.$$

For fixed r and $n \to \infty$,

$$\lambda_{\mathrm{band}} \ge \lambda_1 = 1 - \frac{\pi^2(d^2-1)}{3n^2} + O(n^{-4}),$$

so this spectral certificate is diffusive for fixed-radius local bands. More sharply, for the first nonconstant Fourier cotangent $\eta^{(1)}$,

$$\frac{\|(M^K)^\top \eta^{(1)}\|_2}{\|\eta^{(1)}\|_2} = \lambda_1^K.$$

Thus, if r is fixed and $K = o(n^2)$, then $\lambda_1^K \to 1$. No fixed-radius circular local band can uniformly contract all zero-mass quotient modes in subquadratic depth under this equal-score spectral model.

Proof. Let $h \in \mathbb{R}^n$ be the circular band indicator, $h_s = 1$ for $s \in \{-r, \dots, r\}$ modulo n and $h_s = 0$ otherwise. By Theorem A.2, $M_{\alpha\beta}$ is d^{-2} times the number of rows incident to both columns α and β. Hence M is the circulant cyclic autocorrelation

$$M_{\alpha\beta} = \frac{1}{d^2}\sum_{s \in \mathbb{Z}/n\mathbb{Z}} h_s h_{s+\beta-\alpha}.$$

The discrete Fourier characters diagonalize circulant convolution operators, and the autocorrelation has Fourier transform $|\widehat{h}_k|^2$, where

$$\widehat{h}_k = \sum_{s=-r}^{r} \exp(2\pi \mathrm{i} k s/n).$$

This gives the displayed eigenvalues. For $k \neq 0$, the finite geometric sum is the Dirichlet kernel

$$\widehat{h}_k = \frac{\sin(\pi k d/n)}{\sin(\pi k/n)}$$

up to a phase, whose absolute value gives the second expression. The zero-mass subspace removes the $k = 0$ constant eigenvector, and the spectral Dobrushin bound is Corollary D.1.

Finally, put $x = \pi/n$. For fixed d,

$$\frac{\sin(dx)}{d \sin x} = 1 - \frac{(d^2-1)x^2}{6} + O(x^4),$$

and squaring gives the expansion for λ_1. Since $\lambda_{\mathrm{band}} \geq \lambda_1$, the stated lower asymptotic follows. The first nonconstant Fourier character is an eigenvector with eigenvalue λ_1, so the displayed L^2 ratio after K applications is exact. If $K = o(n^2)$ and r is fixed, the expansion gives $K(1-\lambda_1) \to 0$, hence $\lambda_1^K \to 1$. □

Proposition D.1 (Circular local-band one-step obstruction)**.** *Consider the equal-score biregular baseline on n circular positions where row i is connected to columns $i-r, \ldots, i+r$ modulo n, with $r \geq 0$. If $n \geq 8r+2$, then $\tau(M) = 1$.*

Proof. The support of row α of the derivative kernel M consists exactly of columns within circular distance at most $2r$ from α: to move from column α to column β in one full derivative step, the two columns must share a support row, and this happens precisely when their circular distance is at most $2r$. If $n \geq 8r+2$, there are two columns whose circular distance is at least $4r+1$. Their radius-$2r$ derivative neighborhoods are disjoint. The corresponding rows of M therefore have disjoint supports, so their total-variation distance is one and $\tau(M) = 1$. □

E Registers, Dustbins, and Shared Buses

Global tokens, register tokens, dustbin rows or columns, and expert buses can all create a common component in the derivative kernel. The architectural reading is simple: if every local region sends some reverse-mode mass through the same bus, zero-mass cotangents can no longer remain perfectly isolated.

Proposition E.1 (Register/dustbin mixing certificate)**.** *Consider a fixed support in which a register or dustbin substructure induces a realized derivative kernel M. If the realized rows of M share common mass*

$$\gamma_{\mathrm{cert}} = \sum_j \min_i M_{ij} > 0,$$

then

$$\|M^\top \eta\|_{\mathrm{TV}} \leq (1 - \gamma_{\mathrm{cert}}) \|\eta\|_{\mathrm{TV}}$$

for every zero-mass cotangent η.

Proof. This is the sparse-support overlap certificate with $q_j = \min_i M_{ij}/\gamma_{\mathrm{cert}}$. The register or dustbin interpretation is architectural; the theorem only uses the realized common component. □

Theorem E.1 (Equal-score register-bus certificate)**.** *Let $n \geq 1$, $r \geq 1$, and let the token-token support on the n token rows and columns be d-regular with $d \geq 1$: every token row and every token column has exactly d active token-token edges. Add $r \geq 1$ register rows and r register columns. Token rows connect to their active token columns and to every register column; register rows connect to every token and register column. Let $N = n + r$, use equal active scores, and set uniform marginals $a = b = N^{-1}\mathbf{1}$.*

Define $t > 0$ by

$$rt^2 + (n-r)t - d = 0, \qquad t = \frac{2d}{\sqrt{(n-r)^2 + 4rd} + (n-r)}.$$

Then the Sinkhorn plan has four values:

$$P_{ij} = \begin{cases} A, & i, j \text{ token and } (i,j) \text{ active}, \\ B, & \text{one endpoint token and one endpoint register}, \\ D, & i, j \text{ both registers}, \end{cases}$$

where

$$A = \frac{1}{N(d+rt)}, \qquad B = tA, \qquad D = t^2 A.$$

For the derivative kernel $M = C^\top R$, every source column α and every register target column h satisfy

$$M_{\alpha h} \geq m_{\mathrm{reg}} := \frac{t(d+rt^2)}{(d+rt)^2}.$$

Consequently the r register columns form a common component of mass at least

$$\gamma_{\mathrm{reg}} = r\, m_{\mathrm{reg}} = \frac{rt(d+rt^2)}{(d+rt)^2},$$

and

$$\tau(M) \leq 1 - \gamma_{\mathrm{reg}}.$$

Proof. The proposed plan has Sinkhorn scaling form for equal scores: token row and column scales are constant within the token class, and register row and column scales are constant within the register class. Its token row sum is

$$dA + rB = A(d+rt) = \frac{1}{N}.$$

The token column sum is identical because the token-token support is d-regular and B is used for register-token edges. The register row and register column sums are

$$nB + rD = A(nt + rt^2).$$

The defining quadratic for t is equivalent to $nt + rt^2 = d + rt$, so these sums also equal $1/N$. Thus the candidate is a strictly positive feasible scaling of the equal-score support. By uniqueness of the entropic Sinkhorn plan on the fixed connected support, it is the Sinkhorn plan.

Since $a = b = N^{-1}\mathbf{1}$,

$$M_{\alpha\beta} = N^2 \sum_i P_{i\alpha} P_{i\beta}.$$

If α is a token column and h is a register column, the contributing rows are the d token rows incident to α and the r register rows:

$$M_{\alpha h} = N^2(dAB + rBD) = \frac{t(d+rt^2)}{(d+rt)^2}.$$

If α is itself a register column, then

$$M_{\alpha h} = N^2(nB^2 + rD^2) = \frac{t^2(n+rt^2)}{(d+rt)^2}.$$

Using $d = rt^2 + (n-r)t$, the difference between the register-source value and the token-source value is

$$\frac{rt^2(t-1)^2}{(d+rt)^2} \geq 0.$$

Thus every row of M assigns at least m_{reg} to each register target column. The sparse-support overlap certificate applied to the common mass on the r register columns gives $\tau(M) \leq 1 - \gamma_{\mathrm{reg}}$. □

Equal-score limitation. Theorem E.1 is a pure topology theorem because equal scores and uniform marginals force a four-value scaled plan. With structured scores or nonuniform marginals, the same support still has a register-bus interpretation, but quantitative contraction must be certified from the realized plan or from a separate distortion/mass-floor hypothesis. The degenerate case $d = 0$ is a pure register-mediated support and is not covered by the four-value formula above.

Design implication. A narrow local window may have $\tau(M) = 1$ because distant row distributions do not overlap. Adding a small register budget can reduce $\tau(M)$ if the scaled plan routes enough mass through the shared component. This is a certificate knob, not a proof of task accuracy.

F Certificate-Level Design Rules

The previous theorems describe what can be proved for the fixed-support transport derivative under the stated assumptions. They establish certificate logic rather than accuracy claims.

These results suggest a practical matched-edge workflow. First test an equal-score biregular baseline to understand the pure graph walk. Then compare the candidate support's realized $\tau(M^K)$ against simple graph descriptors and against a local-band control with the same edge budget. Finally, stress the candidate with nonuniform marginals and score perturbations; if the certificate survives, the topology becomes a candidate for a downstream task gate. If it does not survive, the failure is informative: the architecture may still be useful for restricted data, but the fixed-support transport derivative has no global quotient-mixing certificate under those conditions.

G Short DAG Path-Sum Remark

A residual or branching computation graph is outside the core chain theorem, but the same quotient bookkeeping applies once every edge map has an explicitly bounded operator norm on the chosen quotient representative. If a sink-to-source cotangent can be expanded as a finite path sum

$$\eta_{\mathrm{source}} = \sum_p w_p \left(\prod_{e \in p} M_e^\top \right) \eta_{\mathrm{sink}}$$

with zero-mass cotangents throughout, then Theorem B.1 and the triangle inequality give

$$\|\eta_{\mathrm{source}}\|_{\mathrm{TV}} \le \left(\sum_p |w_p| \prod_{e \in p} \tau(M_e) \right) \|\eta_{\mathrm{sink}}\|_{\mathrm{TV}}.$$

Any inhomogeneous source term must be projected and bounded separately, exactly as in Theorem B.2. This appendix remark is a bookkeeping consequence, not an additional architecture claim.