Finite-Step Sinkhorn Certificates for Nonnegative Stream-Mixing Products

Anonymous authors
Paper under double-blind review

Abstract

We develop deterministic finite-step certificates for products of nonnegative stream-mixing matrices in expanded-stream residual architectures, with Manifold-Constrained Hyper-Connections as the motivating example. The certified object is the mixer product

$$T_L = H_{\text{res}}^{(L)} H_{\text{res}}^{(L-1)} \cdots H_{\text{res}}^{(1)},$$

not the full nonlinear network. For entrywise nonnegative mixers, marginal errors are induced-norm certificates: row defects control forward streamwise $A_{\max}$ gain, while column defects control the backward gain. In the column-then-row Sinkhorn order, exact row sums give layerwise forward control, and the output column defect gives

$$\|T_L\|_1 \le \min\Big\{\prod_{\ell=1}^{L}(1+\varepsilon_c^{(\ell)}),\, n\Big\},$$

where n is a structural ceiling. A telescoping prefix-gain bound refines the product certificate, and a Dobrushin coefficient gives a depth-uniform certificate under uniform scrambling. We prove a Hilbert-projective finite-step bound with the half-step loss required for the output column error, converting logit dynamic range, temperature, and a Sinkhorn step cap into a backward mixer-path budget. The global Hilbert estimate is a certificate; near the Sinkhorn limit, the local rate $\sigma_2(P)^2$ is used as a predictor. Numerical diagnostics report marginal defects, product certificates, exact propagated gains, cap failures, and controller costs.

1 Introduction

Expanded-stream residual architectures route information through several parallel streams. At layer ℓ, the residual stream mixer $H_{\text{res}}^{(\ell)} \in \mathbb{R}^{n\times n}$ acts on the stream index, and the depth-wise mixing pathway is the product

$$T_L := H_{\text{res}}^{(L)} H_{\text{res}}^{(L-1)} \cdots H_{\text{res}}^{(1)}. \tag{1.1}$$

Large induced norms of T_L can amplify activations or backpropagated signals through this pathway, even when feature-scale normalization is present. We study this mixer product directly.

Row-stochasticity by itself is insufficient for backward control. For example, $H = \mathbf{1}\mathbf{e}_1^\top$ has $\|H\|_\infty = 1$ but $\|H\|_1 = n$. Thus a row-normalized mixer can preserve forward ℓ_∞ gain while amplifying backward signals at the stream-expansion scale. Finite-step Sinkhorn normalization targets both marginals, but a fixed iteration count alone does not state what depth-wise budget has been certified.

The motivating architecture class is Manifold-Constrained Hyper-Connections (mHC), which parameterizes nonnegative mixers and pushes them toward the Birkhoff polytope of doubly stochastic matrices. Exact projection is expensive, so the implementation-relevant object is a positive logit matrix followed by a finite number of Sinkhorn–Knopp normalizations (Sinkhorn and Knopp, 1967; Knight, 2008). Given a depth L, a backward mixer-path budget B_1, and a step cap $t_{\max}$, we ask what finite-step marginal tolerance suffices to certify $\|T_L\|_1 \le B_1$.

We give a deterministic finite-step certificate. Measured column defects give an induced-norm bound; a Hilbert-metric estimate connects logit dynamic range and Sinkhorn steps to that defect; and a single vector propagation computes the exact realized backward gain. In the ideal column-then-row order, every output has exact row sums, so the forward $A_{\max}$ mixer-path gain is controlled layerwise; the nontrivial design problem is the backward ℓ_1 gain.

Sinkhorn–Knopp scaling and matrix balancing are classical (Sinkhorn and Knopp, 1967; Franklin and Lorenz, 1989; Knight, 2008), and Hilbert's projective metric gives a standard contraction framework for positive maps (Birkhoff, 1957; Bushell, 1973). In machine learning, Sinkhorn iterations are used in entropic optimal transport and Sinkhorn distances (Cuturi, 2013; Peyré and Cuturi, 2019), and Sinkhorn-style normalization has been studied for attention (Sander et al., 2022; Shahbazi et al., 2025). Here Sinkhorn normalization is used to certify finite-step error budgets for products of nonnegative stream mixers. Norm control for residual dynamics is also related to stable deep architectures and continuous-depth models (Haber and Ruthotto, 2018; Chen et al., 2018); our certificates target the finite-step normalization error of the nonnegative mixing pathway.

Scope. The guarantees control the mixer product T_L in induced ℓ_∞ and ℓ_1 matrix norms, equivalently the forward and backward streamwise $A_{\max}$ gains of the mixing pathway. They are not end-to-end stability or accuracy guarantees for the full nonlinear network. Nonlinear residual branches require separate Jacobian control; the corresponding bookkeeping bounds are recorded in Appendix C.

The paper therefore treats Sinkhorn normalization as a finite-step certificate mechanism for the depth-wise nonnegative mixer product, rather than as a generic regularizer.

Contributions.

- **Mixer-product certificates.** For entrywise nonnegative mixers, row and column marginal errors directly control $\|H\|_\infty$ and $\|H\|_1$, giving depth-wise forward and backward $A_{\max}$ budgets for T_L (Theorem 3.1). We also record the exact realized backward-gain diagnostic by vector propagation (Remark 3.2).
- **Sharper product controls.** The backward gain of any product of nonnegative row-stochastic mixers is capped by the expansion rate, $\|T_L\|_1 \leq n$ (Proposition 3.3), independently of the Sinkhorn schedule. A telescoping prefix-gain bound refines the raw product certificate (Proposition 3.5), and a Dobrushin coefficient gives a depth-uniform certificate under uniform scrambling (Theorem 3.7).
- **Finite-step Sinkhorn calibration.** Applying Hilbert-metric contraction to the column-then-row normalization schedule, we derive a column-marginal estimate with the proper one-half-step loss at the output. This maps $(\kappa, t_{\max})$ to a certified tolerance and then to a depth-wise backward budget on $\|T_L\|_1$ (Theorems 4.7 and 4.15).
- **Diagnostics and local rates.** For a fixed positive matrix, the local full-step Sinkhorn rate is $\sigma_2(P)^2$, used as a predictor rather than a global certificate. The experiments track marginal defects, product certificates, exact propagated gains, cap failures, adaptive stopping, and controller costs.

Certificate workflow. Given a backward mixer-path budget B_1, the implementation allocates a column-marginal tolerance across layers and stops Sinkhorn iterations once the measured defect meets the assigned tolerance. It then reports the product certificate, the structural n-ceiling, and the exact propagated gain. The bookkeeping is summarized in Appendix A.

2 Preliminaries and notation

Notation summary

Symbol	Meaning
n	number of streams / expansion rate
C	feature width per stream
$H_{\text{res}}^{(\ell)}$	stream-mixing matrix at layer ℓ
T_L	depth-wise mixer product $H_{\text{res}}^{(L)} \cdots H_{\text{res}}^{(1)}$
$A = S^{(0)}$	positive pre-normalized matrix before Sinkhorn scaling
T_c, T_r	column and row normalization maps
$S^{(s)}$	matrix after s alternating normalization half-steps
t	number of full column-then-row Sinkhorn steps; output is $S^{(2t)}$
$\varepsilon_r, \varepsilon_c$	row and column marginal errors
κ	entrywise dynamic range $\max_{ij} A_{ij} / \min_{ij} A_{ij}$
$\Delta(A)$	projective diameter of A
$\tau(A)$	Birkhoff contraction factor $\tanh(\Delta(A)/4)$
q	dynamic-range bound $(\kappa - 1)/(\kappa + 1)$ for $\tau(A)$
P	limiting doubly stochastic Sinkhorn scaling of A
ρ_*	local full-step rate $\sigma_2(P)^2$
$A_{\max}$	streamwise coordinate maximum functional

2.1 Stream convention and $A_{\max}$ gains

Fix an expansion rate $n \geq 1$ and feature width $C \geq 1$. At depth ℓ, the state is a matrix $x_\ell \in \mathbb{R}^{n \times C}$ whose rows index streams and whose columns index features. A hyper-connected residual layer has the form

$$x_{\ell+1} = H_{\text{res}}^{(\ell)} x_\ell + (H_{\text{post}}^{(\ell)})^\top F\big(H_{\text{pre}}^{(\ell)} x_\ell, W_\ell\big), \tag{2.1}$$

where $H_{\text{res}}^{(\ell)} \in \mathbb{R}^{n \times n}$ mixes streams, $H_{\text{pre}}^{(\ell)}, H_{\text{post}}^{(\ell)} \in \mathbb{R}^{1 \times n}$ are pre/post aggregation row vectors, and $F(\cdot, W_\ell) : \mathbb{R}^{1 \times C} \to \mathbb{R}^{1 \times C}$ is a nonlinear feature block. The certificate below concerns the stream-mixer product generated by $H_{\text{res}}^{(\ell)}$.

For $M \in \mathbb{R}^{n \times n}$ we use the matrix norms induced by the vector ℓ_∞ and ℓ_1 norms:

$$\|M\|_\infty := \max_i \sum_j |M_{ij}|, \qquad \|M\|_1 := \max_j \sum_i |M_{ij}|. \tag{2.2}$$

Throughout, $\|\cdot\|_1$ and $\|\cdot\|_\infty$ refer to these matrix norms; for vector norms we write $\|\cdot\|_{1,\text{vec}}$ and $\|\cdot\|_{\infty,\text{vec}}$.

Definition 2.1 (Streamwise $A_{\max}$ functional)**.** For $X \in \mathbb{R}^{n \times C}$, define

$$A_{\max}(X) := \max_{1 \leq i \leq n} \|X_{i,:}\|_{\infty,\text{vec}} = \max_{i,c} |X_{ic}|. \tag{2.3}$$

For $H \in \mathbb{R}^{n \times n}$ and $X, G \in \mathbb{R}^{n \times C}$,

$$A_{\max}(HX) \leq \|H\|_\infty \, A_{\max}(X), \tag{2.4}$$

while backpropagation through the stream-mixing map uses the adjoint $G \mapsto H^\top G$ and satisfies

$$A_{\max}(H^\top G) \leq \|H\|_1 \, A_{\max}(G). \tag{2.5}$$

Thus forward streamwise $A_{\max}$ gain is governed by $\|H\|_\infty$, while the corresponding backward gain is governed by $\|H\|_1$. If $H \geq 0$, these norms are just the maximum row sum and maximum column sum.

Approximate doubly stochasticity. The Birkhoff polytope of doubly stochastic matrices is

$$\mathcal{B}_n := \left\{ P \in \mathbb{R}^{n\times n} : \; P_{ij} \geq 0, \; \sum_{j=1}^{n} P_{ij} = 1 \; \forall i, \; \sum_{i=1}^{n} P_{ij} = 1 \; \forall j \right\}. \tag{2.6}$$

Since the mixers use a finite number of Sinkhorn–Knopp steps, we quantify closeness to $\mathcal{B}_n$ by marginal errors.

Definition 2.2 (Approximate doubly stochasticity)**.** Let $H \in \mathbb{R}^{n\times n}$ be entrywise nonnegative. We say that H is $(\varepsilon_r, \varepsilon_c)$-doubly stochastic if

$$\max_i \left| \sum_j H_{ij} - 1 \right| \leq \varepsilon_r, \qquad \max_j \left| \sum_i H_{ij} - 1 \right| \leq \varepsilon_c. \tag{2.7}$$

3 Depth-wise $A_{\max}$ certificates

This section converts per-layer marginal errors into worst-case bounds on forward and backward $A_{\max}$ gains for the depth-wise mixer product. For entrywise nonnegative mixers, marginal control directly bounds the induced ℓ_∞ and ℓ_1 matrix norms, with no additional structural assumptions.

If $H \geq 0$ is $(\varepsilon_r, \varepsilon_c)$-doubly stochastic in the sense of Definition 2.2, then

$$\|H\|_\infty \leq 1 + \varepsilon_r, \qquad \|H\|_1 \leq 1 + \varepsilon_c, \tag{3.1}$$

since for nonnegative H the norms reduce to maximum row sum and maximum column sum.

Theorem 3.1 (Marginal-error depth certificate)**.** *Let $H_{\text{res}}^{(1)}, \ldots, H_{\text{res}}^{(L)} \in \mathbb{R}^{n\times n}$ be entrywise nonnegative and define T_L as in Equation (1.1). Assume each $H_{\text{res}}^{(\ell)}$ is $(\varepsilon_r^{(\ell)}, \varepsilon_c^{(\ell)})$-doubly stochastic. Then*

$$\|T_L\|_\infty \leq \prod_{\ell=1}^{L} (1 + \varepsilon_r^{(\ell)}), \qquad \|T_L\|_1 \leq \prod_{\ell=1}^{L} (1 + \varepsilon_c^{(\ell)}). \tag{3.2}$$

In particular, marginal error certificates propagate multiplicatively across depth into a depth-wise $A_{\max}$ budget for the mixer pathway.

Proof. By submultiplicativity of induced norms and (3.1), $\|T_L\|_\infty \leq \prod_{\ell=1}^{L} \left\| H_{\text{res}}^{(\ell)} \right\|_\infty \leq \prod_{\ell=1}^{L} (1 + \varepsilon_r^{(\ell)})$, and similarly for $\|\cdot\|_1$. □

Remark 3.2 (Exact backward gain). For realized nonnegative mixers $H^{(1)}, \ldots, H^{(L)}$, set $T_L = H^{(L)} \cdots H^{(1)}$. The exact backward mixing gain is obtained without forming T_L:

$$w_L = \mathbf{1}, \qquad w_{\ell-1} = (H^{(\ell)})^\top w_\ell, \qquad \ell = L, L-1, \ldots, 1. \tag{3.3}$$

Then

$$w_0 = T_L^\top \mathbf{1}, \qquad \|T_L\|_1 = \|w_0\|_{\infty,\text{vec}}. \tag{3.4}$$

Indeed, for a nonnegative matrix the induced ℓ_1 norm is the maximum column sum. This diagnostic is often much sharper than the layerwise product surrogate.

3.1 A structural ceiling and a telescoping refinement

The product certificate of Theorem 3.1 can grow with depth. For row-stochastic nonnegative mixers, however, the true backward gain is capped by the expansion rate. The next results record this ceiling and a telescoping refinement between the raw product certificate and the exact propagated gain.

Proposition 3.3 (Structural backward ceiling at the expansion rate)**.** *Let $H^{(1)}, \dots, H^{(L)} \in \mathbb{R}^{n\times n}$ be entrywise nonnegative and set $T_L = H^{(L)} \cdots H^{(1)}$.*

(i) If every factor has exact row sums, $H^{(\ell)}\mathbf{1} = \mathbf{1}$ (the ideal column–then–row output of Lemma 4.1), then T_L is nonnegative and row-stochastic, and

$$1 \le \|T_L\|_1 \le n. \tag{3.5}$$

(ii) More generally, if each $H^{(\ell)}$ is $(\varepsilon_r^{(\ell)}, \varepsilon_c^{(\ell)})$-doubly stochastic, then

$$\|T_L\|_1 \le n \prod_{\ell=1}^{L} (1 + \varepsilon_r^{(\ell)}). \tag{3.6}$$

Combining (3.6) with Theorem 3.1,

$$\|T_L\|_1 \le \min\left\{ \prod_{\ell=1}^{L} (1 + \varepsilon_c^{(\ell)}),\ n \prod_{\ell=1}^{L} (1 + \varepsilon_r^{(\ell)}) \right\}, \tag{3.7}$$

and in the ideal `column-then-row` *order $\|T_L\|_1 \le \min\{\prod_\ell (1 + \varepsilon_c^{(\ell)}),\, n\}$.*

Proof. A product of nonnegative row-stochastic matrices is nonnegative and row-stochastic, so $T_L\mathbf{1} = \mathbf{1}$ and the total mass is $\mathbf{1}^\top T_L \mathbf{1} = \mathbf{1}^\top \mathbf{1} = n$. For a nonnegative matrix $\|T_L\|_1$ equals the maximum column sum, which lies between the mean column sum $n/n = 1$ and the total mass n, giving (3.5). For (ii), nonnegativity together with $H^{(\ell)}\mathbf{1} \le (1+\varepsilon_r^{(\ell)})\mathbf{1}$ entrywise gives, by induction on the factors, $T_L\mathbf{1} \le \prod_\ell (1+\varepsilon_r^{(\ell)})\,\mathbf{1}$ entrywise; hence $\mathbf{1}^\top T_L \mathbf{1} \le n \prod_\ell (1 + \varepsilon_r^{(\ell)})$, and the maximum column sum is at most the total mass. □

Remark 3.4 (When the product certificate is informative)*.* The ceiling $\|T_L\|_1 \le n$ holds for any product of nonnegative row-stochastic mixers, independently of the Sinkhorn schedule. Thus finite-step Sinkhorn control is informative when it certifies a budget below this ceiling:

$$\prod_{\ell=1}^{L} (1 + \varepsilon_c^{(\ell)}) < n.$$

Above this threshold the structural cap, rather than the product surrogate, is the operative worst-case bound.

Proposition 3.5 (Telescoping backward certificate)**.** *Assume the ideal* `column-then-row` *order, so each $H^{(\ell)}$ is nonnegative with exact row sums. Write the prefix products $T_0 := I$ and $T_k := H^{(k)} \cdots H^{(1)}$, and let $\varepsilon_c^{(\ell)} = \|\mathbf{1}^\top H^{(\ell)} - \mathbf{1}^\top\|_{\infty,\mathrm{vec}}$. Then*

$$\|T_L\|_1 \le 1 + \sum_{\ell=1}^{L} \|T_{\ell-1}\|_1\, \varepsilon_c^{(\ell)}. \tag{3.8}$$

Bounding the realized prefix gains by Proposition 3.3, $\|T_{\ell-1}\|_1 \le \min\{\prod_{j<\ell}(1 + \varepsilon_c^{(j)}),\, n\}$, yields the a priori refinement

$$\|T_L\|_1 \le 1 + \sum_{\ell=1}^{L} \min\Big\{\textstyle\prod_{j<\ell}(1 + \varepsilon_c^{(j)}),\, n\Big\}\, \varepsilon_c^{(\ell)} \le 1 + n \textstyle\sum_{\ell=1}^{L} \varepsilon_c^{(\ell)}. \tag{3.9}$$

Substituting instead the multiplicative bound $\|T_{\ell-1}\|_1 \le \prod_{j<\ell}(1 + \varepsilon_c^{(j)})$ recovers exactly the product certificate $\prod_\ell (1 + \varepsilon_c^{(\ell)})$ of Theorem 3.1. The telescoping bound is therefore never worse than the product certificate, and is strictly sharper whenever the realized prefix gains grow sub-multiplicatively.

Proof. Let $w_L = \mathbf{1}$ and $w_{\ell-1} = (H^{(\ell)})^\top w_\ell$, so $w_0 = T_L^\top \mathbf{1}$ and $\|T_L\|_1 = \|w_0\|_{\infty,\mathrm{vec}}$ by Remark 3.2. Write $w_\ell = \mathbf{1} + d_\ell$. Since $(H^{(\ell)})^\top \mathbf{1} = \mathbf{1} + c_\ell$ with $c_\ell := (H^{(\ell)})^\top \mathbf{1} - \mathbf{1}$, $\|c_\ell\|_{\infty,\mathrm{vec}} = \varepsilon_c^{(\ell)}$, and $\mathbf{1}^\top c_\ell = \mathbf{1}^\top (H^{(\ell)})^\top \mathbf{1} - n = (H^{(\ell)}\mathbf{1})^\top \mathbf{1} - n = 0$ by exact row sums, the recursion gives

$$d_{\ell-1} = c_\ell + (H^{(\ell)})^\top d_\ell, \qquad d_L = 0,$$

with every $d_\ell \in \mathbf{1}^\perp$. Unrolling,

$$d_0 = \sum_{\ell=1}^{L} (H^{(1)})^\top \cdots (H^{(\ell-1)})^\top c_\ell = \sum_{\ell=1}^{L} T_{\ell-1}^\top c_\ell,$$

using $(H^{(1)})^\top \cdots (H^{(\ell-1)})^\top = (H^{(\ell-1)} \cdots H^{(1)})^\top = T_{\ell-1}^\top$. Because $w_0 = \mathbf{1} + d_0 \geq 0$ entrywise, $\|T_L\|_1 = \max_i (w_0)_i = 1 + \max_i (d_0)_i \leq 1 + \|d_0\|_{\infty,\mathrm{vec}}$, and

$$\|T_{\ell-1}^\top c_\ell\|_{\infty,\mathrm{vec}} \leq \|T_{\ell-1}^\top\|_\infty \, \|c_\ell\|_{\infty,\mathrm{vec}} = \|T_{\ell-1}\|_1 \, \varepsilon_c^{(\ell)},$$

which gives (3.8). The algebraic identity $\prod_\ell (1+\varepsilon_c^{(\ell)}) = 1 + \sum_\ell \varepsilon_c^{(\ell)} \prod_{j<\ell} (1+\varepsilon_c^{(j)})$ shows the product certificate is exactly the case in which each prefix gain is replaced by its multiplicative upper bound. □

Remark 3.6 (Computing prefix gains). The prefix gains in (3.8) are obtained by

$$u_0 = \mathbf{1}, \qquad u_\ell = (H^{(\ell)})^\top u_{\ell-1}, \qquad \|T_\ell\|_1 = \|u_\ell\|_{\infty,\mathrm{vec}}.$$

Thus the telescoping bound is a rigorous intermediate certificate between the a priori product bound and the exact final propagated gain.

3.2 Depth-uniform control under scrambling

The product and telescoping certificates are worst-case bounds driven by marginal defects. A complementary bound is available when the row-stochastic mixers are uniformly scrambling. In that case, the backward recursion contracts mean-zero components in total variation, yielding a certificate independent of the depth.

Theorem 3.7 (Depth-uniform Dobrushin certificate)**.** *Let $H^{(1)}, \ldots, H^{(L)} \in \mathbb{R}^{n\times n}$ be entrywise nonnegative with exact row sums $H^{(\ell)}\mathbf{1} = \mathbf{1}$. For each layer define the column-defect vector*

$$c_\ell := (H^{(\ell)})^\top \mathbf{1} - \mathbf{1}$$

and the Dobrushin coefficient

$$\delta_{\mathrm{D}}(H^{(\ell)}) := \frac{1}{2} \max_{i,k} \sum_j |H_{ij}^{(\ell)} - H_{kj}^{(\ell)}|.$$

If $\bar{\delta} := \max_\ell \delta_{\mathrm{D}}(H^{(\ell)}) < 1$, then the depth-wise backward mixer gain satisfies

$$\|T_L\|_1 \leq 1 + \frac{\max_\ell \|c_\ell\|_{1,\mathrm{vec}}}{1 - \bar{\delta}}. \tag{3.10}$$

This certificate is independent of the depth L.

Proof. Let $w_L = \mathbf{1}$ and $w_{\ell-1} = (H^{(\ell)})^\top w_\ell$. Write $w_\ell = \mathbf{1} + d_\ell$. Since each $H^{(\ell)}$ is row-stochastic, $\mathbf{1}^\top c_\ell = 0$, and the recursion gives

$$d_{\ell-1} = c_\ell + (H^{(\ell)})^\top d_\ell, \qquad d_L = 0.$$

For any zero-sum vector z, the Dobrushin coefficient is the total-variation contraction coefficient:

$$\|(H^{(\ell)})^\top z\|_{1,\mathrm{vec}} \leq \delta_{\mathrm{D}}(H^{(\ell)}) \|z\|_{1,\mathrm{vec}}.$$

Thus

$$\|d_{\ell-1}\|_{1,\mathrm{vec}} \leq \max_j \|c_j\|_{1,\mathrm{vec}} + \bar{\delta} \|d_\ell\|_{1,\mathrm{vec}},$$

and induction yields $\|d_0\|_{1,\mathrm{vec}} \leq \max_j \|c_j\|_{1,\mathrm{vec}}/(1-\bar{\delta})$. Finally, $w_0 = T_L^\top \mathbf{1} = \mathbf{1} + d_0$ is nonnegative, so

$$\|T_L\|_1 = \|w_0\|_{\infty,\mathrm{vec}} \leq 1 + \|d_0\|_{\infty,\mathrm{vec}} \leq 1 + \|d_0\|_{1,\mathrm{vec}},$$

which proves (3.10). □

3.3 Necessity of nonnegativity

Theorem 3.1 uses entrywise nonnegativity through the elementary identities

$$\|H\|_\infty = \max_i \sum_j H_{ij}, \qquad \|H\|_1 = \max_j \sum_i H_{ij}, \qquad H \geq 0.$$

Without nonnegativity, exact row and column sums do not control induced norms. Indeed, for $N > 1$,

$$A_N = \begin{pmatrix} N & 1-N \\ 1-N & N \end{pmatrix}$$

has all row and column sums equal to one, but

$$\|A_N\|_\infty = \|A_N\|_1 = 2N - 1.$$

Since A_N is symmetric, also $\|A_N\|_2 = 2N - 1$. Thus marginal constraints alone do not prevent arbitrary amplification when signs are allowed. Positivity, enforced here by exponentiating logits, is therefore structural for the marginal-error certificate.

The preceding estimates are the mixer-path certificates used throughout the paper. Appendix C records the separate Jacobian bookkeeping needed when these bounds are embedded into a full residual block.

4 Finite-step Sinkhorn: bounds and calibration

Theorem 3.1 bounds depth-wise mixer gains in terms of per-layer marginal errors. We now relate those errors to the finite-step Sinkhorn iterations used in the mHC-motivated mixer parameterization and obtain a conservative rule for choosing $(\kappa, t_{\max})$ to meet a prescribed depth-L budget.

Standing assumptions. Throughout Section 4, the pre-scaling matrix is strictly positive, $S^{(0)} \in \mathbb{R}^{n\times n}_{>0}$, and has bounded dynamic range

$$\kappa := \frac{\max_{i,j} S^{(0)}_{ij}}{\min_{i,j} S^{(0)}_{ij}} < \infty. \tag{4.1}$$

In applications this is enforced by clipped logits, for example $S^{(0)}_{ij} = \exp(\beta C_{ij})$ with $C_{ij} \in [-R, R]$. A full Sinkhorn step means column normalization followed by row normalization. If a layer uses $t^{(\ell)} \in \{1, \dots, t_{\max}\}$ full steps, then

$$H^{(\ell)}_{\text{res}} = S^{(2t^{(\ell)})}. \tag{4.2}$$

The fixed-step variant uses $t^{(\ell)} = t_{\max}$; the early-stopped variant uses $t^{(\ell)} = t^{(\ell)}_{\text{stop}} \leq t_{\max}$. The theorems use exact normalization maps. If an implementation uses denominator stabilizers or floors, the certificate should be evaluated using the actually measured row and column defects of the implemented matrix.

4.1 Alternating normalization and marginal errors

Given logits $\widetilde{H_{\text{res}}}^{(\ell)} \in \mathbb{R}^{n\times n}$, the mixer parameterization forms a positive matrix

$$S^{(0)} := \exp(\beta\, \widetilde{H_{\text{res}}}^{(\ell)}) \qquad \text{(entrywise)}, \tag{4.3}$$

where $\beta > 0$ is an optional temperature. We then apply alternating column and row normalizations:

$$S^{(s+1)} := \begin{cases} T_c(S^{(s)}) & s \text{ even}, \\ T_r(S^{(s)}) & s \text{ odd}, \end{cases} \qquad s = 0, 1, 2, \dots \tag{4.4}$$

where T_c divides each column by its column sum and T_r divides each row by its row sum. After a column normalization we have exact column sums, and after a row normalization we have exact row sums:

$$\mathbf{1}^\top S^{(s)} = \mathbf{1}^\top \quad (s \text{ odd}), \qquad S^{(s)}\mathbf{1} = \mathbf{1} \quad (s \text{ even},\ s \geq 2). \tag{4.5}$$

Accordingly, the nontrivial marginal errors occur on alternating half-steps. We define

$$\varepsilon_r^{(s)} := \left\| S^{(s)}\mathbf{1} - \mathbf{1} \right\|_{\infty,\mathrm{vec}}, \qquad \varepsilon_c^{(s)} := \left\| \mathbf{1}^\top S^{(s)} - \mathbf{1}^\top \right\|_{\infty,\mathrm{vec}}. \tag{4.6}$$

If we run t *full* Sinkhorn steps in the `column-then-row` order, this corresponds to $2t$ alternating normalizations and the output

$$H_{\mathrm{res}}^{(\ell)} := S^{(2t)}. \tag{4.7}$$

We use either (i) a fixed-step choice $t = t_{\max}$, or (ii) an early-stopped choice $t = t_{\mathrm{stop}}^{(\ell)} \le t_{\max}$ where $t_{\mathrm{stop}}^{(\ell)}$ is the smallest full-step count achieving the target column-marginal tolerance in (4.11).

Layer-level errors used in Theorem 3.1. For a layer output $H_{\mathrm{res}}^{(\ell)} = S^{(2t)}$ (with t given by $t^{(\ell)}$ in (4.2)) define $\varepsilon_r^{(\ell)} := \|H_{\mathrm{res}}^{(\ell)}\mathbf{1} - \mathbf{1}\|_{\infty,\mathrm{vec}}$ and $\varepsilon_c^{(\ell)} := \|\mathbf{1}^\top H_{\mathrm{res}}^{(\ell)} - \mathbf{1}^\top\|_{\infty,\mathrm{vec}}$.

Lemma 4.1 (Ideal output row sums and induced ℓ_∞ control)**.** *For the exact arithmetic normalization maps T_c and T_r, under the `column-then-row` order and for any $t \ge 1$, the output $H_{\mathrm{res}}^{(\ell)} := S^{(2t)}$ satisfies*

$$H_{\mathrm{res}}^{(\ell)}\mathbf{1} = \mathbf{1} \tag{4.8}$$

If moreover $H_{\mathrm{res}}^{(\ell)} \ge 0$, then

$$\|H_{\mathrm{res}}^{(\ell)}\|_\infty = 1, \tag{4.9}$$

so the only nontrivial marginal error for the backward certificate is the output column error $\varepsilon_c^{(\ell)} = \|\mathbf{1}^\top H_{\mathrm{res}}^{(\ell)} - \mathbf{1}^\top\|_{\infty,\mathrm{vec}}$, which controls $\|H_{\mathrm{res}}^{(\ell)}\|_1$.

Proof. Each full step ends with a row normalization, hence (4.8). If $H_{\mathrm{res}}^{(\ell)} \ge 0$, then $\|H_{\mathrm{res}}^{(\ell)}\|_\infty = \max_i \sum_j (H_{\mathrm{res}}^{(\ell)})_{ij} = 1$. □

Under the `column-then-row` schedule, the forward mixer-path $A_{\max}$ gain is certified layerwise. The remaining design problem is the backward one: controlling $\|H_{\mathrm{res}}^{(\ell)}\|_1$ through the output column-marginal error $\varepsilon_c^{(\ell)} = \|\mathbf{1}^\top H_{\mathrm{res}}^{(\ell)} - \mathbf{1}^\top\|_{\infty,\mathrm{vec}}$.

Early stopping to a target column tolerance. At each layer, we run full `column-then-row` Sinkhorn steps and measure the output column defect

$$\varepsilon_c^{(\ell)}(t) := \|\mathbf{1}^\top S^{(2t)} - \mathbf{1}^\top\|_{\infty,\mathrm{vec}}. \tag{4.10}$$

The layer stops at the first full-step count

$$t_{\mathrm{stop}}^{(\ell)} = \min\{1 \le t \le t_{\max} : \varepsilon_c^{(\ell)}(t) \le \bar{\varepsilon}_\ell\}, \tag{4.11}$$

if such a step exists, and otherwise returns the $t_{\max}$ output with a failed certificate for that layer. Appendix A summarizes the certificate bookkeeping. The column defect $\varepsilon_c^{(\ell)}(t)$ is always measured after a full column-then-row step, namely at $S^{(2t)}$.

Remark 4.2 (Implemented normalizations). The theorems use exact division by current row and column sums. If code uses denominator stabilizers, floors, or mixed precision, the certificate should be computed from the realized output matrix, using its measured row and column defects. In the reported runs the stabilizer is 10^{-12}, so the measured row defect is negligible.

4.2 Dynamic range and Hilbert-metric control

Assume $S^{(0)}$ has bounded dynamic range: there exist $0 < m \le M$ such that $m \le S_{ij}^{(0)} \le M$ for all i, j, and define $\kappa := M/m$. If $S^{(0)} = \exp(\beta C)$, then

$$\kappa = \exp\big(\beta(\max_{i,j} C_{ij} - \min_{i,j} C_{ij})\big), \tag{4.12}$$

so clipping $C_{ij} \in [-R, R]$ implies $\kappa \leq \exp(2\beta R)$.

Recall that for $u, v \in \mathbb{R}^n_{>0}$, the Hilbert projective metric is

$$d_H(u, v) := \log\left(\max_i \frac{u_i}{v_i}\right) - \log\left(\min_i \frac{u_i}{v_i}\right). \tag{4.13}$$

Lemma 4.3 (Explicit initial spread)**.** *If $m \leq S^{(0)}_{ij} \leq M$ and $u := S^{(0)}\mathbf{1}$, then $d_H(u, \mathbf{1}) \leq \log \kappa$. Similarly, for $v := \mathbf{1}^\top S^{(0)}$ we have $d_H(v^\top, \mathbf{1}) \leq \log \kappa$.*

Proof. Each component $u_i = \sum_j S^{(0)}_{ij}$ lies in $[nm, nM]$, so $\max_i u_i / \min_i u_i \leq \kappa$ and $d_H(u, \mathbf{1}) = \log(\max_i u_i / \min_i u_i) \leq \log \kappa$. The column-sum statement is identical. □

Lemma 4.4 (Hilbert-to-marginal conversion)**.** *Let $u \in \mathbb{R}^n_{>0}$ satisfy $\frac{1}{n}\sum_{i=1}^n u_i = 1$ and set $\delta := d_H(u, \mathbf{1}) = \log\big(\max_i u_i / \min_i u_i\big)$. Then*

$$\|u - \mathbf{1}\|_{\infty,\mathrm{vec}} \leq \Psi_n(\delta) := \frac{(n-1)(e^\delta - 1)}{(n-1) + e^\delta}. \tag{4.14}$$

Moreover, for any $\varepsilon \in [0, n-1)$,

$$\Psi_n(\delta) \leq \varepsilon \quad \Longleftrightarrow \quad \delta \leq \Psi_n^{-1}(\varepsilon) := \log\left(\frac{(n-1)(1+\varepsilon)}{(n-1) - \varepsilon}\right). \tag{4.15}$$

Proof. See Appendix B. □

Lemma 4.5 (Centered-log L^2 conversion to marginal error)**.** *Let $u \in \mathbb{R}^n_{>0}$ satisfy $n^{-1}\sum_i u_i = 1$ and put*

$$z := \Pi \log u, \qquad \eta := \|z\|_2.$$

Then

$$\|u - \mathbf{1}\|_{\infty,\mathrm{vec}} \leq \Gamma_n(\eta), \tag{4.16}$$

where

$$\Gamma_n(\eta) := \frac{n \exp(a)}{\exp(a) + (n-1)\exp(-a/(n-1))} - 1, \qquad a := \sqrt{\frac{n-1}{n}}\,\eta. \tag{4.17}$$

The bound is sharp for the upper tail: equality is attained when one centered log-coordinate equals a and the remaining $n-1$ centered log-coordinates equal $-a/(n-1)$. For small η,

$$\Gamma_n(\eta) = \sqrt{\frac{n-1}{n}}\,\eta + O(\eta^2). \tag{4.18}$$

Proof. See Appendix B. □

4.3 A finite-step sufficient rule

Classical Birkhoff–Bushell contraction theory for positive maps implies geometric contraction in the Hilbert metric with a rate controlled by the dynamic range (Birkhoff, 1957; Bushell, 1973). Specializing to the alternating normalizations above yields the following conservative but explicit design bound.

Half-step indexing. For clarity we record the indexing convention used throughout the finite-step analysis. Starting from $S^{(0)} = A$, the alternating normalizations are

$$S^{(1)} = T_c(A), \qquad S^{(2)} = T_r T_c(A), \qquad S^{(3)} = T_c T_r T_c(A), \qquad \ldots. \tag{4.19}$$

Thus $S^{(2t)}$ is the output after t full column-then-row steps. In exact arithmetic, $S^{(2t)}$ has exact row sums, whereas its column sums are generally only approximate. This is the origin of the one-half-step loss in the output column-error estimate: the column error at $S^{(2t)}$ is controlled at rate τ^{2t-1}, not τ^{2t}.

Lemma 4.6 (Birkhoff–Bushell contraction for positive matrices)**.** *Let $A \in \mathbb{R}^{n\times n}_{>0}$. Define its projective diameter by*

$$\Delta(A) := \log \max_{i,k,j,\ell} \frac{A_{ij}A_{k\ell}}{A_{i\ell}A_{kj}}, \tag{4.20}$$

and set

$$\tau(A) := \tanh\left(\frac{\Delta(A)}{4}\right). \tag{4.21}$$

Then, for all $x, y \in \mathbb{R}^n_{>0}$,

$$d_H(Ax, Ay) \le \tau(A) d_H(x, y), \qquad d_H(A^\top x, A^\top y) \le \tau(A) d_H(x, y). \tag{4.22}$$

Moreover, componentwise inversion is an isometry for the Hilbert metric:

$$d_H(x^{-1}, y^{-1}) = d_H(x, y), \tag{4.23}$$

where x^{-1} denotes the componentwise reciprocal.

Proof. The first estimate is the classical Birkhoff–Bushell contraction theorem for positive linear maps in Hilbert's projective metric. Since A and $A^\top$ have the same projective diameter, the same contraction coefficient applies to $A^\top$. Finally,

$$d_H(x^{-1}, y^{-1}) = \log \frac{\max_i y_i/x_i}{\min_i y_i/x_i} = \log \frac{\max_i x_i/y_i}{\min_i x_i/y_i} = d_H(x, y).$$

□

Theorem 4.7 (Finite-step marginal-error bound from projective contraction)**.** *Let $A := S^{(0)} \in \mathbb{R}^{n\times n}_{>0}$, $n \ge 2$, and let $S^{(s)}$ be the alternating normalization iterates from Subsection 4.1. Define the projective diameter*

$$\Delta(A) := \log \max_{i,k,j,\ell} \frac{A_{ij}A_{k\ell}}{A_{i\ell}A_{kj}}, \tag{4.24}$$

and set

$$\tau(A) := \tanh\left(\frac{\Delta(A)}{4}\right). \tag{4.25}$$

Let

$$P^{(1)} := T_c(A), \qquad \delta_0 := d_H(P^{(1)}\mathbf{1}, \mathbf{1}). \tag{4.26}$$

Then, for every integer $t \ge 0$,

$$\varepsilon_r^{(2t+1)} := \|S^{(2t+1)}\mathbf{1} - \mathbf{1}\|_{\infty,\mathrm{vec}} \le \Psi_n\big(\delta_0 \tau(A)^{2t}\big). \tag{4.27}$$

Moreover, for every integer $t \ge 1$,

$$\varepsilon_c^{(2t)} := \|\mathbf{1}^\top S^{(2t)} - \mathbf{1}^\top\|_{\infty,\mathrm{vec}} \le \Psi_n\big(\delta_0 \tau(A)^{2t-1}\big). \tag{4.28}$$

In particular, if A has entrywise dynamic range

$$\kappa = \frac{\max_{i,j} A_{ij}}{\min_{i,j} A_{ij}},$$

then

$$\tau(A) \le q := \frac{\kappa - 1}{\kappa + 1}, \qquad \delta_0 \le \log \kappa,$$

and hence

$$\varepsilon_r^{(2t+1)} \le \Psi_n\big((\log \kappa) q^{2t}\big), \qquad t \ge 0, \tag{4.29}$$

while

$$\varepsilon_c^{(2t)} \le \Psi_n\big((\log\kappa)q^{2t-1}\big), \qquad t \ge 1. \tag{4.30}$$

Consequently, to guarantee $\varepsilon_c^{(2t_{\max})} \le \bar{\varepsilon}$*, it is sufficient that*

$$t_{\max} \ge \frac{1}{2}\left[1 + \frac{\log\big((\log\kappa)/\Psi_n^{-1}(\bar{\varepsilon})\big)}{\log(1/q)}\right], \tag{4.31}$$

with the right-hand side rounded up and with $t_{\max} \ge 1$ *unless* A *is already normalized. If* $\kappa = 1$*, then* A *is a constant matrix and one full step gives an exactly doubly stochastic output.*

Proof. Let $r^{(t)} \in \mathbb{R}^n_{>0}$ be the accumulated row-scaling vector at the beginning of the t-th full column-then-row step, with $r^{(0)} = \mathbf{1}$, modulo an irrelevant scalar factor. After the column normalization in full step t,

$$S^{(2t+1)} = \operatorname{diag}(r^{(t)})A\operatorname{diag}\Big((A^\top r^{(t)})^{-1}\Big).$$

Its row-sum vector is therefore

$$u_t := S^{(2t+1)}\mathbf{1} = r^{(t)} \odot A(A^\top r^{(t)})^{-1}.$$

The following row normalization divides by this row-sum vector, hence

$$r^{(t+1)} = r^{(t)} \oslash u_t = \Big[A(A^\top r^{(t)})^{-1}\Big]^{-1} =: \Phi(r^{(t)}).$$

We first prove that Φ contracts in Hilbert distance with coefficient $\tau(A)^2$. For $r, s \in \mathbb{R}^n_{>0}$, Lemma 4.6 gives

$$\begin{aligned} d_H(A^\top r, A^\top s) &\le \tau(A) d_H(r, s), \\ d_H((A^\top r)^{-1}, (A^\top s)^{-1}) &= d_H(A^\top r, A^\top s), \\ d_H(A(A^\top r)^{-1}, A(A^\top s)^{-1}) &\le \tau(A) d_H((A^\top r)^{-1}, (A^\top s)^{-1}). \end{aligned}$$

Applying the inversion isometry once more yields

$$d_H(\Phi(r), \Phi(s)) \le \tau(A)^2 d_H(r, s).$$

Consequently,

$$d_H(r^{(t+1)}, r^{(t)}) \le \tau(A)^{2t} d_H(r^{(1)}, r^{(0)}).$$

Since $u_t = r^{(t)} \oslash r^{(t+1)}$,

$$d_H(u_t, \mathbf{1}) = d_H(r^{(t)}, r^{(t+1)}) \le \tau(A)^{2t}\delta_0, \qquad \delta_0 = d_H(u_0, \mathbf{1}) = d_H(P^{(1)}\mathbf{1}, \mathbf{1}).$$

After a column normalization the total mass is n, so $n^{-1}\sum_i (u_t)_i = 1$. Lemma 4.4 gives (4.27).

It remains to control the column sums at the full-step output $S^{(2t)}$. For $t \ge 1$, write

$$v_t := (\mathbf{1}^\top S^{(2t)})^\top.$$

Using the row-scaled representation before and after the preceding row normalization gives

$$v_t = (A^\top r^{(t)}) \oslash (A^\top r^{(t-1)}).$$

Therefore

$$d_H(v_t, \mathbf{1}) = d_H(A^\top r^{(t)}, A^\top r^{(t-1)}) \le \tau(A) d_H(r^{(t)}, r^{(t-1)}) \le \delta_0 \tau(A)^{2t-1}.$$

Since $S^{(2t)}$ has total mass n, Lemma 4.4 gives (4.28).

Finally, if $m \le A_{ij} \le M$ and $\kappa = M/m$, then $\Delta(A) \le 2\log\kappa$, hence

$$\tau(A) \le \tanh\left(\frac{\log\kappa}{2}\right) = \frac{\kappa - 1}{\kappa + 1}.$$

Also, if $P^{(1)} = T_c(A)$, then for any rows i, k,

$$\frac{(P^{(1)}\mathbf{1})_i}{(P^{(1)}\mathbf{1})_k} = \frac{\sum_j P^{(1)}_{ij}}{\sum_j P^{(1)}_{kj}} \le \max_j \frac{P^{(1)}_{ij}}{P^{(1)}_{kj}} = \max_j \frac{A_{ij}}{A_{kj}} \le \kappa.$$

Thus $\delta_0 \le \log \kappa$. The step-count condition follows by solving

$$(\log \kappa) q^{2t_{\max}-1} \le \Psi_n^{-1}(\bar{\varepsilon}).$$

□

Corollary 4.8 (Column–then–row output bound)**.** *Under the ideal exact column–then–row schedule with output $H^{(\ell)}_{\text{res}} = S^{(2t_{\max})}$, the output satisfies $H^{(\ell)}_{\text{res}}\mathbf{1} = \mathbf{1}$ and hence $\|H^{(\ell)}_{\text{res}}\|_\infty = 1$ for $H^{(\ell)}_{\text{res}} \ge 0$. In implemented normalizations with denominator stabilizers, this exact row-sum statement is replaced by the measured row defect described in Remark 4.2. Moreover, in the ideal exact schedule the output column-marginal error obeys*

$$\varepsilon^{(\ell)}_{\text{c}} = \left\|\mathbf{1}^\top H^{(\ell)}_{\text{res}} - \mathbf{1}^\top\right\|_{\infty,\text{vec}} \le \Psi_n\left((\log \kappa) q^{2t_{\max}-1}\right), \qquad q = \frac{\kappa - 1}{\kappa + 1}. \tag{4.32}$$

Thus certifying the backward mixing-path gain reduces to choosing $t_{\max}$ so that the right-hand side of (4.32) is at most the desired per-layer tolerance.

Remark 4.9 (Conservatism of the step-count bound). The κ-only estimate is a uniform worst-case certificate. The matrix-adaptive version using $\Delta(A)$ and δ_0 is sharper for a fixed matrix, while measured-error early stopping is usually sharper still. None of these Hilbert bounds is intended as a numerical decay predictor.

4.4 Matrix-specific local analysis

For a fixed matrix, the nonlinear Sinkhorn map in logarithmic projective coordinates gives matrix-specific information beyond the global dynamic-range bound. Its derivative at the Sinkhorn limit gives the asymptotic rate $\sigma_2(P)^2$, while derivative control on a verified ball gives a local non-asymptotic certificate.

Let

$$P = \operatorname{diag}(r_*) A \operatorname{diag}(c_*) \in \mathcal{B}_n$$

be the limiting doubly stochastic scaling of $A \in \mathbb{R}^{n\times n}_{>0}$, which exists and is unique, with r_*, c_* unique up to reciprocal scaling, by the Sinkhorn–Knopp theorem for strictly positive matrices (Sinkhorn and Knopp, 1967). We work on $\mathbf{1}^\perp$ and define

$$\mathcal{F}(x) := \Pi\left[-\log\left(P\left((P^\top e^x)^{-1}\right)\right)\right], \qquad \mathcal{L}(x) := \Pi \log(P^\top e^x), \tag{4.33}$$

where all exponentials, logarithms, and inverses are componentwise. If

$$x_t := \Pi \log \frac{r^{(t)}}{r_*},$$

then

$$x_{t+1} = \mathcal{F}(x_t), \tag{4.34}$$

and the centered logarithmic output column error after $t \ge 1$ full steps is

$$z_t := \Pi \log v_t = \mathcal{L}(x_t) - \mathcal{L}(x_{t-1}), \qquad v_t := (\mathbf{1}^\top S^{(2t)})^\top. \tag{4.35}$$

Proposition 4.10 (Local spectral rate)**.** *Let $A \in \mathbb{R}^{n\times n}_{>0}$ and let $P = \operatorname{diag}(r_*) A \operatorname{diag}(c_*) \in \mathcal{B}_n$ be its Sinkhorn limit. Set*

$$K := PP^\top, \qquad \rho_* := \|K\Pi\|_{2\to 2}.$$

Then

$$\rho_* = \sigma_2(P)^2. \tag{4.36}$$

Moreover,

$$D\mathcal{F}(0) = K|_{\mathbf{1}^\perp}, \tag{4.37}$$

so, locally,

$$x_{t+1} = Kx_t + O(\|x_t\|^2). \tag{4.38}$$

Consequently, for every $\eta > 0$ there is a neighborhood of zero such that

$$\|x_t\|_2 \le C_\eta(\rho_* + \eta)^t \|x_0\|_2. \tag{4.39}$$

Finally,

$$z_t = \Pi P^\top (K - I)K^{t-1}x_0 + O\big((\rho_* + \eta)^t \|x_0\|_2^2\big) \tag{4.40}$$

locally. In the generic case where x_0 has a nonzero projection onto the leading nontrivial eigenspace of K, the small-error ratio is governed by

$$\frac{\|z_{t+1}\|}{\|z_t\|} \longrightarrow \rho_* = \sigma_2(P)^2$$

up to the usual modification when the leading nontrivial eigenvalue has multiplicity.

Proof. The recurrence (4.34) follows from the identities

$$A^\top r^{(t)} = c_*^{-1} \odot P^\top e^{x_t}, \qquad (A^\top r^{(t)})^{-1} = c_* \odot (P^\top e^{x_t})^{-1},$$

where the irrelevant scalar in e^{x_t} is removed by Π. The row-scaling update gives exactly

$$e^{x_{t+1}} \sim \left[P\left((P^\top e^{x_t})^{-1}\right)\right]^{-1},$$

and hence (4.34). Since $P^\top \mathbf{1} = \mathbf{1}$,

$$P^\top e^x = \mathbf{1} + P^\top x + O(\|x\|^2),$$

so

$$(P^\top e^x)^{-1} = \mathbf{1} - P^\top x + O(\|x\|^2).$$

Applying P and taking $-\log$ gives

$$\mathcal{F}(x) = \Pi P P^\top x + O(\|x\|^2).$$

On $\mathbf{1}^\perp$, $K = PP^\top$ is symmetric and has operator norm $\sigma_2(P)^2$, proving (4.36)–(4.38). The perturbative estimate (4.39) is the standard consequence of a C^2 map whose derivative at the fixed point has spectral radius ρ_*. Formula (4.35) gives

$$z_t = \mathcal{L}(x_t) - \mathcal{L}(x_{t-1}).$$

Since $D\mathcal{L}(0) = \Pi P^\top$ and $x_t = K^t x_0 + O((\rho_* + \eta)^t \|x_0\|^2)$ locally, (4.40) follows. □

Remark 4.11 (Spectral gap and genuine contraction). For strictly positive A the Sinkhorn limit P is strictly positive and hence primitive, so its singular values satisfy $\sigma_2(P) < \sigma_1(P) = 1$; consequently $\rho_* = \sigma_2(P)^2 < 1$ is a strict contraction rate and the local recursion $x_{t+1} = Kx_t + O(\|x_t\|^2)$ is locally attracting. The limiting ratio statement requires only that x_0 has nonzero projection onto the leading nontrivial singular subspace of P; otherwise the rate is governed by the largest nonzero mode actually excited by x_0.

The asymptotic rate does not by itself certify a finite orbit. The following proposition gives a local finite-step envelope.

Proposition 4.12 (Non-asymptotic local spectral envelope)**.** *For $x \in \mathbf{1}^\perp$, define*

$$g(x) := P^\top e^x, \qquad a(x) := P\big(g(x)^{-1}\big),$$

and set

$$J_F(x) := \Pi \operatorname{diag}(a(x))^{-1} P \operatorname{diag}(g(x))^{-2} P^\top \operatorname{diag}(e^x)\Pi, \tag{4.41}$$

$$J_L(x) := \Pi \operatorname{diag}(g(x))^{-1} P^\top \operatorname{diag}(e^x)\Pi. \tag{4.42}$$

For $R > 0$, put

$$\rho_R := \sup_{\substack{x \in \mathbf{1}^\perp \\ \|x\|_2 \le R}} \|J_F(x)\|_{2\to 2}, \qquad \lambda_R := \sup_{\substack{x \in \mathbf{1}^\perp \\ \|x\|_2 \le R}} \|J_L(x)\|_{2\to 2}. \tag{4.43}$$

The quantities ρ_R and λ_R are exact local derivative envelopes. Any computable upper bounds $\widehat{\rho}_R \ge \rho_R$ and $\widehat{\lambda}_R \ge \lambda_R$ may be substituted in the conclusion, provided $\widehat{\rho}_R < 1$. Assume $\rho_R < 1$ and $\|x_0\|_2 \le R$. Then the full Sinkhorn orbit remains in the ball $\|x\|_2 \le R$ and, for every $t \ge 1$,

$$\|z_t\|_2 \le \lambda_R \rho_R^{t-1} \|x_1 - x_0\|_2. \tag{4.44}$$

Consequently, the actual column-marginal error satisfies the following local finite-step bound

$$\varepsilon_c^{(2t)} \le \Gamma_n\big(\lambda_R \rho_R^{t-1} \|x_1 - x_0\|_2\big), \tag{4.45}$$

where Γ_n is defined in Lemma 4.5. The same conclusion holds with ρ_R, λ_R replaced by any certified upper bounds $\widehat{\rho}_R, \widehat{\lambda}_R$ satisfying $\widehat{\rho}_R < 1$.

Proof. The displayed formulae for J_F and J_L are obtained by differentiating (4.33). Because $\mathcal{F}(0) = 0$, the mean-value theorem gives, for any $x \in \mathbf{1}^\perp$ with $\|x\|_2 \le R$,

$$\|\mathcal{F}(x)\|_2 \le \rho_R \|x\|_2.$$

Thus the ball is invariant whenever $\rho_R < 1$. Applying the same mean-value estimate to the differences gives

$$\|x_{s+1} - x_s\|_2 \le \rho_R \|x_s - x_{s-1}\|_2, \qquad s \ge 1,$$

and therefore

$$\|x_t - x_{t-1}\|_2 \le \rho_R^{t-1} \|x_1 - x_0\|_2.$$

Since $z_t = \mathcal{L}(x_t) - \mathcal{L}(x_{t-1})$, another mean-value estimate gives (4.44). Lemma 4.5 converts the centered logarithmic error into the ordinary column-sum error. □

Corollary 4.13 (Orbit-wise derivative certificate)**.** *Let $x_0, x_1, \dots, x_t$ be the actual logarithmic Sinkhorn orbit and define*

$$\rho_s := \sup_{0 \le \theta \le 1} \big\| J_F\big((1-\theta)x_{s-1} + \theta x_s\big) \big\|_{2\to 2}, \qquad s \ge 1,$$

$$\lambda_t^{\mathrm{seg}} := \sup_{0 \le \theta \le 1} \big\| J_L\big((1-\theta)x_{t-1} + \theta x_t\big) \big\|_{2\to 2}.$$

Then

$$\|z_t\|_2 \le \lambda_t^{\mathrm{seg}} \left(\prod_{s=1}^{t-1} \rho_s \right) \|x_1 - x_0\|_2, \tag{4.46}$$

and hence

$$\varepsilon_c^{(2t)} \le \Gamma_n \left(\lambda_t^{\mathrm{seg}} \left(\prod_{s=1}^{t-1} \rho_s \right) \|x_1 - x_0\|_2 \right). \tag{4.47}$$

This is an a posteriori certificate along the realized orbit; it replaces the global dynamic-range rate by the actual derivative norms encountered by the matrix.

Proof. Apply the mean-value theorem on each segment joining x_{s-1} to x_s and then on the segment joining x_{t-1} to x_t for $\mathcal{L}$. The final conversion is Lemma 4.5. □

4.5 Certificates versus predictors

The bounds above have different roles. The κ-only and $(\Delta(A), \delta_0)$ Hilbert estimates are certificates; the local derivative envelope in Proposition 4.12 is a conditional certificate on a verified neighborhood of the Sinkhorn limit. The spectral rate $\rho_* = \sigma_2(P)^2$ and the finite-step expansions are predictors used to interpret observed decay curves.

For the plots we use the first-order local predictor

$$\ell_t^{[1]} = \Pi P^\top (K - I) K^{t-1} x_0, \qquad K = PP^\top, \tag{4.48}$$

followed by exponentiation and recentering,

$$\widehat{v}_t^{[1]} = \frac{n \exp(\ell_t^{[1]})}{\sum_{j=1}^n \exp((\ell_t^{[1]})_j)}, \qquad \widehat{\varepsilon}_{c,1}^{(2t)} = \|\widehat{v}_t^{[1]} - \mathbf{1}\|_{\infty,\mathrm{vec}}. \tag{4.49}$$

The second-order predictor is given in Appendix C; neither predictor enters the deterministic budget certificate.

Remark 4.14 (Stability–expressivity trade-off via logit range). Smaller dynamic range κ, obtained for instance by clipping or smaller temperature β, improves the contraction factor and reduces the step budget needed for a target marginal tolerance. At the same time, as $\kappa \downarrow 1$, the feasible matrices concentrate near $n^{-1}\mathbf{1}\mathbf{1}^\top$, reducing selective routing. Thus κ trades off certificate strength against routing expressivity.

4.6 Mixer-path calibration from $A_{\max}$ budgets

Combining the depth-wise product certificate (Theorem 3.1) with the finite-step Sinkhorn bound (Theorem 4.7) gives a direct calibration rule from target $A_{\max}$ budgets to implementable parameters (logit dynamic range and step count).

Theorem 4.15 (Mixer-path calibration rule: $(\kappa, t_{\max}) \Rightarrow \|T_L\|_1$ budget)**.** *Assume each mixer $H_{\mathrm{res}}^{(\ell)}$ is entrywise nonnegative and produced by exponentiating logits with dynamic range at most κ, followed by $t_{\max} \geq 1$ ideal exact full Sinkhorn steps in the* `column-then-row` *order. Suppose the desired per-layer column-marginal tolerance is ε_c. If*

$$\Psi_n\big((\log \kappa) q^{2t_{\max}-1}\big) \leq \varepsilon_c, \qquad q = \frac{\kappa - 1}{\kappa + 1}, \tag{4.50}$$

then every output mixer satisfies $\varepsilon_c^{(\ell)} \leq \varepsilon_c$ and the backward $A_{\max}$ gain of the depth-L mixer product T_L satisfies

$$\|T_L\|_1 \leq (1 + \varepsilon_c)^L. \tag{4.51}$$

In particular, to enforce a backward budget $\|T_L\|_1 \leq B_1$ it is sufficient to choose

$$\varepsilon_c \leq B_1^{1/L} - 1 \;\approx\; \frac{\log B_1}{L} \qquad (\textit{for small } \varepsilon_c), \tag{4.52}$$

and then choose $t_{\max}$ satisfying (4.50). Equivalently, it is sufficient that

$$t_{\max} \geq \frac{1}{2}\left[1 + \frac{\log\big((\log \kappa)/\Psi_n^{-1}(\varepsilon_c)\big)}{\log(1/q)}\right], \tag{4.53}$$

rounded up, with $t_{\max} \geq 1$ unless the matrix is already normalized.

Proof. The column-marginal estimate (4.30) gives $\varepsilon_c^{(\ell)} \leq \varepsilon_c$ under (4.50). Theorem 3.1 then gives

$$\|T_L\|_1 \leq \prod_{\ell=1}^{L} (1 + \varepsilon_c^{(\ell)}) \leq (1 + \varepsilon_c)^L.$$

The tolerance choice (4.52) is exactly the condition $(1 + \varepsilon_c)^L \leq B_1$. Solving (4.50) for $t_{\max}$ gives (4.53). □

Remark 4.16 (Only targets below the ceiling require a step budget). By Proposition 3.3, in the ideal `column-then-row` order one always has $\|T_L\|_1 \le n$ after a single full step, with no dependence on the dynamic range. Hence the finite-step budget (4.50) is needed only when the target satisfies $B_1 < n$; for $B_1 \ge n$ the structural ceiling already certifies the budget and $t_{\max} = 1$ suffices. Equivalently, the per-layer tolerance (4.52) need only be enforced down to the point where $\prod_\ell (1+\varepsilon_c^{(\ell)}) = n$, after which the ceiling is the binding constraint. The telescoping bound (3.8) can be used in place of the product certificate to certify the same target with larger per-layer tolerances whenever the realized prefix gains are sub-multiplicative.

4.7 Adaptive global-budget early stopping

The uniform rule $\varepsilon_c \le B_1^{1/L} - 1$ is certificate-valid but not step-efficient: some layers reach a much smaller error than required, while other layers need more iterations. The alternative is to allocate the logarithmic product budget online.

Lemma 4.17 (Adaptive log-budget early stopping)**.** *Fix a target backward budget $B_1 \ge 1$ and set $b_1 = \log B_1$. Suppose the layers are processed in order. Define $s_0 = 0$, and before layer ℓ set*

$$\bar{\varepsilon}_\ell := \exp\left(\frac{b_1 - s_{\ell-1}}{L-\ell+1}\right) - 1. \tag{4.54}$$

Run finite-step Sinkhorn at layer ℓ until

$$\varepsilon_c^{(\ell)} \le \bar{\varepsilon}_\ell, \tag{4.55}$$

and then update

$$s_\ell = s_{\ell-1} + \log(1+\varepsilon_c^{(\ell)}). \tag{4.56}$$

If every layer reaches its assigned tolerance before the step cap, then

$$\sum_{\ell=1}^{L} \log(1+\varepsilon_c^{(\ell)}) \le b_1, \qquad \|T_L\|_1 \le B_1. \tag{4.57}$$

Proof. Let $r_{\ell-1} := b_1 - s_{\ell-1}$ be the remaining log-budget. By construction,

$$\log(1+\bar{\varepsilon}_\ell) = \frac{r_{\ell-1}}{L-\ell+1}.$$

Since $\varepsilon_c^{(\ell)} \le \bar{\varepsilon}_\ell$,

$$s_\ell \le s_{\ell-1} + \frac{b_1 - s_{\ell-1}}{L-\ell+1}.$$

Equivalently,

$$b_1 - s_\ell \ge \frac{L-\ell}{L-\ell+1}(b_1 - s_{\ell-1}).$$

Induction gives $s_L \le b_1$. Theorem 3.1 then implies $\|T_L\|_1 \le \exp(s_L) \le B_1$. □

Certificate failure under a finite cap. Lemma 4.17 is a certificate conditional on every layer reaching its allocated tolerance before the step cap. If some layer fails to reach its assigned tolerance by $t_{\max}$, the algorithm returns a failed certificate rather than a false guarantee. The realized matrix may still be used in training, but the deterministic budget $\|T_L\|_1 \le B_1$ is then not certified by the stopping rule. In that case one should increase $t_{\max}$, reduce the logit dynamic range, or accept a larger budget.

For implementation, the rule is evaluated sequentially. After observing

$$s_{\ell-1} = \sum_{j<\ell} \log(1+\varepsilon_c^{(j)}),$$

layer ℓ is assigned the tolerance

$$\bar{\varepsilon}_\ell = \exp\left(\frac{\log B_1 - s_{\ell-1}}{L-\ell+1}\right) - 1. \tag{4.58}$$

The layer then uses the early-stopping rule (4.11). If the cap is reached before $\varepsilon_c^{(\ell)} \le \bar{\varepsilon}_\ell$, the certificate fails for that layer; otherwise $s_\ell = s_{\ell-1} + \log(1+\varepsilon_c^{(\ell)})$. A compact implementation summary is given in Appendix A.

4.8 Temperature–clipping calibration

The dynamic range is controlled in implementations through logit clipping and temperature. Suppose

$$C_{ij} \in [-R, R], \qquad A_{ij} = \exp(\beta C_{ij}). \tag{4.59}$$

Then

$$\kappa(A) \le \exp(2\beta R), \qquad \frac{\kappa(A)-1}{\kappa(A)+1} \le \tanh(\beta R). \tag{4.60}$$

Therefore the conservative finite-step rule becomes

$$\varepsilon_c^{(2t)} \le \Psi_n\big(2\beta R \tanh(\beta R)^{2t-1}\big)\,. \tag{4.61}$$

Corollary 4.18 (Temperature–clipping calibration)**.** *Fix n, a full-step budget $t_{\max} \ge 1$, and a target column tolerance $\bar{\varepsilon}$. If*

$$\Psi_n\big(2\beta R \tanh(\beta R)^{2t_{\max}-1}\big) \le \bar{\varepsilon}, \tag{4.62}$$

then every clipped logit matrix $C \in [-R, R]^{n\times n}$ produces, after $t_{\max}$ full column-then-row Sinkhorn steps, a nonnegative mixer satisfying

$$\varepsilon_c^{(2t_{\max})} \le \bar{\varepsilon}.$$

This follows immediately from (4.60), the monotonicity of Ψ_n, and Theorem 4.7. A depth-L backward budget B_1 is certified whenever $\bar{\varepsilon} \le B_1^{1/L} - 1$.

4.9 Optional certificate-aware penalty

The same certificate quantities can be used as soft training penalties, although the deterministic guarantee still comes from the realized post-normalization marginal errors. For a depth-L stack, define

$$\mathcal{C}_{\text{prod}} := \sum_{\ell=1}^{L} \log(1 + \varepsilon_c^{(\ell)}). \tag{4.63}$$

Given a target backward budget B_1, set

$$\mathcal{R}_{\text{budget}} := [\mathcal{C}_{\text{prod}} - \log B_1]_+^2\,. \tag{4.64}$$

One may train with

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{cert}}\mathcal{R}_{\text{budget}} + \lambda_\kappa \sum_{\ell=1}^{L} [\log \kappa_\ell - \log \kappa_{\max}]_+^2\,. \tag{4.65}$$

The first term penalizes violation of the depth-wise backward budget, while the second prevents the logits from entering a high-dynamic-range regime where many Sinkhorn steps are needed.

If $\mathcal{R}_{\text{budget}} = 0$, then

$$\sum_{\ell=1}^{L} \log(1 + \varepsilon_c^{(\ell)}) \le \log B_1.$$

For entrywise nonnegative mixers, Theorem 3.1 therefore gives

$$\|H^{(L)} \cdots H^{(1)}\|_1 \le B_1. \tag{4.66}$$

Thus zero penalty implies that the realized mixer product satisfies the stated budget certificate. This does not by itself guarantee stable training.

Calibration summary. The workflow described in the Introduction is formalized by Theorem 4.15: choose a column-marginal tolerance from the backward mixer-path budget, then choose or early-stop Sinkhorn steps to meet that tolerance. Table 1 gives conservative examples.

Table 1: Conservative full-step budgets $t_{\max}$ sufficient to achieve $\Psi_n((\log \kappa)q^{2t_{\max}-1}) \le \varepsilon_{\mathrm{tol}}$ (Theorem 4.7). Entries are sufficient $t_{\max}$ values.

κ	$\varepsilon_{\mathrm{tol}}$	$n = 4$	$n = 8$	$n = 16$
2	5×10^{-3}	3	3	3
2	1×10^{-3}	4	4	4
2	5×10^{-4}	4	4	4
5	5×10^{-3}	8	8	8
5	1×10^{-3}	10	10	10
5	5×10^{-4}	11	11	11
10	5×10^{-3}	16	16	16
10	1×10^{-3}	20	20	20
10	5×10^{-4}	21	22	22
20	5×10^{-3}	32	32	33
20	1×10^{-3}	40	40	41
20	5×10^{-4}	43	44	44

Computational note. A full Sinkhorn step (one column and one row normalization) costs $O(n^2)$ time and $O(n^2)$ memory for dense $n \times n$ mixers. The calibration therefore makes the stability–compute trade-off explicit: smaller tolerances or larger dynamic range require larger $t_{\max}$, increasing per-layer cost.

Remark 4.19. The practical motivation for Theorem 4.15 is the $A_{\max}$ behavior reported in (Xie et al., 2025): in an unconstrained baseline, single-layer $A_{\max}$ gains can remain near $O(1)$ while composite backward gains become large at intermediate depth. Our results formalize a deterministic sufficient mechanism: if the implemented finite-step Sinkhorn mixers achieve uniformly small marginal errors, then depth-wise $A_{\max}$ growth is bounded by the corresponding product certificate.

5 Numerical diagnostics

The experiments measure the quantities used in the theory: marginal errors, the product certificate $\prod_{\ell=1}^{L}(1 + \varepsilon_c^{(\ell)})$, the exact backward mixing gain $\|T_L\|_1$, cap failures, and local convergence rates. They test whether the mixer-path gain can be measured and controlled during training, rather than comparing architectures or attributing optimization failures to that gain.

We use two CIFAR-10 configurations: a matched transformer run and a depth-128 stress run. Within each paired comparison, the softmax and Sinkhorn models use identical augmented minibatches, matched non-mixer initialization, and a deterministic per-epoch data order. We record the mixer matrices after the same normalization used in training. Marginal errors are measured as $\varepsilon_c^{(\ell)} = \|\mathbf{1}^\top H^{(\ell)} - \mathbf{1}^\top\|_{\infty,\mathrm{vec}}$ and $\varepsilon_r^{(\ell)} = \|H^{(\ell)}\mathbf{1} - \mathbf{1}\|_{\infty,\mathrm{vec}}$. The exact realized backward mixing gain is computed by the vector recursion $w_L = \mathbf{1}$, $w_{\ell-1} = H^{(\ell)\top} w_\ell$, and $\|T_L\|_1 = \|w_0\|_\infty$, as in Remark 3.2. The same final-epoch weights are also evaluated with fixed $t = t_{\max}$ to separate early-stop behavior from what would be obtained by spending the whole cap.

5.1 Matched transformer run

We run a short-horizon CIFAR-10 transformer-style classifier with $n = 8$ stream copies and trainable stream mixers in each block. The run uses 5000 training images, 2000 test images, eight epochs, and depth $L = 24$. Across seeds $0, 1, 2$, the two rows in Table 2 use identical initial weights, minibatches, optimizer settings, and architecture within each seed; only the stream-mixer normalization changes. The row-only softmax mixer is row-stochastic but has no column control, while the Sinkhorn row uses the adaptive log-budget controller.

The row-only run reaches an exact depth-wise backward mixer gain close to the structural ceiling $n = 8$, whereas the Sinkhorn-budget run keeps $\|T_L\|_1$ near one with comparable task loss and accuracy. Figure 1 shows the same quantities over training. The row-only model has marginally higher mean test accuracy and similar full-network gradient norms, so this run supports mixer-gain controllability without a visible accuracy penalty at this diagnostic scale, not an optimization failure claim.

Policy	Acc.	Loss	Exact $\|T_L\|_1$	Product cert.
row-only softmax	0.350±0.026	1.784±0.045	7.981±0.004	4.46e+21±1.72e+18
Sinkhorn budget	0.339±0.033	1.796±0.049	1.000±2.82e-06	1.006±0.002

Table 2: Matched short-horizon CIFAR-10 transformer diagnostic. Both rows use the same initial weights and minibatches within each seed; only the stream-mixer normalization changes. The row-only mixer is row-stochastic but not column-controlled, while the Sinkhorn row uses the adaptive log-budget controller. Statistics are mean±std over seeds 0, 1, 2.

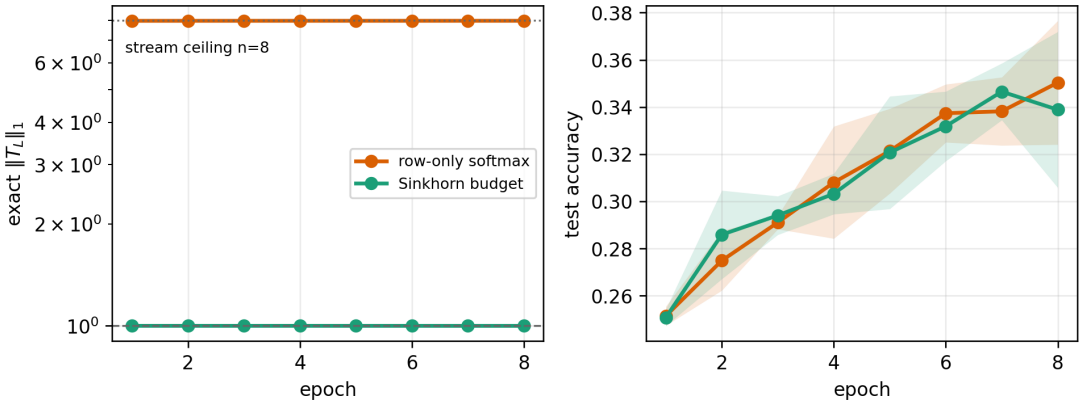


Figure 1: Matched short-horizon transformer run on a CIFAR-10 subset. Row-only softmax mixers nearly saturate the structural backward-gain ceiling $n = 8$, while the column-then-row Sinkhorn controller keeps the exact propagated mixer gain near one. Test accuracy remains comparable under the same training setup.

The remaining diagnostics for the matched transformer run are reported in Appendix A.2. The column-marginal errors confirm that the Sinkhorn controller enforces the intended finite-step tolerance, while the Dobrushin audit gives an additional depth-uniform certificate on the saved final mixer stacks.

5.2 Controller ablation and finite-step diagnostics

The main CIFAR-10 stress setting uses $L = 128$, $n = 8$ streams, feature dimension $d = 128$, SGD with learning rate 0.02, momentum 0.9, weight decay 5×10^{-4}, Sinkhorn tolerance $\varepsilon_{\text{tol}} = 5 \times 10^{-4}$, and cap $t_{\max} = 80$. The softmax branch uses the same architecture but a row-stochastic softmax mixer. The setting uses larger mixer-logit scale and temperature than the regular sweep, making early stopping, cap failures, and dynamic-range effects visible.

The training run records final-epoch dynamic range, early-stopping counts, exact propagated gain, product certificates, local spectral rates, and cap failures. Additional audit tables and raw logs are included in the supplementary archive.

To test the controller in Lemma 4.17 and the soft penalty (4.65), we reran the same CIFAR-10 stress setting with only the Sinkhorn stopping/control policy changed. Table 3 compares fixed cap, uniform early stopping, adaptive log-budget stopping, and an adaptive policy trained with the certificate-aware penalty.

Table 3: Focused controller ablation under the CIFAR-10 stress setting. The adaptive and certificate-aware policies use the shared budget $B_1 = (1 + 5 \cdot 10^{-4})^{128} = 1.066075$. Statistics are computed over seeds $0, 1, 2$; the associated local rates are all approximately $\rho_* = 0.62$.

Policy	Acc.	Time/epoch	Avg. steps	Cert.	Exact $\|T_L\|_1$	Cap fail.
Fixed $t = 80$	0.584 ± 0.037	758.0 ± 6.1s	80.00 ± 0.00	1.0004 ± 0.0003	1.00000 ± 0.00000	$0.00 \pm 0.00\%$
Uniform tol.	0.584 ± 0.037	325.4 ± 17.6s	16.36 ± 0.71	1.0521 ± 0.0012	1.00029 ± 0.00014	$0.26 \pm 0.37\%$
Adaptive budget	0.581 ± 0.034	309.6 ± 19.1s	15.86 ± 0.73	1.0659 ± 0.0001	1.00029 ± 0.00014	$0.26 \pm 0.37\%$
Cert.-aware penalty	0.584 ± 0.038	570.2 ± 34.2s	15.80 ± 0.72	1.0656 ± 0.0000	1.00030 ± 0.00014	$0.26 \pm 0.37\%$

The fixed-cap policy spends substantially more Sinkhorn steps and produces a certificate far below the prescribed budget. Uniform tolerance gives similar accuracy and exact realized backward gain using about one fifth of the Sinkhorn iterations. The adaptive log-budget policy spends the intended product budget more directly and slightly reduces iterations and epoch time relative to uniform tolerance. The certificate-aware penalty is feasible but slower in this run because it adds differentiable certificate terms during training.

5.3 Calibration and local-predictor diagnostics

Figure 2 tests the design inequality (4.61) on random clipped logit matrices with $n = 8$. The measured column errors are always below the finite-precision certificate when the certificate is above the numerical floor. The grid also shows why the temperature–clipping rule is a conservative design certificate rather than a predictor: for large βR, the worst-case certificate is nearly vacuous even though the measured errors decrease rapidly with $t_{\max}$.

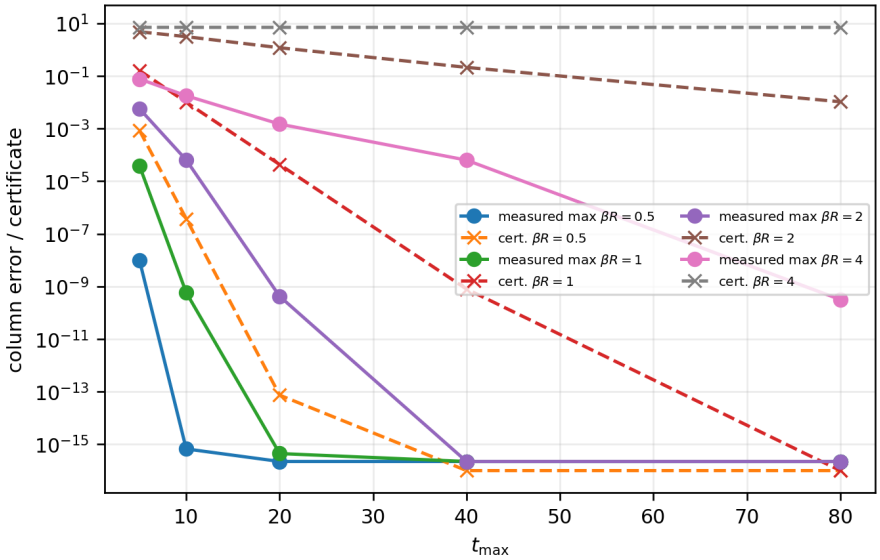


Figure 2: Temperature–clipping diagnostic for the conservative design rule (4.61). Solid curves are measured maximum column errors over random clipped logits; dashed curves are the corresponding worst-case certificates.

For each Sinkhorn layer, we compute the fixed-$t_{\max}$ matrix A^{fix}, use it as the realized doubly stochastic approximation P, and report $\rho_* = \sigma_2(P)^2$. We also report the empirical final-step ratio $\varepsilon_c^{(2t+2)}/\varepsilon_c^{(2t)}$ when the denominator is numerically meaningful. This ratio is diagnostic only; near the early-stopping tolerance or numerical floor it need not estimate the asymptotic local rate.

For trained adaptive Sinkhorn layers, the local first- and second-order predictors track the measured column-error decay much more closely than the global Hilbert certificates. The Hilbert bounds remain conservative, while the local expansions follow the measured decay. Figure 3 shows a representative trained layer.

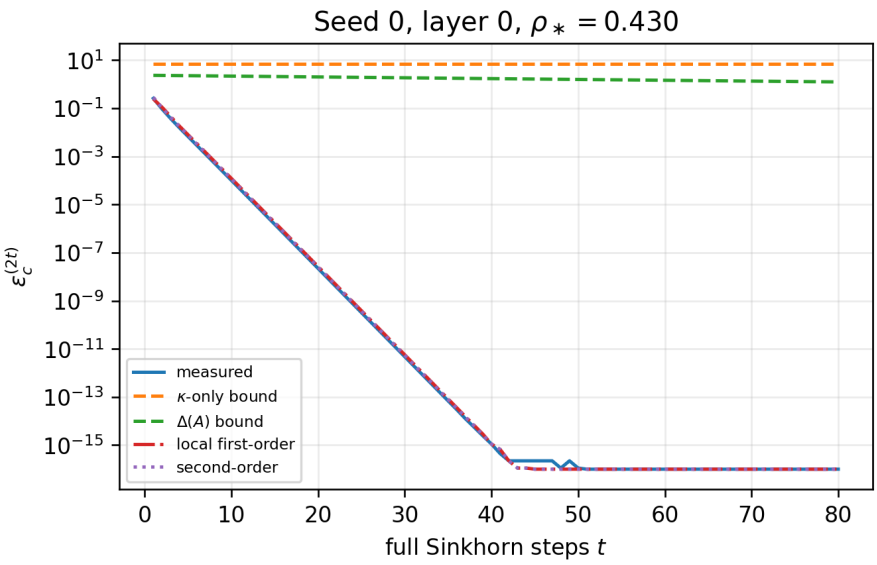


Figure 3: Representative local-predictor comparison on a trained adaptive Sinkhorn layer. The local first- and second-order predictors track the measured Sinkhorn decay closely, while the Hilbert certificates remain conservative.

The stress run operates in a high-dynamic-range regime; the corresponding final-epoch dynamic-range audit is reported in Appendix A.

Appendix A.3 contains a compact stress-run audit figure computed from the same logs.

6 Discussion and practical guidance

6.1 Practical guidance

The results suggest the following use pattern.

- **Keep the mixer nonnegative.** The counterexample in Section 3.3 shows that marginal constraints alone do not control induced norms when signs are allowed.
- **Use the structural ceiling and telescoping refinement.** For row-stochastic nonnegative mixers, $\|T_L\|_1 \leq n$ independently of depth or Sinkhorn schedule. Sinkhorn steps are useful when the desired budget is below this ceiling; the telescoping bound gives a sharper certified estimate using realized prefix gains.
- **Use Dobrushin control as a depth-uniform audit.** When the row-stochastic mixers are uniformly scrambling, Theorem 3.7 gives a backward certificate independent of depth. In the matched transformer run, this bound is close to one for Sinkhorn-controlled mixers, while for row-only softmax the structural ceiling remains the sharper worst-case certificate.
- **Translate depth budgets into finite-step layer tolerances.** For small errors, a composite budget $\|T_L\|_1 \leq B_1$ corresponds to $\sum_{\ell=1}^{L} \varepsilon_c^{(\ell)} \lesssim \log B_1$; the uniform allocation is $\varepsilon_c \approx \log(B_1)/L$. Theorem 4.7 and Corollary 4.18 then convert $(\kappa, t_{\max})$ or $(\beta, R, t_{\max})$ into a conservative tolerance.
- **Use local predictors and adaptive stopping for efficiency.** The κ-only estimate is a worst-case certificate. For realized matrices, local spectral quantities predict the observed marginal-error curve, while the adaptive rule in Lemma 4.17 allocates the remaining log-budget across layers. The exact vector propagation diagnostic in Remark 3.2 then measures the realized gain.

Remark 6.1 (Transfer beyond mHC). The results apply to any depth-wise product of entrywise nonnegative approximately bistochastic mixing matrices, including Sinkhorn-normalized attention on an unmasked support and doubly stochastic message-passing operators. Masked or sparse supports require a support-dependent finite-step analysis.

6.2 Limitations

Worst-case and finite-step bounds. The product and Hilbert-metric bounds are sufficient worst-case certificates and can be conservative. The exact backward gain is computable by Remark 3.2, the gain is structurally capped by Proposition 3.3, and the telescoping bound in Proposition 3.5 sharpens the raw product certificate when realized prefix gains are sub-multiplicative. The local spectral and second-order formulas are predictors of observed finite-step decay, not uniform guarantees.

Nonlinear blocks and training dynamics. The nonlinear block F enters only through Jacobian norm bounds such as Proposition C.2 and Corollary C.3. We do not model how training shapes $\widetilde{H_{\text{res}}}^{(\ell)}$ or how the certificates interact with optimization, data distribution, or generalization.

Architecture and implementation. The statements use streams-as-rows and the column–then–row Sinkhorn order. Streams-as-columns, damping, mixed precision, sparse mixers, and masked supports require explicit translation and can change which marginal error remains at the output.

6.3 Future work

Open directions include certified and inexpensive estimates of the local derivative envelopes ρ_R, λ_R, sharper support- or gap-dependent refinements of the Dobrushin certificate, broader controller and certificate-penalty ablations across datasets and architectures, and full-network bounds that couple mixer certificates to nonlinear-branch Jacobians.

7 Conclusion

We developed deterministic finite-step certificates for nonnegative Sinkhorn-normalized stream mixers. For nonnegative matrices, marginal errors are induced-norm certificates: in the `column-then-row` Sinkhorn order, exact row sums control the forward $A_{\max}$ gain, while the output column defect controls the backward $A_{\max}$ gain of the depth-wise mixer product. Hilbert-projective contraction converts logit dynamic range and finite step count into a conservative design-time tolerance; matrix-specific local analysis explains the faster convergence observed in realized runs. We further show that this backward gain is structurally capped by the expansion rate, $\|T_L\|_1 \leq n$, so the finite-step certificate is informative precisely below that ceiling, and a telescoping refinement tightens the product certificate using realized prefix gains from a companion vector pass.

The theory supports adaptive early stopping, budget tracking, and exact realized-gain diagnostics for the mixing pathway. In the focused ablations, adaptive log-budget stopping matches the uniform tolerance rule in accuracy and exact backward gain while using fewer Sinkhorn steps. The certificate-aware penalty preserves the same stability diagnostics but is slower in the present implementation, indicating that the main compute-saving mechanism is the stopping rule rather than the penalty term. In the matched transformer diagnostic, the mixer-path gain remains near one while test accuracy remains comparable at this scale. This experiment supports controllability of the mixer-path gain; it does not claim that uncontrolled gain is, by itself, the cause of an optimization failure.

Reproducibility Statement

The supplementary archive contains the scripts, fixed seeds, raw CSV/JSON logs, figure-generation scripts, and command manifest needed to regenerate the reported diagnostics and finite-step lookup table. Because the certificates depend on implementation details, the archive records the Sinkhorn dtype, positivity floor $\varepsilon_{\text{floor}}$, clipping/temperature settings, early-stopping tolerance, data path, and preprocessing. An anonymized archive

accompanies the submission; the public repository will be released after review. The archive also includes saved final mixer stacks for the matched transformer run and the script `scripts/dobrushin_diagnostic.py`, which regenerates the depth-uniform Dobrushin audit table.

Ethics and Broader Impact Statement

This work studies finite-step normalization for nonnegative mixer products. The diagnostics use CIFAR-10 and do not involve human-subject data, personal data, or deployment decisions.

References

G. Birkhoff. Extensions of Jentzsch's theorem. *Transactions of the American Mathematical Society*, 85:219–227, 1957.

P. J. Bushell. Hilbert's metric and positive contraction mappings in a Banach space. *Archive for Rational Mechanics and Analysis*, 52:330–338, 1973.

Z. Xie, Y. Wei, H. Cao, et al. mHC: Manifold-Constrained Hyper-Connections. *arXiv preprint arXiv:2512.24880*, 2025. DOI: 10.48550/arXiv.2512.24880. https://arxiv.org/abs/2512.24880.

R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21:343–348, 1967.

J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989.

P. A. Knight. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.

M. Cuturi. Sinkhorn distances: light-speed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

G. Peyré and M. Cuturi. *Computational Optimal Transport*. Foundations and Trends in Machine Learning, 11:355–607, 2019.

M. E. Sander, P. Ablin, M. Blondel, and G. Peyré. Sinkformers: Transformers with Doubly Stochastic Attention. In *Proceedings of AISTATS 2022*, PMLR 151, 2022. arXiv:2110.11773.

A. Shahbazi, E. Akbari, D. Salehi, X. Liu, N. NaderiAlizadeh, and S. Kolouri. ESPFormer: Doubly-Stochastic Attention with Expected Sliced Transport Plans. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:54186–54202, 2025.

E. Haber and L. Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34:014004, 2018.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

A Certificate bookkeeping and extra diagnostics

For each layer, the implementation computes $\bar{\varepsilon}_\ell = \exp((\log B_1 - \sum_{j<\ell} \log(1+\varepsilon_c^{(j)}))/(L-\ell+1)) - 1$ and runs column-then-row Sinkhorn steps until either $\varepsilon_c^{(\ell)}(t) = \|\mathbf{1}^\top S^{(2t)} - \mathbf{1}^\top\|_{\infty,\mathrm{vec}} \le \bar{\varepsilon}_\ell$ or $t = t_{\max}$. The reported product certificate is $\exp(\sum_{\ell=1}^{L} \log(1+\varepsilon_c^{(\ell)}))$, and the exact realized backward gain is computed by the recursion $w_L = \mathbf{1}$, $w_{\ell-1} = (H_{\mathrm{res}}^{(\ell)})^\top w_\ell$, and $\|T_L\|_1 = \|w_0\|_{\infty,\mathrm{vec}}$. If any layer reaches $t_{\max}$ before satisfying its assigned tolerance, the adaptive budget certificate is marked as failed.

This appendix contains post-hoc certificate diagnostics and stress-run audit plots.

A.1 Depth-uniform Dobrushin diagnostic

To audit Theorem 3.7 on trained mixers, we saved the final per-layer residual mixer stacks from the matched transformer run and computed the Dobrushin depth-uniform certificate $1 + (1 - \bar{\delta})^{-1} \max_\ell \|c_\ell\|_{1,\mathrm{vec}}$, where $c_\ell = (H^{(\ell)})^\top \mathbf{1} - \mathbf{1}$ and $\bar{\delta} = \max_\ell \delta_{\mathrm{D}}(H^{(\ell)})$. The diagnostic uses the same final matched-run setting as Figure 1: three seeds, 5000 training images, 2000 test images, eight epochs, depth $L = 24$, and $n = 8$ streams.

Policy	L	n	$\bar{\delta}$	$\max_\ell \|c_\ell\|_1$	Product cert.	Ergodic cert.	Exact $\|T_L\|_1$
softmax seed 0	24	8	9.267×10^{-4}	13.96	4.457×10^{21}	14.98	7.976
sinkhorn seed 0	24	8	0.2345	0.01692	1.0086	1.022	1.0000
softmax seed 1	24	8	7.792×10^{-4}	13.96	4.460×10^{21}	14.97	7.985
sinkhorn seed 1	24	8	0.2082	0.01364	1.0053	1.017	1.0000
softmax seed 2	24	8	9.942×10^{-4}	13.96	4.460×10^{21}	14.98	7.983
sinkhorn seed 2	24	8	0.2006	0.01372	1.0050	1.017	1.0000

Table A.1: Post-hoc Dobrushin diagnostics for saved finite-step mixer stacks.

For the Sinkhorn-controlled mixers, the Dobrushin certificate is far below the structural ceiling $n = 8$ and close to the exact propagated gain. For the row-only softmax mixers, the Dobrushin certificate is much smaller than the raw product certificate but is weaker than the structural ceiling; the exact gain is near the ceiling. The two policies therefore occupy different regimes: Dobrushin control for Sinkhorn and structural capping for row-only softmax.

A.2 Additional matched-run diagnostics

The remaining matched-run diagnostics are shown in Figure A.1. The column-error panel records the finite-step marginal defect, and the gradient-norm panel is included only as an audit diagnostic; the certificates in the main text concern the mixer product rather than full-network gradient stability.

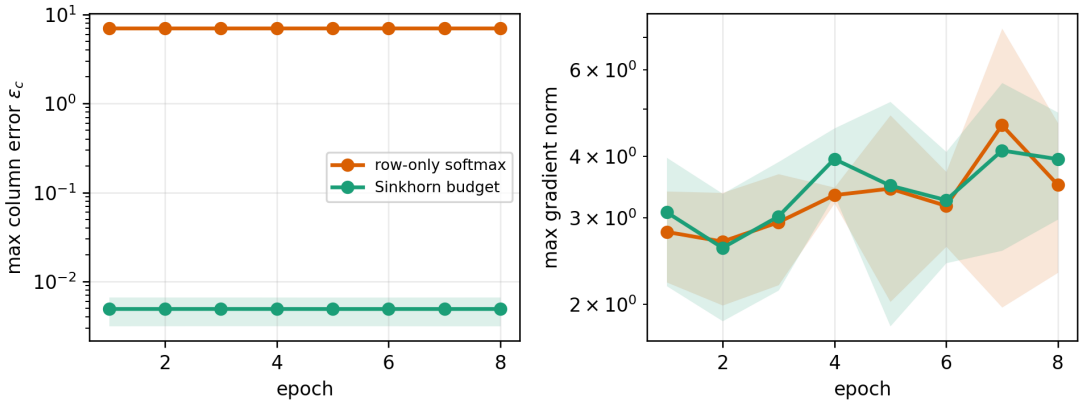


Figure A.1: Additional matched-run diagnostics. The column-error panel records the finite-step marginal defect, and the gradient-norm panel is included as an implementation audit.

A.3 Stress-run audit

These diagnostics use the same stress-run logs as the main certificate tables.

Dynamic-range audit. Table A.2 reports the final-epoch dynamic-range diagnostics for the stress run. For Sinkhorn, the fixed-cap column evaluates the same trained weights using $t = t_{\max}$. For the row-softmax baseline, κ_{obs} is a native pre-softmax diagnostic, not the clipped positive-matrix dynamic range used in the Sinkhorn certificate.

Table A.2: Final-epoch dynamic-range diagnostics for the stress run.

Model	med. κ_{obs}	std	med. $\kappa_{\text{obs}}^{\text{fix}}$	std
softmax	8556629.04	840682.39	—	—
sinkhorn (early)	14489.91	1315.73	14487.00	1324.17

Table A.3: Final-epoch stack certificate diagnostics. The exact backward gain is computed by vector propagation, while the certificate is $\prod_\ell (1 + \varepsilon_c^{(\ell)})$.

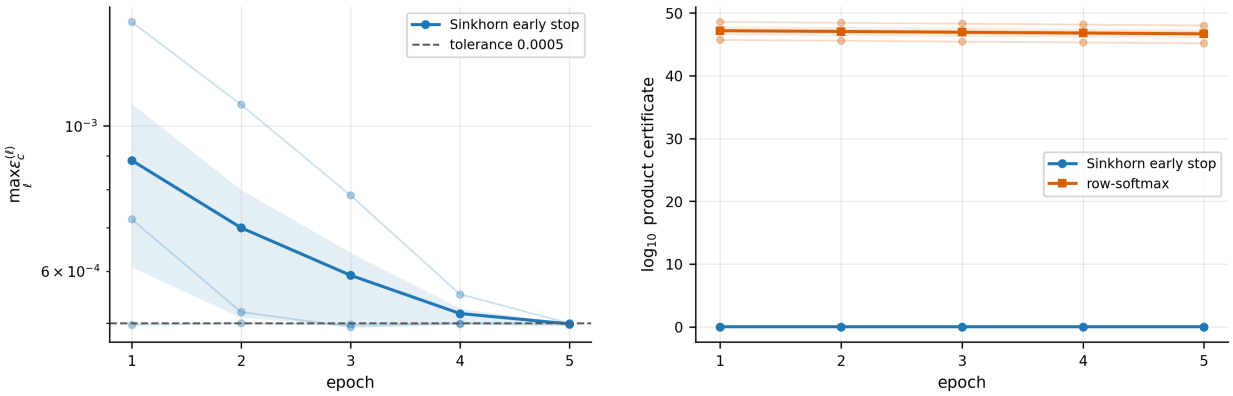
Model	cert.	exact $\|T_L\|_1$	exact/cert.
softmax	$3.61\times10^{47} \pm 5\times10^{47}$	5.72 ± 2	—
sinkhorn (early)	1.05 ± 0.0009	$1 \pm 8\times10^{-5}$	0.951 ± 0.0009

For the early-stopped Sinkhorn run, the final epoch reaches the tolerance $\varepsilon_{\text{tol}} = 5 \times 10^{-4}$ with median stopping time $t_{\text{stop}} = 14$. Re-evaluating the same trained weights with the full cap $t_{\max} = 80$ gives a smaller column defect, about 1.9×10^{-4}, confirming that the reported stopping behavior is tolerance-driven rather than cap-limited.

The row-softmax baseline illustrates that row stochasticity alone does not give a useful backward product certificate: the product certificate is enormous, although the realized gain is structurally capped and remains $O(n)$. The Sinkhorn controller keeps both the certificate and the exact propagated gain near one.

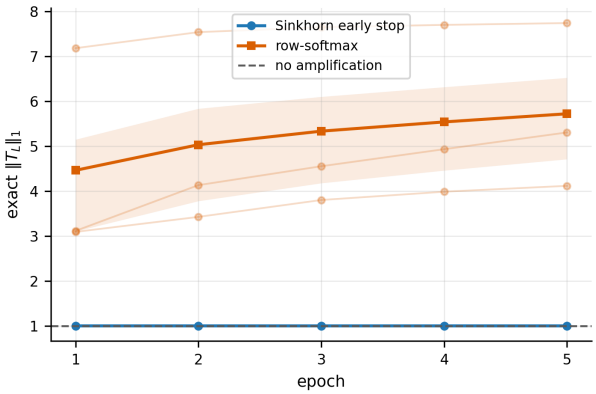
For the Sinkhorn run, the associated local diagnostic is $\rho_* = 0.631 \pm 0.01$, and the cap-failure rate is zero.

Stress-run audit trajectories. Figure A.2 summarizes the main stress-run diagnostics over training. The maximum column defect stays near the early-stopping tolerance, the product certificate remains below the prescribed budget, and the exact propagated gain stays close to one. The stopping-count panel records the compute savings from early stopping. Cap failures are zero in the reported run.

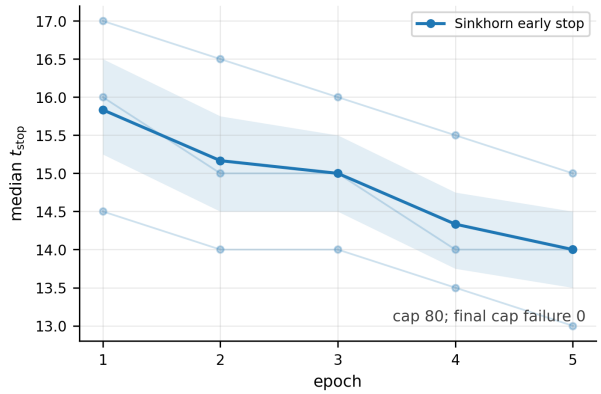


(a) Maximum column defect.

(b) Product certificate.



(c) Exact propagated gain.



(d) Early-stop step counts.

Figure A.2: Stress-run audit diagnostics for the finite-step Sinkhorn controller. These plots use the same logs as the main certificate tables.

B Auxiliary finite-step estimates

Proof of Lemma 4.4. Let $a := \max_i u_i$, $b := \min_i u_i$, and $r := a/b = e^{\delta} \geq 1$. The mean constraint gives $b \leq 1 \leq a$. For the upper deviation, fixing the ratio $a/b = r$ and the mean, the largest possible value of a occurs when one entry is a and the remaining $n-1$ entries are b. Hence

$$a \leq \frac{nr}{r+n-1}, \qquad a - 1 \leq \frac{(n-1)(r-1)}{r+n-1}.$$

For the lower deviation, the smallest possible value of b occurs when one entry is b and the remaining $n-1$ entries are a. Thus

$$b \geq \frac{n}{1+(n-1)r}, \qquad 1 - b \leq \frac{(n-1)(r-1)}{1+(n-1)r}.$$

Since $1 + (n-1)r \geq r + n - 1$ for $n \geq 2$ and $r \geq 1$, the lower-tail bound is no larger than the upper-tail bound. Therefore

$$\|u - \mathbf{1}\|_{\infty,\mathrm{vec}} \leq \frac{(n-1)(r-1)}{r+n-1} = \Psi_n(\delta).$$

Solving $\Psi_n(\delta) \leq \varepsilon$ for $r = e^{\delta}$ gives (4.15).

Proof of Lemma 4.5. Since $z = \Pi \log u$ and $n^{-1} \sum_i u_i = 1$, there is a unique scalar normalization such that

$$u_i = \frac{ne^{z_i}}{\sum_{k=1}^{n} e^{z_k}}.$$

We first maximize a single coordinate. Under the constraints $\sum_i z_i = 0$ and $\|z\|_2 \leq \eta$, the largest possible value of one coordinate is

$$a = \sqrt{\frac{n-1}{n}}\,\eta,$$

attained when the remaining coordinates all equal $-a/(n-1)$. Indeed, if $z_i = a$ is fixed, the sum of the other coordinates is $-a$; Jensen's inequality gives

$$\sum_{k \neq i} e^{z_k} \geq (n-1)e^{-a/(n-1)},$$

with equality exactly in the equal-tail configuration. Therefore

$$u_i \leq \frac{ne^{a}}{e^{a} + (n-1)e^{-a/(n-1)}}.$$

This gives the upper-tail deviation $\Gamma_n(\eta)$.

It remains to check that the lower-tail deviation is not larger. Put

$$r := \exp\left(\frac{n}{n-1}a\right) \geq 1.$$

For the upper-tail extremizer, the maximal coordinate is

$$U = \frac{nr}{r+n-1}.$$

For the lower-tail extremizer, where one coordinate is $-a$ and the other $n-1$ coordinates are $a/(n-1)$, the minimal coordinate is

$$L = \frac{n}{1+(n-1)r}.$$

The upper and lower deviations are $U-1$ and $1-L$. A direct calculation gives

$$(U-1)-(1-L) = U+L-2 = \frac{(n-1)(n-2)(r-1)^2}{(r+n-1)(1+(n-1)r)} \geq 0. \tag{B.1}$$

Thus the one-spike upper-tail value controls the full ℓ_∞ deviation. Expanding (4.17) at $\eta = 0$ gives (4.18).

C Predictor expansion and full-block bookkeeping

C.1 Second-order local predictor formulas

The second-order diagnostic predictor keeps the quadratic terms in the logarithmic Sinkhorn map. With $K = PP^\top$, define

$$Q_F(x) := -P\big((P^\top x)^{\odot 2}\big) + \frac{1}{2}K(x^{\odot 2}) + \frac{1}{2}(Kx)^{\odot 2}, \tag{C.1}$$

$$Q_L(x) := \frac{1}{2}P^\top(x^{\odot 2}) - \frac{1}{2}(P^\top x)^{\odot 2}. \tag{C.2}$$

Then

$$\mathcal{F}(x) = Kx + \Pi Q_F(x) + O(\|x\|^3), \qquad \mathcal{L}(x) = \Pi P^\top x + \Pi Q_L(x) + O(\|x\|^3). \tag{C.3}$$

Set

$$x_t^{[1]} := K^t x_0, \qquad x_t^{[2]} := K^t x_0 + \sum_{s=0}^{t-1} K^{t-1-s}\Pi Q_F(K^s x_0). \tag{C.4}$$

The second-order centered log-column predictor is

$$\ell_t^{[2]} := \Pi\Big[P^\top x_t^{[2]} + Q_L(x_t^{[1]}) - P^\top x_{t-1}^{[2]} - Q_L(x_{t-1}^{[1]})\Big]. \tag{C.5}$$

Finally define

$$\widehat{v}_t^{[2]} := \frac{n\exp(\ell_t^{[2]})}{\sum_{j=1}^n \exp((\ell_t^{[2]})_j)}, \qquad \widehat{\varepsilon}_{c,2}^{(2t)} := \|\widehat{v}_t^{[2]} - \mathbf{1}\|_{\infty,\mathrm{vec}}. \tag{C.6}$$

For each fixed t,

$$z_t = \ell_t^{[2]} + O_t(\|x_0\|^3). \tag{C.7}$$

Thus (C.6) is the finite-step prediction to compare with early numerical points, while (4.49) gives the asymptotic slope.

C.2 Full-layer Jacobian $A_{\max}$ bounds

The mixer-product certificates in Section 3 isolate the contribution of stream mixing. Full-network certificates additionally require independent control of the nonlinear residual branch. We first record the generic bookkeeping bound, then give the row-major Jacobian calculation for the mHC-style layer in (2.1).

Proposition C.1 (Full-block $A_{\max}$ composition bound)**.** *Consider a depth-L network of the form*

$$X_{\ell+1} = H^{(\ell)} X_\ell + \mathcal{F}_\ell(X_\ell), \qquad \ell = 1, \dots, L, \tag{C.8}$$

where $H^{(\ell)} \in \mathbb{R}^{n\times n}$ mixes streams and $\mathcal{F}_\ell$ denotes the remaining nonlinear part of the block. Suppose

$$A_{\max}(\mathcal{F}_\ell(X) - \mathcal{F}_\ell(Y)) \le a_\ell A_{\max}(X - Y) \tag{C.9}$$

for all relevant X, Y. Then

$$A_{\max}(X_L - Y_L) \le \prod_{\ell=1}^{L} \big(\|H^{(\ell)}\|_\infty + a_\ell\big) A_{\max}(X_0 - Y_0). \tag{C.10}$$

If, in addition, along the realized trajectory the adjoint Jacobian of the nonlinear branch satisfies

$$A_{\max}(D\mathcal{F}_\ell(X_\ell)^* G) \le b_\ell A_{\max}(G), \tag{C.11}$$

then the input-gradient amplification obeys

$$A_{\max}(\nabla_{X_0}\mathcal{L}) \le \prod_{\ell=1}^{L} \big(\|H^{(\ell)}\|_1 + b_\ell\big) A_{\max}(\nabla_{X_L}\mathcal{L}). \tag{C.12}$$

In particular, if $H^{(\ell)} \ge 0$ is $(\varepsilon_r^{(\ell)}, \varepsilon_c^{(\ell)})$-doubly stochastic, then

$$A_{\max}(\nabla_{X_0}\mathcal{L}) \le \prod_{\ell=1}^{L} \big(1 + \varepsilon_c^{(\ell)} + b_\ell\big) A_{\max}(\nabla_{X_L}\mathcal{L}). \tag{C.13}$$

Proof. Subtracting two trajectories and using the triangle inequality gives

$$\begin{aligned} A_{\max}(X_{\ell+1} - Y_{\ell+1}) &\le A_{\max}\big(H^{(\ell)}(X_\ell - Y_\ell)\big) + A_{\max}\big(\mathcal{F}_\ell(X_\ell) - \mathcal{F}_\ell(Y_\ell)\big) \\ &\le \big(\|H^{(\ell)}\|_\infty + a_\ell\big) A_{\max}(X_\ell - Y_\ell), \end{aligned}$$

which iterates to (C.10). The backward recursion has the form

$$G_\ell = (H^{(\ell)})^\top G_{\ell+1} + D\mathcal{F}_\ell(X_\ell)^* G_{\ell+1}.$$

Using (2.5) and (C.11) gives

$$A_{\max}(G_\ell) \le \big(\|H^{(\ell)}\|_1 + b_\ell\big) A_{\max}(G_{\ell+1}),$$

and iteration yields (C.12). The marginal version follows from (3.1). □

Let $X \in \mathbb{R}^{n\times C}$ and define the row-major vectorization

$$\mathrm{vec}_r(X) := \mathrm{vec}(X^\top) \in \mathbb{R}^{nC}, \tag{C.14}$$

where vec is the standard column-stacking operator. With this convention one has, whenever products are well-defined,

$$\mathrm{vec}_r(AXB) = (A \otimes B^\top)\,\mathrm{vec}_r(X), \tag{C.15}$$

and in particular

$$\mathrm{vec}_r(HX) = (H \otimes I_C)\,\mathrm{vec}_r(X). \tag{C.16}$$

Layer Jacobian and induced-norm bounds Fix a layer ℓ. Let $H_{\text{res}}^{(\ell)} \in \mathbb{R}^{n\times n}$ be the residual mixer and $H_{\text{pre}}^{(\ell)}, H_{\text{post}}^{(\ell)} \in \mathbb{R}^{1\times n}$ the pre/post aggregation row vectors. Let $F^{(\ell)} : \mathbb{R}^C \to \mathbb{R}^C$ denote the feature map applied to the aggregated feature vector $u := H_{\text{pre}}^{(\ell)} X \in \mathbb{R}^C$, and let $J_F^{(\ell)} \in \mathbb{R}^{C\times C}$ be the Jacobian of $F^{(\ell)}$ with respect to u.

The layer update is

$$X \mapsto H_{\text{res}}^{(\ell)} X + H_{\text{post}}^{(\ell)\top} F^{(\ell)}(H_{\text{pre}}^{(\ell)} X). \tag{C.17}$$

Applying Equation (C.15) and the chain rule yields the Jacobian (with respect to $\text{vec}_r(X)$):

$$J^{(\ell)} = H_{\text{res}}^{(\ell)} \otimes I_C + \left(H_{\text{post}}^{(\ell)\top} \otimes I_C\right) J_F^{(\ell)} \left(H_{\text{pre}}^{(\ell)} \otimes I_C\right). \tag{C.18}$$

Proposition C.2 (Layer Jacobian induced-norm bounds)**.** *With the notation above,*

$$\left\|J^{(\ell)}\right\|_\infty \le \left\|H_{\text{res}}^{(\ell)}\right\|_\infty + \left\|H_{\text{post}}^{(\ell)}\right\|_{\infty,\text{vec}} \left\|J_F^{(\ell)}\right\|_\infty \left\|H_{\text{pre}}^{(\ell)}\right\|_{1,\text{vec}}. \tag{C.19}$$

Similarly,

$$\left\|J^{(\ell)}\right\|_1 \le \left\|H_{\text{res}}^{(\ell)}\right\|_1 + \left\|H_{\text{post}}^{(\ell)}\right\|_{1,\text{vec}} \left\|J_F^{(\ell)}\right\|_1 \left\|H_{\text{pre}}^{(\ell)}\right\|_{\infty,\text{vec}}. \tag{C.20}$$

Proof. Using Equation (C.18) and submultiplicativity of induced norms,

$$\left\|J^{(\ell)}\right\|_\infty \le \left\|H_{\text{res}}^{(\ell)} \otimes I_C\right\|_\infty + \left\|H_{\text{post}}^{(\ell)\top} \otimes I_C\right\|_\infty \left\|J_F^{(\ell)}\right\|_\infty \left\|H_{\text{pre}}^{(\ell)} \otimes I_C\right\|_\infty. \tag{C.21}$$

Since $\|A \otimes I_C\|_\infty = \|A\|_\infty$ for any A, this becomes

$$\left\|J^{(\ell)}\right\|_\infty \le \left\|H_{\text{res}}^{(\ell)}\right\|_\infty + \left\|H_{\text{post}}^{(\ell)\top}\right\|_\infty \left\|J_F^{(\ell)}\right\|_\infty \left\|H_{\text{pre}}^{(\ell)}\right\|_\infty. \tag{C.22}$$

Finally, because $H_{\text{post}}^{(\ell)\top}$ is a column vector, $\|H_{\text{post}}^{(\ell)\top}\|_\infty = \|H_{\text{post}}^{(\ell)}\|_{\infty,\text{vec}}$, and because $H_{\text{pre}}^{(\ell)}$ is a row vector, $\|H_{\text{pre}}^{(\ell)}\|_\infty = \|H_{\text{pre}}^{(\ell)}\|_{1,\text{vec}}$, yielding Equation (C.19). The $\|\cdot\|_1$ bound Equation (C.20) follows analogously. □

Corollary C.3 (Conditional full-block Jacobian certification)**.** *Assume each residual mixer is entrywise nonnegative and $(\varepsilon_r^{(\ell)}, \varepsilon_c^{(\ell)})$-doubly stochastic so that $\left\|H_{\text{res}}^{(\ell)}\right\|_\infty \le 1 + \varepsilon_r^{(\ell)}$. If, in addition, there exist constants $b_{\text{pre}}, b_{\text{post}}, L_F \ge 0$ such that for all layers ℓ,*

$$\left\|H_{\text{pre}}^{(\ell)}\right\|_{1,\text{vec}} \le b_{\text{pre}}, \qquad \left\|H_{\text{post}}^{(\ell)}\right\|_{\infty,\text{vec}} \le b_{\text{post}}, \qquad \left\|J_F^{(\ell)}\right\|_\infty \le L_F, \tag{C.23}$$

then

$$\left\|J^{(\ell)}\right\|_\infty \le (1 + \varepsilon_r^{(\ell)}) + b_{\text{post}}\, L_F\, b_{\text{pre}}. \tag{C.24}$$

Thus

$$\left\|J^{(L)} \cdots J^{(1)}\right\|_\infty \le \prod_{\ell=1}^{L} \Big((1 + \varepsilon_r^{(\ell)}) + b_{\text{post}}\, L_F\, b_{\text{pre}}\Big). \tag{C.25}$$

Remark C.4 (On bounding L_F). The constant L_F depends on the block F. For pointwise MLP blocks, one may bound L_F by products of induced operator norms of the affine maps times activation Lipschitz constants (in $\|\cdot\|_\infty$). For attention-style blocks, L_F depends additionally on projection norms and the (data-dependent) attention map; in such settings, we treat L_F as an architecture- and data-dependent constant and report explicitly how the certificate scales with it.