ONE MARGINAL IS ENOUGH: THE PRICE OF BALANCED ROUTING IN HYPER-CONNECTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hyper-Connections widen the residual pathway into n parallel streams and put a learned $n \times n$ matrix in front of every layer to mix them. Those matrices multiply along depth, so they have to be constrained, and Manifold-Constrained Hyper-Connections (mHC) constrain them to the Birkhoff polytope of doubly stochastic matrices. That is two constraints, one on the row sums and one on the column sums, and we ask what each marginal is worth. Fixing either one alone already bounds the spectral norm of the whole product by $\sqrt{n}$, at any depth, so a single marginal buys the depth-uniform stability the constraint was introduced for. The second only tightens that bound from $\sqrt{n}$ to 1, and it is not free: the mixer can then neither replicate a source nor aggregate at a destination, and it can only preserve or dissipate the differences among the streams, never enlarge them. We therefore keep one marginal and release the other, giving **One-Sided Hyper-Connections (osHC)**. We instantiate four variants — two that fix the same marginal at every layer, and two that switch marginals partway through the network — and prove that all four are stable at every depth. Across ten routing constructions, two pretraining corpora and three scales, osHC attains the best validation loss in five of the six settings and beats every exactly doubly stochastic construction in all six, and at the largest scale it takes the best and the second-best zero-shot perplexity on each of five held-out corpora, under either pretraining corpus. Trained osHC routers use the released capability at almost every site, while their depth-composed gain stays inside the proved ceiling.

1 INTRODUCTION

A residual connection (He et al., 2016b;a) gives a network a single pathway for carrying information forward: layer l reads one hidden state $\mathbf{x}_l \in \mathbb{R}^d$, adds its own output to it, and passes on $\mathbf{x}_{l+1} = \mathbf{x}_l + F_l(\mathbf{x}_l)$. Hyper-Connections (HC) (Zhu et al., 2024) widen that pathway into n parallel states, called *streams* and stacked as $X_l \in \mathbb{R}^{n \times d}$, and let every layer mix them:

$$X_{l+1} = H_l^{\mathrm{res}} X_l + H_l^{\mathrm{post}} F_l(H_l^{\mathrm{pre}} X_l)\,. \tag{1}$$

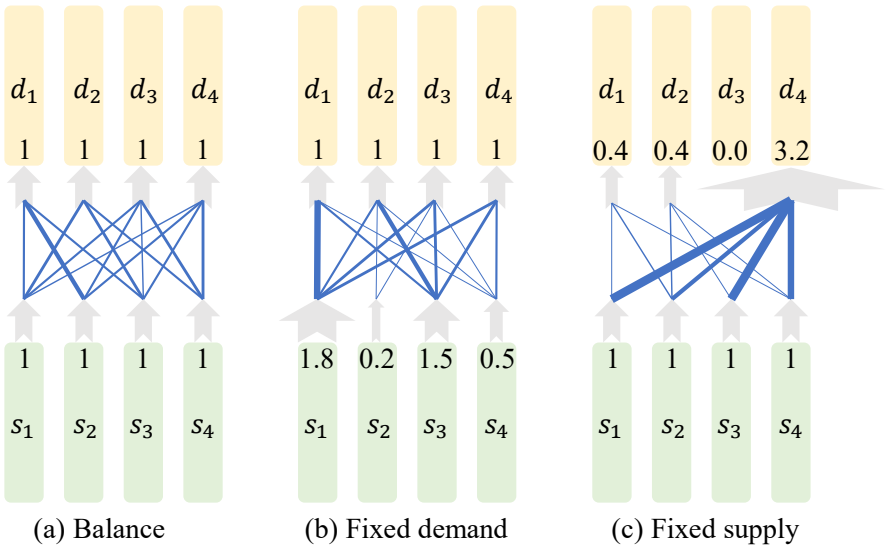
Here $H_l^{\mathrm{pre}} \in \mathbb{R}^{1 \times n}$ reads the streams into the branch F_l and $H_l^{\mathrm{post}} \in \mathbb{R}^{n \times 1}$ writes its output back to them, while the *residual mixer* $H_l^{\mathrm{res}} \in \mathbb{R}^{n \times n}$ routes the old streams directly to the new ones. Throughout the paper we use *router* and *residual mixer* interchangeably. We write $H := H_l^{\mathrm{res}}$ when the layer is clear. Its action on the streams is $(HX)_i = \sum_{j=1}^{n} H_{ij} X_j$, so H_{ij} is the weight routed from *source* stream j to *destination* stream i.

Unrolling Eq. (1) over the L layers, with X_1 the input and X_{L+1} the output, gives

$$X_{L+1} = H_L^{\mathrm{res}} \cdots H_1^{\mathrm{res}} X_1 + \sum_{l=1}^{L} H_L^{\mathrm{res}} \cdots H_{l+1}^{\mathrm{res}}\, H_l^{\mathrm{post}} F_l(H_l^{\mathrm{pre}} X_l)\,. \tag{2}$$

The residual mixers, and they alone, multiply into a single *composite mixer* $\prod_{k=1}^{L} H_{L+1-k}^{\mathrm{res}}$ standing between the input and the output. Left unconstrained, that product amplifies activations and gradients geometrically in L. Manifold-Constrained Hyper-Connections (mHC) (Xie et al., 2025)

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.



(a) Balance (b) Fixed demand (c) Fixed supply

Figure 1: **Balanced routing vs. osHC routing: two marginals or one.** Blue line width is the weight H_{ij} routed from a source stream s_j (green, below) to a destination stream d_i (yellow, above); a missing line is a zero weight. Grey arrow width is the supply leaving a source or the demand entering a destination, printed beside it. **(a)** Balance fixes both at 1. **(b)** osHC-R fixes the demand alone. **(c)** osHC-C fixes the supply alone.

prevent this by making H doubly stochastic: $H \geq 0$ with every row and every column summing to one. These row and column sums are the *marginals* of H: the row sums are the *demand*, the total weight arriving at each destination, and the column sums the *supply*, the total weight leaving each source. A unit demand makes each destination a convex average of its sources, a unit supply makes each source contribute one unit of weight in total, and we call routing that fixes both *balanced*. Such matrices form the Birkhoff polytope $\mathcal{B}_n$, which is closed under multiplication and on which the spectral norm is exactly one — so the composite mixer is Euclidean non-expansive at any depth.

Balance buys real stability, but the price is high. Being doubly stochastic is not one restriction but two, one per marginal, and each costs the router a capability: with every column sum pinned to one, no source can be *replicated* — written onward with more than its own mass — however much the rest of the network wants it; with every row sum pinned, no destination can *aggregate*, taking in more than one stream's worth of weight (Figure 1). Balance also limits how distinct the streams can become: a balanced mixer can only preserve or reduce the differences among them, never enlarge them. We therefore look for a constraint that keeps the stability without the rigidity, and find that one marginal already suffices: fixing either marginal alone still bounds the gain of a composition of any depth, while returning one of the two capabilities to the router.

Contributions.

- **We price balance.** Fixing each marginal costs the router a capability, and balance, being both, costs both: a balanced mixer can neither replicate a source nor aggregate at a destination, and it can never enlarge the differences among the streams (Theorem 3.2).
- **We propose One-Sided Hyper-Connections (osHC).** To keep the stability without the rigidity, we constrain the residual mixer to be one-sided stochastic, and instantiate four variants: two that fix the same marginal at every layer, and two that switch marginals partway through the network. We prove the resulting router is stable at every depth (Theorems 4.1 and 4.2), and that the bound survives a switch (Theorem B.4).
- **We test it.** We compare ten routing constructions on two pretraining corpora at three scales, tracking training dynamics, validation loss, zero-shot transfer to five out-of-distribution corpora, and the routing behaviour of the trained models. osHC attains the best validation loss in five of the six settings, and every osHC variant beats both exact Birkhoff construc-

tions in all six. The trained routers use the released capability at almost every site, and the depth-composed gain stays inside the proved ceiling.

2 Related Work

Manifold-Constrained Hyper-Connections. Building on HC (Zhu et al., 2024), mHC (Xie et al., 2025) constrains all three maps of Eq. (1): the residual mixer is projected onto the Birkhoff polytope, and the read-in and write-out vectors are made non-negative. Write $\vec{X}_l := \mathrm{vec}(X_l) \in \mathbb{R}^{nd}$ for the flattened streams and $\vec{X}'_l := \mathrm{RMSNorm}(\vec{X}_l)$. mHC first forms unconstrained logits

$$\tilde{H}^\star_l = \alpha^\star_l \big(\vec{X}'_l \, \varphi^\star_l\big) + \mathbf{b}^\star_l, \qquad \star \in \{\mathrm{pre}, \mathrm{post}, \mathrm{res}\}, \tag{3}$$

where $\varphi^{\mathrm{pre}}_l, \varphi^{\mathrm{post}}_l \in \mathbb{R}^{nd\times n}$ and $\varphi^{\mathrm{res}}_l \in \mathbb{R}^{nd\times n^2}$ are learnable projections, the $\alpha^\star_l$ are learnable scalars initialised small, and the biases $\mathbf{b}^{\mathrm{pre}}_l, \mathbf{b}^{\mathrm{post}}_l \in \mathbb{R}^n$ and $\mathbf{b}^{\mathrm{res}}_l \in \mathbb{R}^{n^2}$ carry the static part of each map. The three logits are then turned into the three maps of Eq. (1) by

$$H^{\mathrm{pre}}_l = \sigma\big(\tilde{H}^{\mathrm{pre}}_l\big), \quad H^{\mathrm{post}}_l = 2\,\sigma\big(\tilde{H}^{\mathrm{post}}_l\big), \quad H^{\mathrm{res}}_l = \mathrm{SK}_T\big(\mathrm{mat}(\tilde{H}^{\mathrm{res}}_l)\big), \tag{4}$$

where σ is the Sigmoid function and mat reshapes a vector in $\mathbb{R}^{n^2}$ into an $n \times n$ matrix. Here SK_T denotes T Sinkhorn–Knopp iterations (Sinkhorn & Knopp, 1967): its argument is first exponentiated entrywise, and each iteration then normalises the columns and the rows to sum to one. H^{res}_l reaches $\mathcal{B}_n$ as $T \to \infty$, and mHC takes $T = 20$.

Reaching the Birkhoff polytope. Work following mHC has concentrated on attaining $\mathcal{B}_n$ exactly or cheaply rather than on whether $\mathcal{B}_n$ is the right target. mHC-lite (Yang & Gao, 2026) averages permutation matrices, KromHC (Zhou et al., 2026) takes Kronecker products of small doubly stochastic factors, go-mHC (Dandachi & Diggs-Galligan, 2026) and TBP-mHC (Lyubinin, 2026) parameterise $\mathcal{B}_n$ directly, and Wang & Qiu (2026) replace the unrolled Sinkhorn–Knopp solver with a Newton method. These differ in how much of $\mathcal{B}_n$ they cover and at what cost; none varies the constraint itself.

Leaving it. Three methods change the feasible set, and all three pay for the change with non-negativity: sHC (Liu et al., 2026) keeps both marginals but moves H onto the spectral sphere, while JPmHC (Sengupta et al., 2026) and EΔ-MHC-Geo (Shahmansoori, 2026) drop the marginals for orthogonality. What unites them with the exact constructions is that every one targets $\|H\|_2 \le 1$, deliberately, since Euclidean non-expansiveness is the guarantee the line was built on. Section 3.2 shows what insisting on it costs, and Section 4.2 shows that the guarantee is sufficient for depth-uniform control but not necessary. No prior method keeps $H \ge 0$ while fixing exactly one marginal. Appendix A gives a fuller survey.

3 What Balance Forbids

Balance is two constraints rather than one, so we separate them and ask what each is doing.

Definition 3.1 (Marginal families)**.** Let $\mathbf{1} \in \mathbb{R}^n$ denote the all-ones column vector. The two one-sided stochastic families and their intersection are

$$\begin{aligned}
\mathcal{R}_n &:= \{H \in \mathbb{R}^{n\times n} : H \ge 0,\ H\mathbf{1} = \mathbf{1}\} && \text{(row-stochastic / fixed destination demand)},\\
\mathcal{C}_n &:= \{H \in \mathbb{R}^{n\times n} : H \ge 0,\ \mathbf{1}^\top H = \mathbf{1}^\top\} && \text{(column-stochastic / fixed source supply)},\\
\mathcal{B}_n &:= \mathcal{R}_n \cap \mathcal{C}_n && \text{(doubly stochastic / balanced routing)}.
\end{aligned} \tag{5}$$

3.1 Balance flattens the supply and demand profiles

Each one-sided family fixes one marginal and learns the other. For $H \in \mathcal{R}_n$ every destination is a convex combination of its sources, while the source-supply profile $\mathbf{s}(H) := H^\top\mathbf{1}$ is learned; for $H \in \mathcal{C}_n$ every source is spent in full across the destinations, while the destination-demand profile $\mathbf{d}(H) := H\mathbf{1}$ is learned. Each profile has n entries summing to $\sum_{ij} H_{ij} = n$ in all three families.

Balance constrains how that total is distributed, pinning both profiles to the uniform vector rather than letting either be spread unevenly.

On the source side, $s_j(H) = \sum_i H_{ij}$ is the total weight with which the layer writes stream j into the next block. Holding every s_j at 1 sends each source onward with exactly one unit of weight, none with more and none with less: a stream carrying what the rest of the network needs cannot be written more widely, and a stream carrying little cannot be held back. When $\mathbf{s}(H)$ is learned rather than fixed, the layer can grade its sources: $s_j > 1$ *replicates* stream j, writing it onward with more than its own mass, while $s_j < 1$ passes a stream on only in part, damping a source the rest of the network has less use for. On the destination side, $d_i(H) = \sum_j H_{ij}$ is the total weight arriving at stream i. Holding every d_i at 1 makes every destination an interpolation of its sources. When $\mathbf{d}(H)$ is learned rather than fixed, the same grading applies to the destinations: $d_i > 1$ lets stream i *aggregate*, summing its sources rather than interpolating between them, while $d_i < 1$ leaves it at reduced amplitude, an auxiliary channel rather than another full-strength mixture. Balance denies the mixer any means of differentiating its streams, whether by how widely a source is reused or by how much a destination takes in: every stream sends out exactly one unit of weight and takes in exactly one.

Birkhoff's theorem (Birkhoff, 1946) explains why: $\mathcal{B}_n$ is the convex hull of the permutation matrices, so balanced routing is an average of relabelings of the streams. A relabeling is a bijection — it can move what a stream holds to a different stream, but it can neither copy one stream into two nor merge two into one — and averaging bijections does not create either ability.

3.2 Balance confines disagreement

Let $J := \mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{n\times n}$ and let $\Pi := I_n - J/n$ be the orthogonal projector that subtracts the mean stream. Applied to the stream matrix, ΠX replaces each stream by its deviation from the mean stream, so we measure how far the streams have separated by the energy

$$\mathcal{E}(X) := \|\Pi X\|_F^2, \tag{6}$$

which we call the *disagreement*. It vanishes exactly on the *consensus* states $X = \mathbf{1}\mathbf{c}^\top$, $\mathbf{c} \in \mathbb{R}^d$, in which all streams coincide, and it is large when the streams carry distinct content. Sustaining it is what gives n streams more capacity than one, since streams that agree carry no more information than a single stream does. Here $\|\cdot\|_F$ is the Frobenius norm, while $\|\cdot\|_p$ throughout denotes the matrix norm induced by the vector ℓ_p norm.

Theorem 3.2 (Disagreement is non-increasing under balanced routing)**.** *For every balanced residual mixer $H \in \mathcal{B}_n$ and every stream matrix X,*

$$\mathcal{E}(HX) \le \mathcal{E}(X). \tag{7}$$

Balanced routing can therefore only preserve or dissipate the differences among the streams, never enlarge them, and since $\mathcal{B}_n$ is closed under products the same holds however many layers are composed (proof in Appendix B, page 15). Within the relative interior of $\mathcal{B}_n$ the contraction is strict in every non-consensus direction, with equality confined to the relative boundary (Appendix Proposition B.5). The same contraction is documented in the graph networks that share the averaging structure, where repeated averaging drives the representations together and the effect is known as over-smoothing (Oono & Suzuki, 2020; Cai & Wang, 2020; Rusch et al., 2023).

Both prohibitions trace to the same pair of constraints, and the derivation of each uses both members of the pair. Dropping either one lifts both prohibitions at once, which is the design Section 4 adopts.

4 One-Sided Hyper-Connections

We write $Z_l := \mathrm{mat}(\tilde{H}_l^{\mathrm{res}}) \in \mathbb{R}^{n\times n}$ for the residual-mixing logits of Eq. (4), so that $Z_{l,ij}$ scores routing from source j to destination i.

Because either normalisation in SK_T disturbs the sums fixed by the other, mHC attains $\mathcal{B}_n$ only in the limit. At any finite budget the residual mixer it applies still departs from balance, and on trained logits that departure is large: across all mHC checkpoints the marginal normalised last is exact up to 1.8×10^{-7}, while the other reaches a maximum error of 0.999995. The routing mHC performs

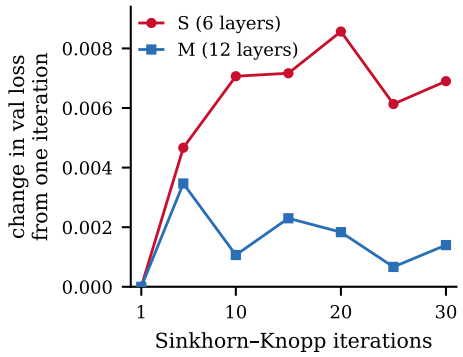


Figure 2: Validation loss of mHC against its Sinkhorn–Knopp iteration count, on FineWeb-Edu at two scales, averaged over three random seeds.

is therefore frequently one-sided rather than balanced. The budget thus sets how nearly balanced the routing is, and we sweep it under the protocol of Section 5. Every larger budget we tested is worse than a single iteration at both scales (Figure 2). Enforcing balance more tightly therefore does not pay for itself, which motivates us to stop approximating it and impose a single marginal exactly instead: the one-sided hyper-connections (osHC) we now define.

4.1 Parameterisation

We drop one marginal and keep everything else. On the same residual-mixing logits Z_l that mHC uses, a single softmax replaces the iterative projection:

$$\text{osHC-R}: \quad H_{ij} = \frac{\exp Z_{ij}}{\sum_k \exp Z_{ik}}, \qquad \text{osHC-C}: \quad H_{ij} = \frac{\exp Z_{ij}}{\sum_k \exp Z_{kj}}. \tag{8}$$

osHC-R normalises each row and so lies exactly in $\mathcal{R}_n$: every destination receives unit total weight, source usage is free, and the mixer can replicate (Figure 1b). **osHC-C** normalises each column and lies exactly in $\mathcal{C}_n$: every source supplies unit total weight, destination totals are free, and the mixer can aggregate (Figure 1c). Only the map from Z_l to H_l^{res} changes; everything else is inherited from mHC.

4.2 One-sided routing is stable at every depth

A single marginal already bounds the Euclidean gain of one mixer, and the family is closed under multiplication, so the same bound covers the composed residual path $H_L^{\text{res}} \cdots H_1^{\text{res}}$ of Eq. (2) at any depth.

Theorem 4.1 (Bounded Euclidean gain under one marginal)**.** *Let $H \in \mathcal{R}_n$. Then*

$$\|H\|_\infty = 1, \qquad 1 \le \|H\|_2 \le \sqrt{\|H\|_1} \le \sqrt{n}. \tag{9}$$

Dually, $H \in \mathcal{C}_n$ gives $\|H\|_1 = 1$ and $1 \le \|H\|_2 \le \sqrt{\|H\|_\infty} \le \sqrt{n}$.

Proof. Non-negativity turns the row-sum condition into an identity for the induced ℓ_∞ norm, $\|H\|_\infty = \max_i \sum_j |H_{ij}| = \max_i \sum_j H_{ij} = 1$, while $H\mathbf{1} = \mathbf{1}$ exhibits 1 as an eigenvalue, so the largest singular value obeys $\|H\|_2 \ge 1$. Since the n row sums are each 1, all entries together sum to n, and no one column sum can exceed that total:

$$\|H\|_1 = \max_j \sum_i |H_{ij}| = \max_j \sum_i H_{ij} \le \sum_j \sum_i H_{ij} = n.$$

Riesz–Thorin interpolation at $p = 2$ between the two endpoint norms gives

$$\|H\|_2 \le \sqrt{\|H\|_1 \, \|H\|_\infty}, \tag{10}$$

which here reads $\|H\|_2 \le \sqrt{\|H\|_1} \le \sqrt{n}$. Transposition exchanges the two endpoint norms and carries $\mathcal{R}_n$ to $\mathcal{C}_n$, giving the dual case. □

Theorem 4.2 (Depth-uniform norm control)**.** *Let $H_1, \dots, H_L \in \mathcal{R}_n$ and $P_L := H_L \cdots H_1$. Then $P_L \in \mathcal{R}_n$, so*

$$\|P_L\|_\infty = 1, \qquad 1 \le \|P_L\|_2 \le \sqrt{n}, \tag{11}$$

with no dependence on L. The same holds for $H_1, \dots, H_L \in \mathcal{C}_n$, with ℓ_1 in place of ℓ_∞.

Proof. $\mathcal{R}_n$ is closed under multiplication: $H_2 H_1 \mathbf{1} = H_2 \mathbf{1} = \mathbf{1}$, and a product of non-negative matrices is non-negative. Induction gives $P_L \in \mathcal{R}_n$, to which Theorem 4.1 applies. In $\mathcal{C}_n$ the same argument runs on $\mathbf{1}^\top H_2 H_1 = \mathbf{1}^\top H_1 = \mathbf{1}^\top$. □

4.3 Depth schedules

Each one-sided family recovers one of the two capabilities, so a network built from a single family can replicate at every layer or aggregate at every layer, but never both. The obvious way to ask for both is to use one family for part of the depth and the other for the rest, and we evaluate four schedules in all. The two *pure* schedules, **osHC-R** and **osHC-C**, apply the same normalisation at all L layers. The two *switch* schedules divide the depth into two contiguous blocks: **osHC-C→R** uses osHC-C for the first $\lfloor L/2 \rfloor$ layers and osHC-R thereafter, and **osHC-R→C** reverses the order.

A pure schedule places every layer's mixer in one family, which is closed under multiplication, so the product $P_L = H_L^{\text{res}} \cdots H_1^{\text{res}}$ inherits both the exact endpoint norm and the ceiling $\|P_L\|_2 \le \sqrt{n}$. A switch breaks the closure: the product of a $\mathcal{C}_n$ block with an $\mathcal{R}_n$ block need not lie in either family, and neither endpoint norm is pinned any longer. What survives is the bound itself. Each block contributes at most $\sqrt{n}$, so one switch gives $\|P_L\|_2 \le n$, and in general m contiguous blocks give $n^{m/2}$ — still a constant fixed by the number of switches rather than by the number of layers (Appendix Theorem B.4).

5 Experiments

Setup. We build on the public mHC-lite implementation and adopt its training setup without modification (Yang & Gao, 2026): nanoGPT decoder-only Transformers at three scales — **Small** (6 layers, 512 width, 45M parameters), **Medium** (12/768, 125M) and **Large** (24/1024, 354M); $n = 4$ streams; 10,000 AdamW updates of 131,072 tokens (1.31B in total) from either OpenWebText (OWT) or FineWeb-Edu; and the initialisation under which every HC block begins as an ordinary residual connection. We run seeds $\{0, 1, 1337\}$ and share optimiser, schedule, and data pipeline across methods. Evaluation differs in one respect: the best-validation checkpoint (estimated on 200 minibatches every 500 updates) is then rescored on the *complete* validation split (4.43M OWT, 49.35M FineWeb-Edu tokens) rather than on a sampled subset. Appendix C gives all hyperparameters, the per-seed losses, and the routing diagnostics and Paloma suite in full.

5.1 Training dynamics

osHC trains to a lower loss, and just as stably. Figure 3 plots the training loss (left) and gradient norm (right) of the five constrained hyper-connection methods. osHC-C represents all four osHC variants, which train almost identically (Figure 4), so plotting one avoids crowding the figure. On the left, osHC-C sits below all three Birkhoff constructions at every plotted step, while those three are indistinguishable from one another: what separates the curves is the constraint, not the projection realising it. sHC, which keeps both marginals but gives up non-negativity, trains above all four. On the right, the four non-negative methods settle at the same level: after about 4000 updates their mean gradient norms lie between 0.464 and 0.479 and no update of any of them exceeds 0.55. sHC never joins them — it averages 3.27 over the same window and individual updates reach 107. Releasing a marginal thus buys a lower loss without losing control of the gradient norm.

The four osHC variants are interchangeable during training. Figure 4 repeats both panels for the osHC variants in isolation. Their losses interleave and end within 0.015 nats of each other. Their gradient norms stay just as close, with means after step 4000 between 0.458 and 0.487. The two switch schedules (osHC-C→R and osHC-R→C) are no exception, although alternating the released marginal relaxes the composed bound from $\sqrt{n}$ to n (Theorem B.4).

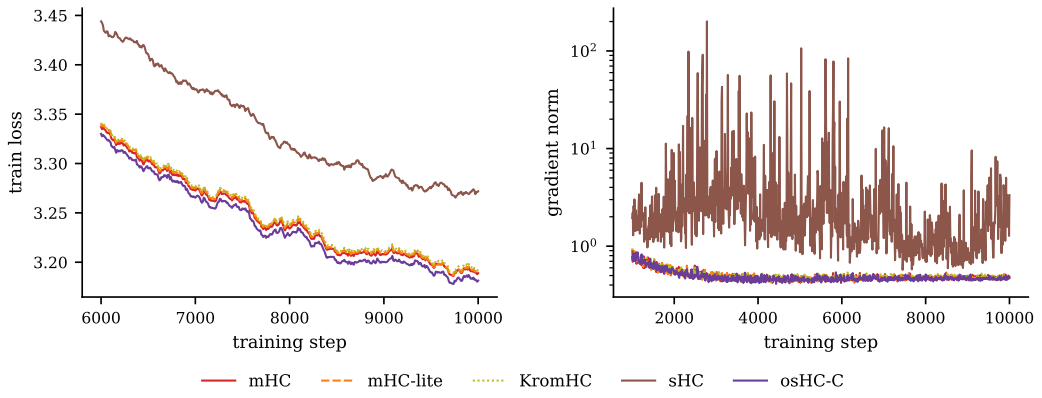


Figure 3: **One marginal against two, during training** (Large, FineWeb-Edu). Training loss on the left, gradient norm on the right, each curve a mean over three seeds. The loss panel is smoothed by an exponential moving average with weight 0.03. The panels start at step 6000 and 1000 respectively, so that the large values of the opening updates do not compress the scale and hide the differences.

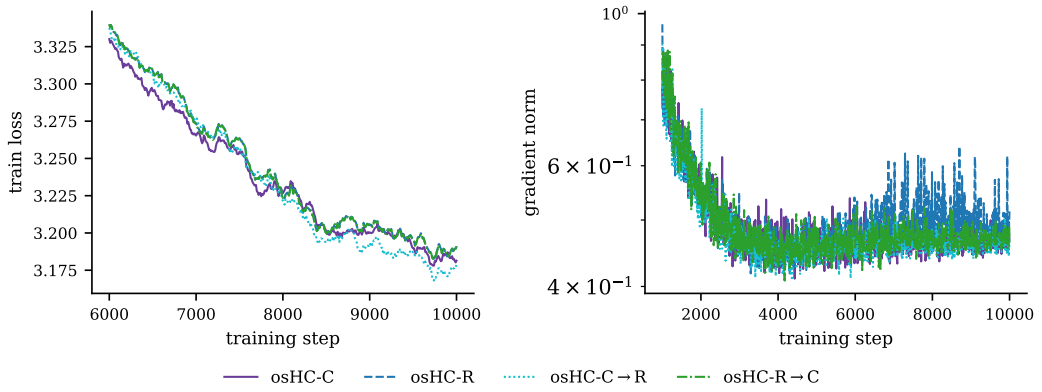


Figure 4: **The four osHC parameterisations against each other** (Large, FineWeb-Edu). Built as in Figure 3, with both y-axes rescaled to the range the four variants occupy.

Both figures show the FineWeb-Edu runs; Appendix C.4 repeats them on OpenWebText.

5.2 Validation loss and zero-shot transfer

One marginal lowers validation loss at nearly every setting. Table 1 reports validation loss across the six dataset–scale cells. In five of the six cells the best and the second-best of the ten methods are both osHC, the only exception being OWT-Medium, where sHC is best and osHC takes the next two places. On FineWeb-Edu the four osHC variants sweep the first four places at all three scales. The single sHC win does not survive scale: at Large it is 0.0600 behind the best osHC on OWT and 0.0933 behind on FineWeb-Edu. Against the two constructions that reach $\mathcal{B}_n$ exactly, mHC-lite and KromHC, the separation is complete: every osHC variant is ahead in all six cells.

The advantage carries over out of domain. The in-domain gaps above are measured on the corpus each model was trained on, so they leave open whether the ordering survives a change of distribution. To test that, we score the Large checkpoints from both pretraining corpora zero-shot on five held-out corpora — C4, Dolma, Falcon RefinedWeb, RedPajama and WikiText-103 — drawn from Paloma (Magnusson et al., 2023), a benchmark built on the premise that perplexity on one held-out set need not predict perplexity on another, and which therefore reports each domain separately. We report word-normalised perplexity; Appendix C.6 gives the protocol.

Table 1: **Full-split validation loss** (lower is better). Mean over three seeds. Bold marks the best mean in each column, underline the second best, ranked before rounding.

Method	OpenWebText			FineWeb-Edu		
	Small	Medium	Large	Small	Medium	Large
Residual	3.5622	3.3453	3.2484	3.5287	3.3220	3.2308
HC	3.4739	3.2771	3.3115	3.4720	3.2780	3.2241
mHC	3.4793	3.2579	3.1934	3.4663	3.2412	3.1816
mHC-lite	3.4787	3.2624	3.1984	3.4630	3.2426	3.1829
KromHC	3.4807	3.2621	3.1957	3.4663	3.2426	3.1851
sHC	3.4857	**3.2512**	3.2447	3.4631	3.2401	3.2649
osHC-C **(ours)**	<u>3.4731</u>	3.2603	<u>3.1847</u>	3.4579	3.2374	<u>3.1737</u>
osHC-R **(ours)**	3.4731	<u>3.2551</u>	3.1904	3.4595	3.2372	3.1795
osHC-C→R **(ours)**	3.4741	3.2562	**3.1847**	**3.4559**	**3.2369**	**3.1716**
osHC-R→C **(ours)**	**3.4699**	3.2595	3.1943	<u>3.4573</u>	<u>3.2370</u>	3.1791

Table 2: **Zero-shot word perplexity on out-of-distribution Paloma corpora** for Large models, pretrained on FineWeb-Edu (left) and OpenWebText (right); lower is better. Means over three seeds; bold marks the best mean in each column, underline the second best.

Method	FineWeb-Edu pretraining					OpenWebText pretraining				
	C4	Dolma	Falcon	RedPajama	WikiText	C4	Dolma	Falcon	RedPajama	WikiText
Residual	130.3	352.0	167.2	2846.6	80.9	173.1	457.2	253.8	2164.6	101.9
HC	130.0	325.8	167.1	2185.4	81.2	201.2	453.1	296.6	1836.5	105.4
mHC	122.1	307.2	156.9	2125.4	74.4	174.0	402.0	254.8	1703.4	93.8
mHC-lite	122.3	307.4	156.9	2057.9	75.0	179.5	432.5	273.6	1983.1	95.4
KromHC	122.6	310.1	157.4	2057.9	75.1	180.0	421.4	274.3	1892.7	92.6
sHC	135.4	337.7	174.1	2233.6	84.0	175.5	424.8	255.9	1872.5	98.4
osHC-C	<u>120.8</u>	**296.1**	<u>154.8</u>	<u>1927.6</u>	**72.4**	<u>170.0</u>	**387.4**	<u>247.8</u>	<u>1649.4</u>	**87.7**
osHC-R	121.8	299.4	156.3	1947.5	73.6	**165.8**	<u>389.1</u>	**242.0**	**1630.3**	89.5
osHC-C→R	**120.3**	<u>296.6</u>	**154.7**	**1892.1**	<u>72.7</u>	172.1	409.7	257.2	1819.3	<u>89.4</u>
osHC-R→C	121.8	300.7	156.2	1965.4	73.4	174.6	420.3	256.8	1858.4	96.2

Table 2 reports the results on both pretraining corpora. osHC holds the lowest and the second-lowest perplexity in all ten columns. On FineWeb-Edu every osHC variant is ahead of every baseline on every corpus, and the two best methods, osHC-C and osHC-C→R, reduce word perplexity over mHC by 1.06%–9.31% and 1.40%–10.98% respectively. On OWT the two pure variants lead: osHC-C and osHC-R reduce it by 2.30%–6.50% and 3.21%–5.02%.

5.3 Routing diagnostics

Trained osHC routers do replicate and aggregate. We feed two 256-token validation sequences through each pretrained checkpoint and record the residual mixer H_l^{res} at every attention and MLP mixing site. The two profiles that balance flattens are directly measurable, and we summarise each by its largest entry: the *replication factor* $\mathrm{Rep}(H) := \max_j s_j(H) = \|H\|_1$, the largest total weight any one source sends out, and the *aggregation factor* $\mathrm{Agg}(H) := \max_i d_i(H) = \|H\|_\infty$, the largest total any one destination takes in. Balance pins both at 1; each one-sided family frees exactly one.

Figure 5 plots the cumulative distribution of the larger of the two factors, for exactly doubly stochastic mixers and for osHC mixers. When H is doubly stochastic that value is 1 at every site; under osHC it exceeds 1 at 96% of osHC-R sites and at 99% of osHC-C sites. Averaged over sites, Rep is 1.280 under osHC-R and Agg is 1.441 under osHC-C. At the extreme, osHC-R reaches $\mathrm{Rep} = 3.149$ and osHC-C reaches $\mathrm{Agg} = 3.987$, against a structural maximum of $n = 4$ for either factor.

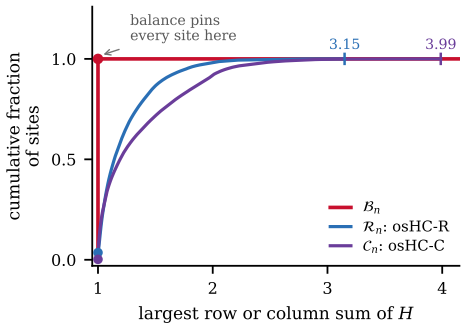
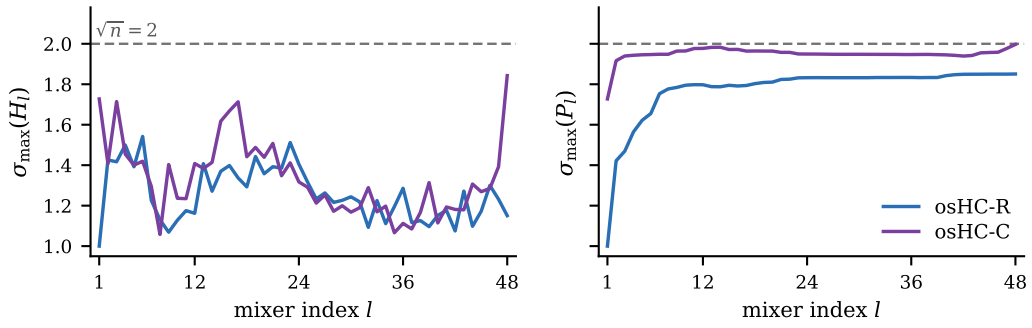


Figure 5: **How far trained routers depart from balance.** Cumulative distribution of the largest row or column sum of H, one value per mixer. It equals 1 exactly when H is doubly stochastic, so the two exact constructions, mHC-lite and KromHC, never leave 1.



(a) Single-layer mixer

(b) Composite mixer

Figure 6: **Spectral norm of the single-layer and composite mixers.** Panel (a) is the mixer H_l acting alone at each residual-mixing site (attention or MLP), panel (b) the composite mixer P_l built from all sites up to l. Each point plotted is the maximum over the probed tokens and the six checkpoints (two corpora by three seeds). The two switch schedules, osHC-C→R and osHC-R→C, whose bound on the composite mixer is $n = 4$ instead, are in Appendix Figure 9.

The measured gain stays below its ceiling at every depth. To test whether the Euclidean gain of the osHC residual mixer is bounded, and bounded independently of depth, we take the mixers recorded above for the Large checkpoints. At each site l we compute the largest singular value $\sigma_{\max}$ of H_l^{res} and of the composite mixer $P_l = \prod_{k=1}^{l} H_{l+1-k}^{\text{res}}$, plotted against l in Figure 6. In both panels every curve stays below the ceiling proved for it in Theorems 4.1 and 4.2, at every depth.

6 Conclusion

Balance pins both the row sums and the column sums of H. Pinning either one alone already holds the gain of the composite mixer to $\sqrt{n}$ at any depth, so a single marginal buys the depth-uniform stability the constraint was introduced for. Pinning the second one as well tightens that bound from $\sqrt{n}$ to 1, and costs the router the ability to replicate a source or to aggregate at a destination, together with the ability to make the streams more different from one another. One-Sided Hyper-Connections keep one marginal and release the other. Every osHC variant reaches a lower validation loss than every construction that is exactly doubly stochastic, on both pretraining corpora and at all three scales, and in the trained models the released capability is used at almost every routing site.

Limitations. Our evidence covers $n = 4$ streams, models up to 354M parameters, 10,000 updates, and decoder-only language modelling; whether the ordering survives wider streams or longer training is untested. Nor do we know how it behaves at substantially larger model scale.

AI USE STATEMENT

We used generative AI tools as writing and coding assistants: for copy-editing and condensing prose that we drafted, and for routine implementation and refactoring of analysis scripts. The research question, the theoretical results and their proofs, the experimental design, and all claims made in this paper are our own. Every proof was verified by hand and, where stated, cross-checked numerically by the verification scripts released with the code; all AI-assisted code was reviewed and tested by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies which constraint to impose on the residual mixer of a hyper-connection architecture and uses only publicly available pretraining corpora (OpenWebText, FineWeb-Edu) and public evaluation suites. It involves no human subjects, no new data collection, and no personally identifying information. The models we train are small relative to deployed language models and are released only as research artifacts. We report all runs we conducted, including the unstable ones, rather than selecting them away.

REPRODUCIBILITY STATEMENT

The training protocol, hyperparameters, checkpoint-selection rule, and full-split evaluation procedure are given in Appendix C, which also lists every one of the 180 runs individually in Table 3 rather than as cell means, so that any aggregation other than the one we report can be recomputed. Proofs of all theoretical statements are in Appendix B. Anonymous source code for the models, the training configurations, and the routing-diagnostic pipeline is submitted as supplementary material.

REFERENCES

Ekaterina Alimaskina, Gleb Molodtsov, and Aleksandr Beznosikov. Analyzing stream collapse in hyper-connections: From diagnosis to mitigation. In *ICML Workshop on Weight-Space Symmetries*, 2026.

Garrett Birkhoff. Tres observaciones sobre el algebra lineal. *Universidad Nacional de Tucumán, Revista A*, 5:147–151, 1946.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006. doi: 10.1109/TIT.2006.874516.

Chen Cai and Yusu Wang. A note on over-smoothing for graph neural networks. *arXiv preprint arXiv:2006.13318*, 2020.

Torque Dandachi and Sophia Diggs-Galligan. go-mhc: Direct parameterization of manifold-constrained hyper-connections via generalized orthostochastic matrices. *arXiv preprint arXiv:2604.02309*, 2026.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *European Conference on Computer Vision (ECCV)*, pp. 630–645, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016b.

Zhaoyi Liu, Haichuan Zhang, and Ang Li. Beyond the birkhoff polytope: Spectral-sphere-constrained hyper-connections. *arXiv preprint arXiv:2603.20896*, 2026.

Anton Lyubinin. Tbp-mhc: Full expressivity for manifold-constrained hyper-connections through transportation polytopes. *arXiv preprint arXiv:2605.21724*, 2026.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A. Smith, Kyle Richardson, and Jesse Dodge. Paloma: A benchmark for evaluating language model fit. *arXiv preprint arXiv:2312.10523*, 2023.

Albert W. Marshall, Ingram Olkin, and Barry C. Arnold. *Inequalities: Theory of Majorization and Its Applications*. Springer, 2nd edition, 2011.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations (ICLR)*, 2020.

William Gao Peng, Josheev Rai, Kevin Tseng, Siwei Wang, and Sean Wu. Ablate and rescue: A causal analysis of residual stream hyper-connections. *arXiv preprint arXiv:2603.14833*, 2026.

T. Konstantin Rusch, Michael M. Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151, pp. 3515–3530, 2022.

E. Seneta. *Non-negative Matrices and Markov Chains*. Springer, 2nd edition, 1981.

Biswa Sengupta, Jinhua Wang, and Léo Brunswic. Jpmhc: Dynamical isometry via orthogonal hyper-connections. *arXiv preprint arXiv:2602.18308*, 2026.

Arash Shahmansoori. The EΔ-MHC-Geo transformer: Adaptive geodesic operations with guaranteed orthogonality. *arXiv preprint arXiv:2605.06729*, 2026.

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

Chenrui Wang and Yixuan Qiu. Accelerating birkhoff projection for manifold-constrained hyper-connections. *arXiv preprint arXiv:2606.07574*, 2026.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. mhc: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025.

Yongyi Yang and Jianyang Gao. mhc-lite: You don't need 20 sinkhorn-knopp iterations. *arXiv preprint arXiv:2601.05732*, 2026.

Wuyang Zhou, Yuxuan Gu, Giorgos Iacovides, and Danilo Mandic. Kromhc: Manifold-constrained hyper-connections with kronecker-product residual matrices. *International Conference on Machine Learning (ICML)*, 2026.

Defa Zhu, Hongzhi Huang, Zihao Huang, Yutao Zeng, Yunyao Mao, Banggu Wu, Qiyang Min, and Xun Zhou. Hyper-connections. *arXiv preprint arXiv:2409.19606*, 2024.

Defa Zhu, Hongzhi Huang, Jundong Zhou, Zihao Huang, Yutao Zeng, Banggu Wu, Qiyang Min, and Xun Zhou. Frac-connections: Fractional extension of hyper-connections. *arXiv preprint arXiv:2503.14125*, 2025.

A Extended Related Work

This appendix supplies the detail deferred by Section 2.

Scope of the comparison. Throughout, our claims concern H^{res}, the only factor that composes across depth. Orthogonal to the constraint question, Frac-Connections (Zhu et al., 2025) reduce the memory cost of widening the residual stream rather than changing what the mixer is allowed to be, so the analysis applies to them unchanged.

Why the three relaxations do not reach the released mechanism. sHC (Liu et al., 2026) is the closest to ours: it reports that Birkhoff-constrained matrices concentrate near the identity during training and that non-negativity precludes subtractive mixing — we adopt both observations — and moves H onto the spectral sphere, keeping both marginals and dropping non-negativity. JPmHC (Sengupta et al., 2026) drops both marginals instead, requiring $H^\top H = I_n$ on the Stiefel manifold. EΔ-MHC-Geo (Shahmansoori, 2026) gates a data-dependent Cayley rotation against a Householder reflection to reach both components of $\mathrm{O}(n)$.

None of the three releases a marginal, and none therefore reaches the mechanism we study. The proof of Theorem 3.2 never uses $H \geq 0$, so it holds verbatim on the signed set sHC inhabits; orthogonal mixers are non-expansive by construction, and a gate between two orthogonal operators is a convex combination of them, hence non-expansive as well. The coverage is no coincidence, because $\{\|H\|_2 \leq 1\}$ is exactly $\mathrm{conv}\,\mathcal{O}(n)$ (Appendix B): any mixer assembled from orthogonal primitives by convex combination already satisfies the Euclidean bound. Among these constraints, relaxing a marginal is the only route that lets routing exceed it.

Empirical analyses of trained models. Two studies probe HC models directly rather than proposing a constraint. Peng et al. (2026) release an open mHC language model and introduce an ablation-and-rescue protocol that separates functional redundancy from asymmetric utilisation, finding that streams are not interchangeable even when their representations are highly similar; we adopt that caution when comparing osHC-R against osHC-C. Alimaskina et al. (2026) report *stream collapse* — signal, representation norm and interpretable features concentrating in one dominant stream, with residual mixing staying near the identity after an early seeding phase — and mitigate it by breaking the stream-permutation symmetry at the expansion interface. Their collapse is not the contraction of Theorem 3.2; in our coordinates it is the opposite extreme, since a dominant-stream configuration, $X = \mathbf{e}_1\mathbf{c}^\top$, has $\mathcal{E}(X)/\|X\|_F^2 = 1 - 1/n$, the largest value attainable, whereas contraction drives $\mathcal{E}$ to zero. Their near-identity finding, which Liu et al. (2026) also report, is the more directly relevant of the two: it indicates that the mixing operator is largely idle in trained mHC models, and Theorem 3.2 suggests one reason. Inside $\mathcal{B}_n$, exact preservation of disagreement is confined to non-generic matrices — permutations, and block forms such as $\mathrm{diag}(J_2/2,\, I_{n-2})$ — while generic motion away from them contracts, $(1-t)I_n + tJ/n$ giving $\sigma_2 = 1 - t$ (Proposition B.5(iv)). A model that must not lose stream diversity therefore has little room to move inside $\mathcal{B}_n$.

Doubly stochastic constraints elsewhere. The contraction studied here is classical rather than specific to HC. Doubly stochastic multiplication moves a vector downward in the majorization order (Marshall et al., 2011); equivalently, its Birkhoff–von Neumann decomposition averages permutations and cannot increase centered Euclidean spread. Distributed averaging and gossip algorithms exploit this behaviour to drive agents towards consensus (Boyd et al., 2006), where the contraction is the objective rather than a defect.

The same design choice appears in attention, in the opposite direction. Sinkformers (Sander et al., 2022) replace the row-stochastic softmax of attention with Sinkhorn normalisation, moving attention from one marginal to two, and report a gain, whereas we argue for the reverse in residual routing. The two are consistent because attention mixes tokens inside a layer while residual routing composes along a fixed index set, so a per-layer contraction accumulates there and nowhere else. Their case rests on two supports, which are worth separating. The interpretation of the resulting attention layer as a Wasserstein gradient flow is intrinsic to attention and transfers to nothing here. The empirical support is the interesting one, and its analogue in our setting points the other way: Sinkformers observe that the *unconstrained* optimum drifts towards double stochasticity, which is a reason to impose it, whereas for residual mixing what has been reported is that the *constrained* optimum

moves towards the degenerate corner of $\mathcal{B}_n$ rather than its interior (Liu et al., 2026; Alimaskina et al., 2026). The two observations are made on different objects, so this is a contrast rather than a proof. It does mean the empirical argument that recommends $\mathcal{B}_n$ for attention has no counterpart for residual routing, and that the closest available evidence runs against it.

Diversity contraction and over-smoothing. Our disagreement energy $\mathcal{E}(X) = \|\Pi X\|_F^2$ is, up to the factor n, the complete-graph Dirichlet energy $\frac{1}{2}\sum_{i,j}\|X_i - X_j\|_2^2$ of the streams, and Theorem 3.2 is a statement of the kind studied extensively for graph neural networks (Oono & Suzuki, 2020; Cai & Wang, 2020; Rusch et al., 2023). Two differences matter. First, the graph results are asymptotic in depth and require conditions on the propagation operator, whereas Theorem 3.2 is exact, per-layer, and assumption-free on $\mathcal{B}_n$. Second, the graph results apply equally to row-stochastic propagation, so one-sided constraints do not by themselves prevent over-smoothing. We confirm this numerically: for random row-stochastic matrices with dense positive entries, the composed map drives $\mathcal{E}$ below 10^{-6} of its input value within twelve layers.

Norm control in deep networks. Bounding the spectral norm of learned operators is a standard stabilisation device, from spectral normalisation (Miyato et al., 2018) to Lipschitz- and orthogonality-constrained layers. mHC is commonly read as a member of this family, but the reading is imprecise in a way that matters here: on $\mathcal{B}_n$ the spectral norm is not bounded by 1 but equal to 1 for every feasible matrix (Eq. (13)). There is consequently no norm budget to allocate, and $\sigma_{\max}$ is uninformative for an exactly doubly stochastic model. Under osHC the budget is nontrivial and observable: $\sigma_{\max}$ lies between $\mathrm{Rep}(H)/\sqrt{n}$ and $\sqrt{\mathrm{Rep}(H)}$, so the replication factor the network selects is visible in its Euclidean gain. The classical apparatus for the composed behaviour of one-sided stochastic products is the theory of ergodicity coefficients (Seneta, 1981), which bounds the contraction rate of such products under positivity assumptions that trained matrices need not satisfy; relating those bounds to Rep is a natural direction we do not pursue here.

B Deferred Statements and Proofs

B.1 The transfer principle

Theorem 4.2 states its conclusion for $\mathcal{R}_n$ only. The following lemma is what licenses that convention.

Lemma B.1 (Transfer principle)**.** *Transposition $H \mapsto H^\top$ is an involution carrying $\mathcal{R}_n$ onto $\mathcal{C}_n$ and $\mathcal{B}_n$ onto itself. It preserves the Euclidean gain,*

$$\|H^\top\|_2 = \|H\|_2, \tag{12}$$

and exchanges the endpoint norms, $\|H^\top\|_1 = \|H\|_\infty$ and $\|H^\top\|_\infty = \|H\|_1 = \mathrm{Rep}(H)$.

Proof. The map is involutive, and $H \geq 0$ is transpose-invariant while $H\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top H = \mathbf{1}^\top$ exchange, giving the stated action on $\mathcal{R}_n$, $\mathcal{C}_n$ and $\mathcal{B}_n$. Transposition preserves singular values, hence $\|H^\top\|_2 = \|H\|_2$. The endpoint identities are the standard duality $\|A^\top\|_1 = \|A\|_\infty$ together with $\mathrm{Rep} = \|\cdot\|_1$ and $\mathrm{Agg} = \|\cdot\|_\infty$, which give $\mathrm{Rep}(H^\top) = \mathrm{Agg}(H)$ and $\mathrm{Agg}(H^\top) = \mathrm{Rep}(H)$. □

B.2 Geometry of the directions removed by the second marginal

Proposition B.2 (Geometry of the removed directions)**.**

(i) $\mathcal{B}_n = \mathcal{R}_n \cap \mathcal{C}_n$ *and the inclusions* $\mathcal{B}_n \subsetneq \mathcal{R}_n$, $\mathcal{B}_n \subsetneq \mathcal{C}_n$ *are strict for all* $n \geq 2$.

(ii) $\dim \mathcal{B}_n = (n-1)^2$ *while* $\dim \mathcal{R}_n = \dim \mathcal{C}_n = n(n-1)$, *so* $\dim \mathcal{R}_n - \dim \mathcal{B}_n = \dim \mathcal{C}_n - \dim \mathcal{B}_n = n-1$. *The* $n-1$ *removed degrees of freedom are exactly the free entries of the supply profile* $\mathbf{s}(H)$ *(resp. the demand profile* $\mathbf{d}(H)$*) on the simplex* $\{\mathbf{1}^\top \mathbf{s} = n\}$.

(iii) The extreme points of $\mathcal{B}_n$ *are the* $n!$ *permutation matrices (Birkhoff–von Neumann theorem), whereas the extreme points of* $\mathcal{R}_n$ *are the* n^n *row-wise one-hot matrices. Every permutation matrix is row-wise one-hot, but not conversely.*

(iv) The unique rank-one element of $\mathcal{B}_n$ is the uniform averaging matrix J/n, whereas $\mathcal{R}_n$ contains the entire $(n-1)$-dimensional family of selective broadcast *maps $\{\mathbf{1}\mathbf{c}^\top : \mathbf{c} \geq 0,\ \mathbf{1}^\top\mathbf{c} = 1\}$. Dually, $\mathcal{C}_n$ contains the family of* selective aggregation *maps $\{\mathbf{r}\mathbf{1}^\top : \mathbf{r} \geq 0,\ \mathbf{1}^\top\mathbf{r} = 1\}$.*

Proof. (i) The inclusion is Definition 3.1; strictness follows from (iv). (ii) $\mathcal{R}_n$ is a product of n simplices Δ^{n-1}, of dimension $n(n-1)$; $\dim \mathcal{B}_n = (n-1)^2$ is classical, the $2n$ marginal equations having exactly one linear dependency. (iii) $\mathcal{R}_n$ is a product of simplices, whose extreme points are products of vertices, i.e. row-wise one-hot matrices; the statement for $\mathcal{B}_n$ is the Birkhoff–von Neumann theorem. (iv) If $H = \mathbf{u}\mathbf{v}^\top \in \mathcal{B}_n$ then $\mathbf{u}(\mathbf{v}^\top\mathbf{1}) = \mathbf{1}$ forces $\mathbf{u} \propto \mathbf{1}$, and $(\mathbf{1}^\top\mathbf{u})\mathbf{v}^\top = \mathbf{1}^\top$ forces $\mathbf{v} \propto \mathbf{1}$; normalising gives $H = J/n$. Conversely $\mathbf{1}\mathbf{c}^\top \in \mathcal{R}_n$ for any $\mathbf{c}$ in the simplex, and $\mathbf{1}\mathbf{c}^\top \in \mathcal{B}_n$ only for $\mathbf{c} = \mathbf{1}/n$. □

B.3 Interpolation and endpoint norms

Proposition B.3 (One marginal fixes one endpoint norm)**.** *Let $H \geq 0$.*

(i) If $H \in \mathcal{R}_n$, then $\mathrm{Agg}(H) = \|H\|_\infty = 1$ and every output is a convex combination of the inputs.

(ii) If $H \in \mathcal{C}_n$, then $\mathrm{Rep}(H) = \|H\|_1 = 1$ and aggregate stream mass is conserved, $\mathbf{1}^\top HX = \mathbf{1}^\top X$.

(iii) $\mathcal{R}_n$, $\mathcal{C}_n$, and $\mathcal{B}_n$ are closed under matrix multiplication.

(iv) Either unit marginal forces $\|H\|_2 \geq 1$, without a sign assumption.

(v) Riesz–Thorin gives $\|H\|_p \leq \|H\|_1^{1/p}\|H\|_\infty^{1-1/p}$, whose $p = 2$ case is the joint bound Eq. (10) of the main text. Balance sets both factors to 1, which with (iv) gives

$$H \in \mathcal{B}_n \quad \Longrightarrow \quad \|H\|_2 = 1. \tag{13}$$

(vi) On $\mathcal{R}_n$, $\mathrm{Agg} \equiv 1$, so Eq. (10) sharpens to

$$\max\left\{1, \frac{\mathrm{Rep}(H)}{\sqrt{n}}\right\} \leq \|H\|_2 \leq \sqrt{\mathrm{Rep}(H)}, \qquad \|H\|_p \leq \mathrm{Rep}(H)^{1/p}, \tag{14}$$

and dually on $\mathcal{C}_n$ with Agg *in place of* Rep.

(vii) Each free factor certifies departure from balance on its own side: $\mathrm{Rep}(H) = 1 \iff H \in \mathcal{B}_n$ *for $H \in \mathcal{R}_n$, and* $\mathrm{Agg}(H) = 1 \iff H \in \mathcal{B}_n$ *for $H \in \mathcal{C}_n$. The two are not interchangeable in effect. For $H \in \mathcal{C}_n$,* $\mathrm{Agg}(H) > 1$ *forces the row sums to be unequal and hence $\mu(H) > 0$, whereas* $\mathrm{Rep}(H) > 1$ *on $\mathcal{R}_n$ leaves $H\mathbf{1} = \mathbf{1}$ and so $\mu(H) = 0$.*

Proof of Proposition B.3. (i) $\|H\|_\infty$ is the maximum absolute row sum, which equals 1 by $H \geq 0$ and $H\mathbf{1} = \mathbf{1}$; the convex-combination statement is immediate from $H_{ij} \geq 0$, $\sum_j H_{ij} = 1$. (ii) is dual, with $\|H\|_1$ the maximum absolute column sum. (iii) If $H_1\mathbf{1} = \mathbf{1}$ and $H_2\mathbf{1} = \mathbf{1}$ then $H_2H_1\mathbf{1} = H_2\mathbf{1} = \mathbf{1}$, and $H_2H_1 \geq 0$; dually for $\mathcal{C}_n$; intersecting gives $\mathcal{B}_n$. (iv) $\|H\|_2 \geq \|H\mathbf{1}\|_2/\|\mathbf{1}\|_2 = 1$, and dually $\mathbf{1}^\top H = \mathbf{1}^\top$ gives $\|H^\top\mathbf{1}\|_2 = \|\mathbf{1}\|_2$. (v) Riesz–Thorin at $p = 2$ with the endpoint norms $\|H\|_1$ and $\|H\|_\infty$ is Eq. (10); on $\mathcal{B}_n$ both are 1, so $\|H\|_2 \leq 1$, which combined with (iv) is an equality. (vi) Substitute $\mathrm{Agg}(H) = 1$ into Eq. (10) for the upper bounds; the lower bound is (iv) together with $\|H\|_1 \leq \sqrt{n}\,\|H\|_2$ for $n \times n$ matrices. (vii) For $H \in \mathcal{R}_n$ the entries sum to n, so the n column sums average 1; if the largest is 1 then all are 1 and $H \in \mathcal{C}_n$, and the converse is (ii). The $\mathcal{C}_n$ statement is its transpose by Lemma B.1. For the second claim, $H \in \mathcal{C}_n$ again makes the row sums average 1, so $\mathrm{Agg}(H) > 1$ leaves them unequal, $H\mathbf{1} \notin \mathrm{span}(\mathbf{1})$, and $\mu(H) = \|\Pi H\mathbf{1}\|_2 > 0$; on $\mathcal{R}_n$ the definition gives $H\mathbf{1} = \mathbf{1}$ and $\mu(H) = \|\Pi\mathbf{1}\|_2 = 0$ whatever $\mathrm{Rep}(H)$ is. □

Equation 13 shows that the spectral norm of an exact balanced mixer is fixed, rather than merely upper-bounded, at one, so a balanced model has no Euclidean budget to allocate. Equation 10 locates what each one-sided family buys in a single inequality: the product $\mathrm{Rep} \cdot \mathrm{Agg}$ bounds the

Euclidean gain in all three cases, balance pins both factors at 1, and each one-sided family releases exactly one of them. That is the precise sense in which $\mathcal{R}_n$ and $\mathcal{C}_n$ are mirror images, and Eq. (14) is the row-side reading of it. Part (vii) is where the mirror stops. The two factors are exchanged by transposition and each certifies departure from $\mathcal{B}_n$, but only the aggregation factor is tied to the consensus channel, because only it reads the marginal that decides whether $\mathbf{1}$ is fixed.

B.4 Proofs of the main-text theorems

Proof of Theorem 3.2. The two marginals enter one step each. The column condition $\mathbf{1}^\top H = \mathbf{1}^\top$ gives $JH = \mathbf{1}(\mathbf{1}^\top H) = \mathbf{1}\mathbf{1}^\top = J$, the row condition $H\mathbf{1} = \mathbf{1}$ gives $HJ = (H\mathbf{1})\mathbf{1}^\top = \mathbf{1}\mathbf{1}^\top = J$, and substituting these into $\Pi = I_n - J/n$ on each side makes H commute with the projector:

$$\Pi H = H - \tfrac{1}{n}JH = H - \tfrac{1}{n}J = H - \tfrac{1}{n}HJ = H\Pi.$$

Next, since $H \geq 0$, the same two conditions give $\|H\|_1 = \max_j \sum_i |H_{ij}| = \max_j \sum_i H_{ij} = 1$ and $\|H\|_\infty = \max_i \sum_j |H_{ij}| = \max_i \sum_j H_{ij} = 1$, so Riesz–Thorin at $p = 2$, Proposition B.3(v), gives $\|H\|_2 \leq \sqrt{\|H\|_1 \|H\|_\infty} = 1$. Combining the two, and using $\|AB\|_F \leq \|A\|_2 \|B\|_F$,

$$\mathcal{E}(HX) = \|\Pi H X\|_F^2 = \|H\,\Pi X\|_F^2 \leq \|H\|_2^2\, \|\Pi X\|_F^2 \leq \|\Pi X\|_F^2 = \mathcal{E}(X). \tag{15}$$

The bound applies to a composed path by induction, since $\mathcal{B}_n$ is closed under composition (Proposition B.3(iii)). □

B.5 Hybrid schedules

Theorem B.4 (The hybrid ceiling is set by switches, not depth)**.** *Let $H_1, \ldots, H_L \in \mathcal{R}_n \cup \mathcal{C}_n$ and $P := H_L \cdots H_1$. Partition the sequence into m maximal runs drawn from the same family, so $m-1$ is the number of switches. Then*

$$\|P\|_2 \leq n^{m/2}, \tag{16}$$

independently of L, and the bound is attainable.

Proof of Theorem B.4. Each maximal run is a product of factors from one family, and both families are closed under multiplication (Proposition B.3(iii)), so the run lies in that family. A run in $\mathcal{R}_n$ has $\|\cdot\|_\infty = 1$ and $\|\cdot\|_1 = \mathrm{Rep} \leq n$; a run in $\mathcal{C}_n$ has $\|\cdot\|_1 = 1$ and $\|\cdot\|_\infty \leq n$ by Lemma B.1. Either way Riesz–Thorin gives a run norm of at most $\sqrt{n}$, and submultiplicativity across the m runs yields the bound. For attainment, $R_0 \in \mathcal{R}_n$ and $C_0 \in \mathcal{C}_n$ are immediate, and $C_0 R_0 = \mathbf{e}_1(\mathbf{1}^\top \mathbf{1})\mathbf{e}_1^\top = n\,\mathbf{e}_1\mathbf{e}_1^\top$. Since $\mathbf{e}_1\mathbf{e}_1^\top$ is idempotent, $(C_0R_0)^k = n^k \mathbf{e}_1\mathbf{e}_1^\top$, whose spectral norm is n^k, attaining the bound for $m = 2k$. Since $R_0(C_0R_0)^k = n^k R_0$ and $\|R_0\|_2 = \sqrt{n}$, the bound is also attained for $m = 2k+1$. □

For tightness, let $R_0 := \mathbf{1}\mathbf{e}_1^\top \in \mathcal{R}_n$ and $C_0 := \mathbf{e}_1\mathbf{1}^\top \in \mathcal{C}_n$. Then $C_0R_0 = n\mathbf{e}_1\mathbf{e}_1^\top$, so $(C_0R_0)^k$ attains n^k with $m = 2k$, while $R_0(C_0R_0)^k = n^k R_0$ attains $n^{k+1/2}$ with $m = 2k+1$. Thus the bound is tight for every m. A pure schedule has $m = 1$ and recovers Theorem 4.2; a single-switch block schedule has $m = 2$ and ceiling n. Alternating families at every layer has $m = L$ and therefore loses depth-uniform control. This is why our hybrid variants use one contiguous block of each family rather than layer-wise alternation.

B.6 Strict dissipation inside the Birkhoff polytope

Proposition B.5 (Balanced routing is strictly dissipative in the interior)**.** *Let $H \in \mathcal{B}_n$.*

- *(i) H preserves both $\mathrm{span}(\mathbf{1})$ and $\mathbf{1}^\perp$, and $\mathcal{E}(HX) \leq \sigma_2(H)^2\, \mathcal{E}(X)$ for every X, where $\sigma_2(H)$ is the second largest singular value of H.*
- *(ii) If $H > 0$ entrywise, then $\sigma_2(H) < 1$ and $\mathcal{E}(HX) < \mathcal{E}(X)$ for every non-consensus X, that is every X with $\mathcal{E}(X) > 0$.*
- *(iii) The set $\{H \in \mathcal{B}_n : \sigma_2(H) = 1\}$ lies on the relative boundary of $\mathcal{B}_n$ and has measure zero within it. It includes all permutation matrices.*

(iv) Along $H_t = (1-t)I_n + tJ/n$, one has $\sigma_2(H_t) = 1 - t$.

Proof of Proposition B.5. (i) $H\mathbf{1} = \mathbf{1}$ gives invariance of $\mathrm{span}(\mathbf{1})$ and shows H acts as the identity there. For $\mathbf{v} \perp \mathbf{1}$, the second marginal gives $\mathbf{1}^\top H\mathbf{v} = (H^\top\mathbf{1})^\top\mathbf{v} = \mathbf{1}^\top\mathbf{v} = 0$, so $\mathbf{1}^\perp$ is invariant as well. Thus H is block diagonal with respect to the orthogonal decomposition $\mathbb{R}^n = \mathrm{span}(\mathbf{1}) \oplus \mathbf{1}^\perp$, and its singular values are $\{1\}$ together with those of the restriction $H|_{\mathbf{1}^\perp}$. By Eq. (13) the largest is 1, so $\|H|_{\mathbf{1}^\perp}\|_2 = \sigma_2(H)$. Block diagonality also makes H commute with Π, and the columns of ΠX lie in $\mathbf{1}^\perp$, so $\mathcal{E}(HX) = \|H\Pi X\|_F^2 \le \|H|_{\mathbf{1}^\perp}\|_2^2 \, \|\Pi X\|_F^2 = \sigma_2(H)^2 \, \mathcal{E}(X)$.

(ii) $H^\top H \in \mathcal{B}_n$ by Proposition B.3(iii); it is symmetric, and entrywise positive whenever $H > 0$. Its row sums are 1, so its spectral radius is 1, attained at $\mathbf{1}$. Perron–Frobenius applied to a positive matrix makes the spectral radius a *simple* eigenvalue, so every other eigenvalue of $H^\top H$ is strictly below 1; hence $\sigma_2(H) < 1$. Strict decrease of $\mathcal{E}$ then follows from (i): $\mathcal{E}(HX) \le \sigma_2(H)^2\mathcal{E}(X) < \mathcal{E}(X)$ whenever $\mathcal{E}(X) > 0$.

(iii) By (ii), $\sigma_2(H) = 1$ forces at least one entry of H to vanish, and $\{H \in \mathcal{B}_n : H_{ij} = 0 \text{ for some } i, j\}$ is the relative boundary of $\mathcal{B}_n$, a finite union of proper faces and hence of measure zero in $\mathcal{B}_n$. A permutation matrix P satisfies $P^\top P = I_n$ and leaves $\mathbf{1}^\perp$ invariant by (i), so $\sigma_2(P) = \|P|_{\mathbf{1}^\perp}\|_2 = 1$.

(iv) $\Pi H_t \Pi = (1-t)\Pi + \frac{t}{n}\Pi J\Pi = (1-t)\Pi$, whose spectral norm is $1 - t$; by (i) this spectral norm is $\sigma_2(H_t)$. □

Thus a positive balanced mixer necessarily dissipates stream disagreement. Boundary matrices can preserve selected disagreement directions, but doing so requires degenerate support; permutations preserve all disagreement while performing only a relabeling.

C Extended Experiments

C.1 Training details

All models use sequence length 1024, microbatch size 16, eight gradient accumulation steps, and hence 131,072 tokens per optimiser update. We train for 10,000 updates with AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, gradient clipping at 1.0, 200 warmup updates, cosine decay, bfloat16 arithmetic, and no dropout. The peak/minimum learning rates are respectively $10^{-3}/10^{-4}$ (Small), $6 \cdot 10^{-4}/6 \cdot 10^{-5}$ (Medium), and $3 \cdot 10^{-4}/3 \cdot 10^{-5}$ (Large). The three architectures have $(L, d, h) = (6, 512, 8), (12, 768, 12), (24, 1024, 16)$ and approximately 45M, 125M, and 354M parameters. All Hyper-Connection variants use four streams and are parameter-matched within a scale. Runs use seeds 0, 1, and 1337.

Checkpoint selection and full-split evaluation. Training evaluates 200 randomly sampled validation minibatches every 500 updates and retains the checkpoint with the lowest such estimate. After training, that selected checkpoint is evaluated once on every non-overlapping 1024-token validation window: 4,330 windows (4,433,920 tokens) for OWT and 48,195 windows (49,351,680 tokens) for FineWeb-Edu. Selection is not a formality at this budget: in 72 of the 180 runs the retained checkpoint is not the final one, falling at update 9,500 in 60 runs, 9,000 in 11, and 8,500 in one. Thus the reported number is deterministic conditional on the selected checkpoint, but checkpoint selection uses the same validation split. This protocol is identical across methods; we nevertheless record the resulting selection bias as a limitation and will use a disjoint selection split in larger-scale follow-up experiments.

C.2 Reading the main table

Table 1 marks the lowest mean in each column in bold and the second lowest by underline, ranking on the unrounded means. Three columns have a gap of $3.3 \cdot 10^{-5}$ across one of those two boundaries: between first and second on OpenWebText Large and on FineWeb-Edu Medium, and between second and third on OpenWebText Small. That number is not a coincidence. Per-seed losses are logged to four decimals, so a three-seed mean is quantised in steps of $\frac{1}{3} \cdot 10^{-4} = 3.3 \cdot 10^{-5}$, and these gaps are the smallest nonzero difference the logs can express. In two of the three columns

the two entries even print the same four decimals. The highlighting therefore records a rank, not a separation.

Every method is trained at the same three seeds on the same corpora and scales, sharing optimiser, schedule, and data pipeline, so cells of Table 1 differ only in the residual mixer. We report the individual runs in Table 3 rather than only their summaries, so that a reader who prefers a different aggregation can compute it.

C.3 Per-scale results

Table 1 reports cell means; Table 3 lists all 180 full-split validation losses individually, one per method, corpus, scale, and seed. It also makes the two instabilities legible as single-seed events: the underlined entries are the three runs whose loss exceeds that of their two siblings by more than 0.03 nats, one for HC and two for sHC, and no other cell in the study contains one.

Table 3: Complete per-seed full-split validation losses. Each cell lists seeds $0/1/1337$ in that order. The final four rows are our methods. Underlining marks a seed whose loss exceeds the larger of its two siblings by more than 0.03 nats; all three such runs are retained rather than discarded, and are discussed in Section 5.

	OpenWebText			FineWeb-Edu		
Method	Small	Medium	Large	Small	Medium	Large
Residual	3.5623/3.5644/3.5599	3.3502/3.3443/3.3414	3.2462/3.2560/3.2431	3.5238/3.5366/3.5257	3.3255/3.3239/3.3167	3.2279/3.2256/3.2390
HC	3.4668/3.4783/3.4767	3.2699/3.2806/3.2809	3.4375/3.2510/3.2461	3.4759/3.4692/3.4709	3.2774/3.2694/3.2873	3.2227/3.2280/3.2215
mHC	3.4787/3.4769/3.4822	3.2554/3.2506/3.2676	3.1831/3.1997/3.1974	3.4715/3.4568/3.4705	3.2422/3.2352/3.2463	3.1746/3.1868/3.1834
mHC-lite	3.4803/3.4754/3.4805	3.2665/3.2568/3.2638	3.1997/3.2034/3.1921	3.4693/3.4528/3.4670	3.2456/3.2346/3.2475	3.1802/3.1837/3.1848
KromHC	3.4797/3.4804/3.4819	3.2607/3.2631/3.2624	3.1902/3.1999/3.1971	3.4702/3.4587/3.4699	3.2458/3.2360/3.2460	3.1836/3.1867/3.1850
sHC	3.4989/3.4810/3.4773	3.2428/3.2545/3.2563	3.2390/3.2079/3.2873	3.4687/3.4490/3.4717	3.2396/3.2434/3.2374	3.2434/3.2163/3.3350
osHC-C **(ours)**	3.4683/3.4750/3.4759	3.2528/3.2630/3.2650	3.1791/3.1899/3.1852	3.4588/3.4503/3.4645	3.2369/3.2319/3.2434	3.1671/3.1754/3.1786
osHC-R **(ours)**	3.4692/3.4746/3.4755	3.2504/3.2545/3.2603	3.1812/3.1958/3.1942	3.4652/3.4481/3.4653	3.2409/3.2276/3.2430	3.1756/3.1808/3.1822
osHC-C→R **(ours)**	3.4721/3.4735/3.4767	3.2589/3.2536/3.2562	3.1754/3.1958/3.1829	3.4602/3.4473/3.4602	3.2400/3.2324/3.2384	3.1687/3.1742/3.1719
osHC-R→C **(ours)**	3.4706/3.4736/3.4654	3.2466/3.2703/3.2617	3.1833/3.2005/3.1991	3.4628/3.4493/3.4597	3.2339/3.2318/3.2452	3.1719/3.1839/3.1814

C.4 Training dynamics on OpenWebText

Figures 3 and 4 show the Large FineWeb-Edu runs. Figures 7 and 8 repeat them on OpenWebText, on the same axes and under the same treatment.

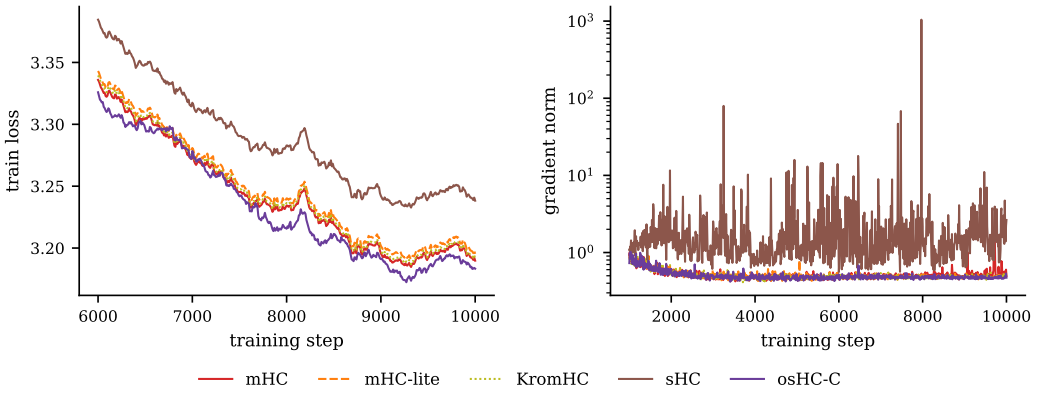


Figure 7: **One marginal against two, during training** (Large, OpenWebText). The OpenWebText counterpart of Figure 3. Training loss on the left, gradient norm on the right, each curve a mean over three seeds. The loss panel is smoothed by an exponential moving average with weight 0.03. The panels start at step 6000 and 1000 respectively, so that the large values of the opening updates do not compress the scale and hide the differences.

C.5 Block schedules and depth

The completed Large runs settle the question the block schedules left open: they do retain their advantage with depth. osHC-C→R has the best or joint-best mean in four of the six cells of Table 1,

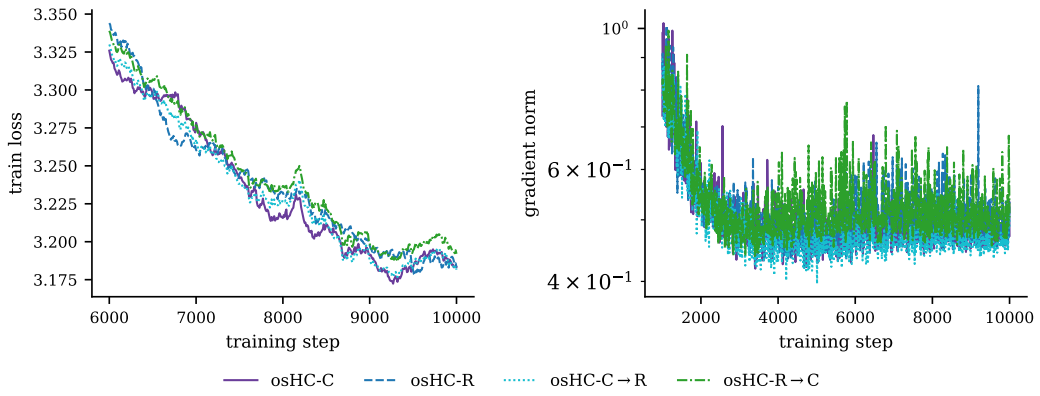
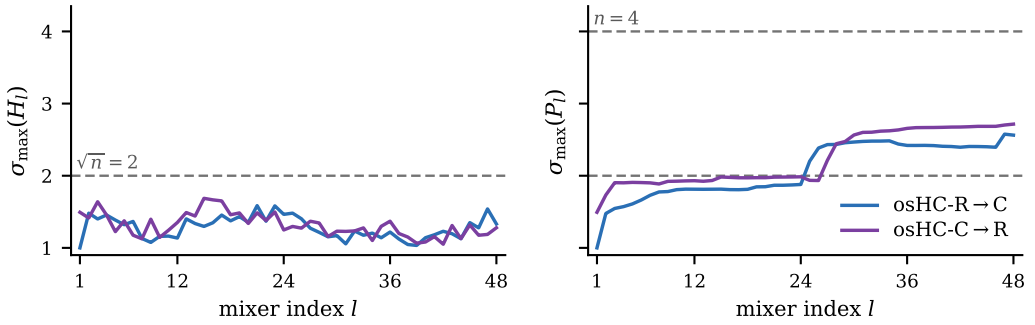


Figure 8: **The four osHC parameterisations against each other** (Large, OpenWebText). The OpenWebText counterpart of Figure 4. Built as in Figure 7, with both y-axes rescaled to the range the four variants occupy.



(a) Single-layer mixer (b) Composite mixer

Figure 9: **Spectral norm of the single-layer and composite mixers, for the switch schedules.** Built exactly as Figure 6. The ceiling on the composition is the two-block bound $n = 4$ of Theorem B.4, not the pure-schedule bound $\sqrt{n} = 2$.

including both Large cells, and it is the only variant that separates from all three Birkhoff constructions at $16/18$ or better. We do not read this as establishing that alternating the marginal is the right design. Theorem B.4 shows that a block schedule pays a ceiling set by the number of switches, and one switch is the cheapest such schedule; a sweep over switch counts is the experiment that would settle it, and we have not run it.

Figure 9 is the switch-schedule counterpart of Figure 6, and it shows that ceiling being paid rather than merely allowed. The single mixers stay under $\sqrt{n} = 2$ throughout (panel a), each one still belonging to one family. The composition is where the switch tells: each curve tracks the pure schedule of its opening family up to mixer index 24, then leaves $\sqrt{n} = 2$ and climbs to a new plateau (panel b). The largest value reached is 2.72, inside the two-block bound $n = 4$ but well above what either pure schedule attains, and the plateau after the switch is again flat in depth.

C.6 Out-of-distribution generalisation

Following the Paloma protocol (Magnusson et al., 2023), we evaluate the Large checkpoints from both pretraining corpora on C4, Dolma v1.5, Falcon RefinedWeb, RedPajama, and WikiText-103. We keep documents separate, prepend GPT-2 end-of-text at each document boundary, and score every token exactly once in non-overlapping prediction windows of at most 1024 tokens. The primary

metric is word-normalised perplexity,

$$\mathrm{PPL}_{\mathrm{word}} = \exp\left(-\frac{\sum_d \log p_\theta(x^{(d)})}{\sum_d \#\,\mathrm{words}(x^{(d)})}\right),$$

matching `lm-evaluation-harness` (Gao et al., 2023). Each entry aggregates the same three seeds as the in-domain table. We pin our implementation and data revision and use the checkpoints' architectural context limit of 1024.

Section 5.2 quotes the reduction $100(1 - \mathrm{PPL}_{\mathrm{method},d}/\mathrm{PPL}_{\mathrm{mHC},d})$ on each corpus d separately, computed from the corpus-level three-seed means printed in Table 2. Averaging the five weights the corpora equally and gives 3.60% for osHC-C and 3.92% for osHC-C$\rightarrow$R on FineWeb-Edu, and 3.67% for osHC-C and 4.36% for osHC-R on OpenWebText. We reduce each corpus to a ratio before averaging rather than averaging the perplexities themselves, as Paloma does across the domains within one source, because the five differ by more than an order of magnitude in absolute perplexity and RedPajama would otherwise decide the result almost alone. Aggregating the same five ratios geometrically moves each of these averages by at most 0.07 percentage points.

The two Large OOD grids together contain 60 checkpoints (two corpora, ten methods, three seeds) and 300 complete checkpoint–corpus evaluations.