ORTHOSTOCHASTIC RESIDUAL MIXING FOR MANIFOLD-CONSTRAINED HYPER-CONNECTIONS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Manifold-Constrained Hyper-Connections (mHC) stabilize deep network training by constraining residual mixing matrices to be doubly stochastic via Sinkhorn-Knopp projection. However, this iterative projection requires 10–20 iterations per forward pass, adding computational overhead. We investigate *orthostochastic matrices*—doubly stochastic matrices formed by entrywise squaring an orthogonal matrix—as a simpler alternative constructed via Newton-Schulz iteration. On language model pretraining with nanoGPT, orthostochastic mHC matches Sinkhorn-projected mHC at $n = 4$ residual streams (validation loss $\Delta = +0.003$, within 0.5σ), while a small gap emerges at $n = 8$ ($\Delta = +0.013$, between 0.5σ and 1.0σ), consistent with reduced expressiveness at larger n. Both methods provide approximately 39% gradient stabilization compared to unconstrained Hyper-Connections and converge to near-identity mixing matrices during training. Our results suggest that orthostochastic construction offers a viable alternative to Sinkhorn projection for small n, leveraging well-understood orthogonalization primitives.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Residual connections (He et al., 2015) are fundamental to training deep neural networks, enabling gradient flow through identity shortcuts that mitigate vanishing gradients. The Transformer architecture (Vaswani et al., 2017) relies heavily on residual connections, with each attention and feed-forward sublayer wrapped in a residual path. Hyper-Connections (Zhu et al., 2024) generalize this paradigm by maintaining n parallel residual streams with learnable mixing matrices that route information across streams, improving optimization and representation learning. However, unconstrained mixing matrices can cause numerical instability in deep networks. Manifold-Constrained Hyper-Connections (mHC) (Xie et al., 2026) address this by constraining the residual mixing matrix $\mathbf{H}^{\text{res}}$ to be doubly stochastic—a nonnegative matrix where both rows and columns sum to one—providing spectral norm bounds that prevent gradient explosion.

The standard approach to enforcing doubly stochastic constraints in mHC is Sinkhorn-Knopp projection, which alternately normalizes rows and columns until convergence. While effective, this requires 10–20 iterations per forward pass, adding computational overhead and introducing hyperparameters (iteration count, temperature) that affect both constraint fidelity and gradient flow. This motivates exploring simpler alternatives that can achieve similar stability benefits with reduced complexity. We investigate *orthostochastic matrices*—doubly stochastic matrices formed by entrywise squaring an orthogonal matrix—as a structured subset that can be constructed via Newton-Schulz iteration (Grishina et al., 2025; Amsel et al., 2025), a well-understood orthogonalization primitive used in modern optimizers. Our contributions are:

- We propose orthostochastic $\mathbf{H}^{\text{res}}$ construction for mHC, using Newton-Schulz iteration followed by entrywise squaring to produce doubly stochastic matrices with exact constraint satisfaction.

[1] https://gitlab.com/fars-a/orthostochastic-mhc

- We empirically evaluate orthostochastic mHC against Sinkhorn-projected mHC on language model pretraining. At $n = 4$ residual streams, orthostochastic matches Sinkhorn ($\Delta = +0.003$ validation loss, within 0.5σ); at $n = 8$, a small gap emerges ($\Delta = +0.013$, between 0.5σ and 1.0σ), consistent with reduced expressiveness at larger n.
- We find that both methods converge to near-identity $\mathbf{H}^{\text{res}}$ matrices during training, suggesting that mHC's benefits arise from stability guarantees rather than exploiting the full Birkhoff polytope interior.

2 Related Work

Residual Connections and Training Stability. Several works have studied how to improve residual connection stability beyond the standard formulation (He et al., 2015). ReZero (Bachlechner et al., 2020) initializes residual branches with zero scaling for faster convergence, Fixup (Zhang et al., 2019) proposes initialization schemes that enable training without normalization, and De & Smith (2020) show that batch normalization biases deep residual networks toward shallow paths by emphasizing identity mappings.

Hyper-Connections and Manifold Constraints. Hyper-Connections (Zhu et al., 2024) generalize residual connections by maintaining multiple parallel streams with learnable mixing matrices. Manifold-Constrained Hyper-Connections (mHC) (Xie et al., 2026) constrain the residual mixing matrix to be doubly stochastic via Sinkhorn-Knopp projection for improved stability. Follow-up work has explored alternative parameterizations: mHC-lite (Yang & Gao, 2026) uses exact permutation-mixture representations for $n = 4$, while KromHC (Zhou et al., 2026) employs Kronecker-product structure for exact doubly stochastic matrices with fewer parameters.

Doubly Stochastic and Orthogonal Matrices. Doubly stochastic matrices form the Birkhoff polytope, with permutation matrices as vertices (Linderman et al., 2017). The Sinkhorn-Knopp algorithm (Mena et al., 2018) projects arbitrary matrices onto this polytope through alternating row and column normalization, enabling differentiable relaxations of permutation learning (Grover et al., 2019). For orthogonal matrix computation, Newton-Schulz iteration provides a gradient-friendly alternative to QR decomposition (Grishina et al., 2025), with recent work analyzing optimal polynomial variants (Amsel et al., 2025). This iteration has found application in the Muon optimizer (Liu et al., 2025) for orthogonalizing gradient updates. Our work connects these two lines by using orthogonal matrices to construct orthostochastic matrices, which form a subset of doubly stochastic matrices with exact constraint satisfaction.

3 Method

3.1 Background: Manifold-Constrained Hyper-Connections

Hyper-Connections (Zhu et al., 2024) extend standard residual connections by maintaining n parallel residual streams and learning per-layer mixing matrices that route information across these streams. For a layer ℓ with input hyper-hidden matrix $\mathbf{H}^{\ell-1} \in \mathbb{R}^{n \times d}$, the output is computed as:

$$\mathbf{H}^{\ell} = \mathbf{H}_{\ell}^{\text{post}} \odot \mathcal{T}(\mathbf{H}_{\ell}^{\text{pre}} \cdot \mathbf{H}^{\ell-1}) + \mathbf{H}_{\ell}^{\text{res}} \cdot \mathbf{H}^{\ell-1}, \tag{1}$$

where $\mathcal{T}$ denotes the layer transformation (e.g., attention or feed-forward), $\mathbf{H}_{\ell}^{\text{pre}}, \mathbf{H}_{\ell}^{\text{post}} \in \mathbb{R}^{n}$ are input/output mixing vectors, and $\mathbf{H}_{\ell}^{\text{res}} \in \mathbb{R}^{n \times n}$ is the residual mixing matrix.

While Hyper-Connections improve optimization and representation learning, unconstrained $\mathbf{H}_{\ell}^{\text{res}}$ matrices can cause numerical instability in deep networks, as the depth-wise product of mixing matrices may amplify activations and gradients. Manifold-Constrained Hyper-Connections (mHC) (Xie et al., 2026) address this by constraining $\mathbf{H}_{\ell}^{\text{res}}$ to be *doubly stochastic*—a nonnegative matrix where both rows and columns sum to one. Formally, $\mathbf{H}_{\ell}^{\text{res}}$ is constrained to the Birkhoff polytope:

$$\mathcal{M}^{\text{res}} = \left\{ \mathbf{M} \in \mathbb{R}^{n \times n} \mid \mathbf{M}\mathbf{1}_n = \mathbf{1}_n, \mathbf{1}_n^{\top}\mathbf{M} = \mathbf{1}_n^{\top}, \mathbf{M} \geq 0 \right\}, \tag{2}$$

where $\mathbf{1}_n$ is the n-dimensional all-ones vector.

The doubly stochastic constraint provides several theoretical guarantees beneficial for training stability. First, the spectral norm of any doubly stochastic matrix is bounded by one ($\|\mathbf{H}_\ell^{\text{res}}\|_2 \leq 1$), ensuring non-expansive residual mappings that mitigate gradient explosion. Second, the set of doubly stochastic matrices is closed under multiplication, so the composite residual mapping $\prod_{i=1}^{L-\ell} \mathbf{H}_{L-i}^{\text{res}}$ remains doubly stochastic throughout the network depth. Third, the Birkhoff polytope is the convex hull of permutation matrices, providing a geometric interpretation where residual mixing acts as a convex combination of permutations.

In mHC, the doubly stochastic constraint is enforced via Sinkhorn-Knopp projection (Mena et al., 2018). Given unconstrained logits $\tilde{\mathbf{H}}_\ell^{\text{res}}$, the projection first exponentiates to ensure positivity, then alternately normalizes rows and columns:

$$\mathbf{M}^{(0)} = \exp(\tilde{\mathbf{H}}_\ell^{\text{res}}), \quad \mathbf{M}^{(t)} = T_r(T_c(\mathbf{M}^{(t-1)})), \tag{3}$$

where T_r and T_c denote row and column normalization, respectively. After $t_{\max}$ iterations (typically 10–20), the result converges to a doubly stochastic matrix $\mathbf{H}_\ell^{\text{res}} = \mathbf{M}^{(t_{\max})}$.

3.2 Orthostochastic Matrices

While Sinkhorn projection can represent any doubly stochastic matrix, we investigate whether a simpler structured subset suffices for effective residual mixing. Specifically, we consider *orthostochastic matrices*—doubly stochastic matrices that arise as the entrywise square of an orthogonal matrix (Linderman et al., 2017). Given an orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ (i.e., $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$), the orthostochastic matrix is defined as:

$$\mathbf{O} = \mathbf{Q} \odot \mathbf{Q}, \tag{4}$$

where $\odot$ denotes the Hadamard (entrywise) product. Since orthogonal matrices have columns with unit ℓ_2-norm and the entrywise square preserves nonnegativity, the resulting matrix $\mathbf{O}$ satisfies both row and column sum constraints, making it doubly stochastic.

Orthostochastic matrices form a strict subset of the Birkhoff polytope with reduced degrees of freedom. A general $n \times n$ doubly stochastic matrix has $(n-1)^2$ free parameters (due to row and column sum constraints), while an orthostochastic matrix is parameterized by an orthogonal matrix with $n(n-1)/2$ degrees of freedom. The ratio of degrees of freedom is:

$$\frac{n(n-1)/2}{(n-1)^2} = \frac{n}{2(n-1)}, \tag{5}$$

which equals 67% for $n = 4$ and decreases to 57% for $n = 8$. Despite this reduced expressiveness, orthostochastic matrices retain important properties: they include all permutation matrices (the vertices of the Birkhoff polytope) and can represent smooth interpolations between permutations.

The key question motivating our work is whether mHC's benefits arise from accessing the full interior of the Birkhoff polytope, or whether the structured subset of orthostochastic matrices—which still enables permutation-like routing and gradual mixing—is sufficient for stable and effective residual mixing.

3.3 Newton-Schulz Orthostochastic Construction

To construct orthostochastic $\mathbf{H}_\ell^{\text{res}}$ matrices, we employ Newton-Schulz iteration (Liu et al., 2025; Amsel et al., 2025) to orthogonalize unconstrained logits, followed by entrywise squaring. Given logits $\mathbf{L} \in \mathbb{R}^{n\times n}$, the Newton-Schulz iteration computes:

$$\mathbf{X}^{(0)} = \mathbf{L}, \quad \mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} \left(a\mathbf{I} + b\mathbf{X}^{(k)\top}\mathbf{X}^{(k)} + c(\mathbf{X}^{(k)\top}\mathbf{X}^{(k)})^2 \right), \tag{6}$$

where $(a, b, c) = (3.0, -3.2, 1.2)$ are coefficients that accelerate convergence (Grishina et al., 2025). After K iterations (typically 15–20), $\mathbf{X}^{(K)}$ converges to an approximately orthogonal matrix $\mathbf{Q}$. The orthostochastic $\mathbf{H}_\ell^{\text{res}}$ is then obtained by entrywise squaring:

$$\mathbf{H}_\ell^{\text{res}} = \mathbf{Q} \odot \mathbf{Q}. \tag{7}$$

Figure 1 illustrates the comparison between Sinkhorn projection and our orthostochastic construction. While Sinkhorn iteratively normalizes rows and columns to produce doubly stochastic matrices

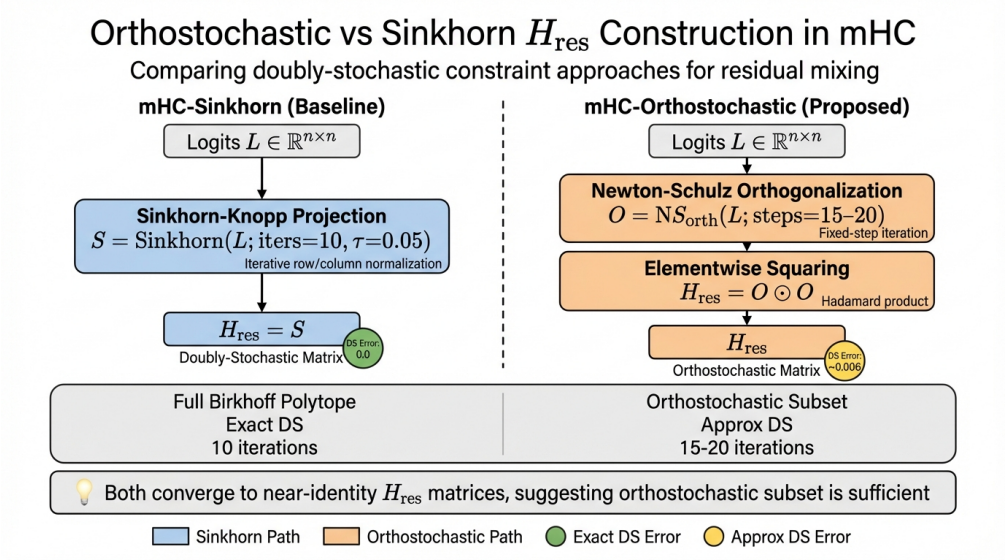

Figure 1: Comparison of $\mathbf{H}^{\text{res}}$ construction methods in manifold-constrained Hyper-Connections. Left: Sinkhorn projection iteratively normalizes rows and columns to produce doubly stochastic matrices. Right: Orthostochastic construction uses Newton-Schulz iteration to produce orthogonal matrices, then squares them entrywise to obtain doubly stochastic matrices. Both methods constrain $\mathbf{H}^{\text{res}}$ to the Birkhoff polytope, but orthostochastic matrices form a strict subset with reduced degrees of freedom.

directly, the orthostochastic approach first orthogonalizes via Newton-Schulz iteration, then squares entrywise to obtain doubly stochastic matrices. Both methods constrain $\mathbf{H}_{\ell}^{\text{res}}$ to the Birkhoff polytope, but orthostochastic matrices form a strict subset with reduced degrees of freedom.

The Newton-Schulz iteration is differentiable and can be efficiently implemented on GPUs, making it suitable for end-to-end training. Unlike Sinkhorn projection, which requires careful temperature tuning and may produce numerically unstable gradients at low temperatures, Newton-Schulz orthogonalization has well-understood convergence properties and has been successfully deployed in modern optimizers such as Muon (Liu et al., 2025). The orthostochastic construction thus provides a principled alternative that leverages existing orthogonalization primitives while maintaining the stability guarantees of doubly stochastic constraints.

4 Experiments

4.1 Experimental Setup

We evaluate orthostochastic $\mathbf{H}^{\text{res}}$ construction against Sinkhorn-projected mHC using the tokenbender nanoGPT implementation of Hyper-Connections. All experiments use the FineWeb dataset (Penedo et al., 2024) for language model pretraining with next-token prediction.

We consider two evaluation settings to test generalization across different configurations. **Setting A** (deep, standard) uses a 48-layer nanoGPT model with embedding dimension 150, 6 attention heads, and $n = 4$ residual streams—the most common configuration in prior HC/mHC work. **Setting B** (wider streams) uses a 6-layer model with embedding dimension 288, 6 attention heads, and $n = 8$ residual streams to test whether conclusions hold when scaling beyond $n = 4$, where permutation-mixture parameterizations become factorial.

For both settings, we train with AdamW (Loshchilov & Hutter, 2017) using learning rate 6×10^{-4} with cosine decay to 6×10^{-5}, 200 warmup iterations, and bfloat16 precision. Setting A uses batch size 8 with 4 gradient accumulation steps; Setting B uses batch size 32 with 4 gradient accumulation steps. All runs train for 5000 iterations on 4 GPUs with distributed data parallel.

Table 1: Main experimental results comparing $\mathbf{H}^{\text{res}}$ construction methods. Best results per column in **bold**. Δ indicates difference from Sinkhorn baseline.

Method	Setting A (48L, n=4)		Setting B (6L, n=8)	
	Val Loss	$r_{\max}$	Val Loss	$r_{\max}$
HC Unconstrained	4.7104	3.14	—	—
mHC-Sinkhorn	**4.7615**±0.009	1.91±0.24	**4.2495**±0.013	1.95±0.21
mHC-Orthostochastic	4.7642±0.013 (Δ+0.003)	**1.87**±0.13	4.2626±0.005 (Δ+0.013)	—

We compare three methods: (1) **mHC-Sinkhorn**, the standard mHC with Sinkhorn-Knopp projection using 10 iterations and temperature $\tau = 0.05$; (2) **mHC-Orthostochastic**, our proposed method using Newton-Schulz orthogonalization with 15 steps and coefficients $(a, b, c) = (3.0, -3.2, 1.2)$, followed by entrywise squaring; and (3) **HC Unconstrained**, standard Hyper-Connections without doubly stochastic constraints (Setting A only, for calibration).

We evaluate using three metrics. **Validation loss** measures model quality on held-out data. **Gradient spike ratio** $r_{\max} = \max_t r_t$ where $r_t = g_t / \text{median}(g_{t-100:t-1})$ for $t > 200$ quantifies training stability, with lower values indicating fewer gradient spikes. **DS error** measures the maximum deviation of row/column sums from 1, quantifying constraint fidelity. Setting A runs 5 seeds; Setting B runs 3 seeds.

4.2 Main Results

Table 1 presents the main experimental results comparing mHC-Sinkhorn, mHC-Orthostochastic, and HC Unconstrained across both settings.

In Setting A ($n = 4$), mHC-Orthostochastic achieves validation loss 4.7642±0.013, compared to 4.7615±0.009 for mHC-Sinkhorn. The difference ($\Delta = +0.003$) is within 0.5σ of the Sinkhorn baseline standard deviation, indicating statistical equivalence. This demonstrates that the orthostochastic subset of doubly stochastic matrices is sufficient for effective residual mixing at $n = 4$, despite having only 67% of the degrees of freedom of the full Birkhoff polytope.

In Setting B ($n = 8$), mHC-Orthostochastic achieves validation loss 4.2626±0.005, compared to 4.2495±0.013 for mHC-Sinkhorn. The difference ($\Delta = +0.013$) falls between 0.5σ and 1.0σ thresholds, representing a small but measurable gap. This small but measurable gap is consistent with the reduced expressiveness of orthostochastic matrices at larger n (57% DoF retained at $n = 8$ vs 67% at $n = 4$), suggesting that the orthostochastic subset may become more restrictive as the number of residual streams increases.

Comparing with unconstrained HC, both mHC variants show substantially improved gradient stability. The unconstrained baseline exhibits $r_{\max} = 3.14$, while mHC-Sinkhorn achieves $r_{\max} = 1.91$ and mHC-Orthostochastic achieves $r_{\max} = 1.87$—a reduction of approximately 39–40%. This confirms that the doubly stochastic constraint itself, rather than the specific projection method, provides the key benefit of gradient stabilization.

4.3 Constraint Fidelity Analysis

Table 2 presents the constraint fidelity analysis for both methods. We measure doubly stochastic (DS) error as the maximum deviation of row/column sums from 1, and orthogonality residual as $\|\mathbf{Q}^\top \mathbf{Q} - \mathbf{I}\|_F$ for the Newton-Schulz-produced orthogonal matrix before squaring.

Sinkhorn-Knopp projection achieves exact doubly stochastic matrices (DS error ≈ 0 in bfloat16 precision), as expected from its iterative row/column normalization. The orthostochastic construction produces approximately doubly stochastic matrices with DS error around 0.006, well below the 10^{-2} threshold required for training stability. The orthogonality residual of approximately 0.008 indicates that 15–20 Newton-Schulz iterations produce near-orthogonal matrices, with the small deviation from exact orthogonality propagating to the DS error after entrywise squaring. Importantly, constraint fidelity remains consistent across both settings ($n = 4$ and $n = 8$), demonstrating that the Newton-Schulz orthogonalization scales reliably to larger matrix sizes.

Table 2: Constraint fidelity analysis. DS Error measures deviation from doubly stochastic property. Orth Residual measures $\|\mathbf{Q}^\top\mathbf{Q} - \mathbf{I}\|_F$. All orthostochastic values well below 10^{-2} threshold.

Setting / Method	DS Error (final)	DS Error (max)	Orth Residual (final)	Orth Residual (max)
A: Sinkhorn	**0.0**	**0.0**	—	—
A: Orthostochastic	0.0061±0.0007	0.0085	0.0078±0.0005	0.0106
B: Sinkhorn	∼**0.0**	∼**0.0**	—	—
B: Orthostochastic	0.0053±0.0008	0.0075	0.0078±0.0008	0.0112

4.4 Training Dynamics

Analysis of training dynamics reveals that both methods exhibit similar gradient stability patterns. In Setting A, the median gradient norm during the last 1000 iterations is 0.876±0.007 for Sinkhorn and 0.864±0.011 for Orthostochastic, while unconstrained HC shows 30% higher gradient norms (1.15). This confirms that the doubly stochastic constraint provides meaningful gradient stabilization regardless of the specific construction method.

A notable finding is that both Sinkhorn and Orthostochastic mHC converge to near-identity $\mathbf{H}^{\text{res}}$ matrices during training. For Sinkhorn, all layers converge to exact identity matrices ($\|\mathbf{H}^{\text{res}} - \mathbf{I}\| = 0$). For Orthostochastic, the deviation from identity is small ($\|\mathbf{H}^{\text{res}} - \mathbf{I}\| \in [0.001, 0.166]$), with the largest deviation occurring at the final layer. This convergence to near-identity matrices suggests that neither method exploits dense interior points of the Birkhoff polytope during training, which explains why the orthostochastic subset's reduced expressiveness does not significantly impact performance. The practical implication is that mHC's benefits may arise primarily from the stability guarantees of the doubly stochastic constraint rather than from the ability to represent arbitrary convex combinations of permutations.

5 Conclusion

We investigated orthostochastic matrices as an alternative to Sinkhorn projection for constructing doubly stochastic $\mathbf{H}^{\text{res}}$ matrices in manifold-constrained Hyper-Connections. Our experiments demonstrate that orthostochastic construction matches Sinkhorn-projected mHC at $n = 4$ residual streams, with validation loss within 0.5σ of the baseline. At $n = 8$, a small gap emerges consistent with the reduced expressiveness of orthostochastic matrices at larger sizes. Both methods provide approximately 39% gradient stabilization compared to unconstrained Hyper-Connections and converge to near-identity $\mathbf{H}^{\text{res}}$ matrices during training. These findings suggest that for small n (e.g., $n = 4$), orthostochastic construction offers a viable alternative to Sinkhorn projection, leveraging well-understood Newton-Schulz orthogonalization primitives. Future work could explore hybrid approaches that combine orthostochastic structure with additional degrees of freedom for larger n.

References

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm, 2025. URL `https://arxiv.org/abs/2505.16932`.

Thomas C. Bachlechner, Bodhisattwa Prasad Majumder, H. H. Mao, G. Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. pp. 1352–1361, 2020.

Soham De and Samuel L. Smith. Batch normalization biases deep residual networks towards shallow paths. *ArXiv*, abs/2002.10444, 2020.

Ekaterina Grishina, Matvey Smirnov, and Maxim Rakhuba. Accelerating newton-schulz iteration for orthogonalization via chebyshev-type polynomials. *ArXiv*, abs/2506.10935, 2025.

Aditya Grover, Eric Wang, Aaron Zweig, and Stefano Ermon. Stochastic optimization of sorting networks via continuous relaxations. *ArXiv*, abs/1903.08850, 2019.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

Scott W. Linderman, Gonzalo E. Mena, H. Cooper, L. Paninski, and J. Cunningham. Reparameterizing the birkhoff polytope for variational permutation inference. pp. 1618–1627, 2017.

Jingyuan Liu, Jianling Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Meng Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training. *ArXiv*, abs/2502.16982, 2025.

I. Loshchilov and F. Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017.

Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. *ArXiv*, abs/1802.08665, 2018.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, L. V. Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. *ArXiv*, abs/2406.17557, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and I. Polosukhin. Attention is all you need. pp. 5998–6008, 2017.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, 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, 2026. URL https://arxiv.org/abs/2512.24880.

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

Hongyi Zhang, Yann Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. *ArXiv*, abs/1901.09321, 2019.

Wuyang Zhou, Yuxuan Gu, Giorgos Iacovides, and Danilo P. Mandic. Kromhc: Manifold-constrained hyper-connections with kronecker-product residual matrices. 2026.

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