Rook: Global Permutation Geometry for Neural Routing

author names withheld

Under Review for OPT 2026

Abstract

Modern optimizer stacks match geometry to parameter class, giving dense hidden matrices a spectral duality map and vectors coordinatewise adaptivity, yet doubly stochastic routing blocks, whose natural atoms are permutations, are still trained by applying AdamW to an unconstrained Sinkhorn score. We construct the missing member of the stack from those atoms. The central symmetrization of the Birkhoff polytope is the unit ball of a *matching norm* whose exact smooth duality map is a Gibbs expectation over permutations with a matrix permanent as its potential, and an exact identity splits that potential into Sinkhorn free energy plus permanent structure of the scaled matrix. Rook replaces the intractable part with a fractional-Bethe map whose coefficient $\gamma_n = -(n-1)/n$ is uniquely fixed by exact permutation covariance, and which stays Lipschitz in the score where the hard assignment oracle jumps across ties. The construction pays off where routing has to adapt. After the routed streams are rewired from a shared checkpoint, Rook lowers post-shift cross-entropy from 0.415 to 0.156 against Muon+AdamW and recovers $2.96\times$ sooner, on all ten paired seeds and at 7.1% lower step time. Under batch-32 gradient noise it reduces direction chatter by 62.4% relative to that oracle and halves final test loss. On stationary training AdamW on the scores stays ahead.

1. Introduction

The premise behind geometry-aware optimization is that storage shape is not parameter semantics. Muon treats dense hidden matrices through a spectral duality map rather than coordinatewise adaptivity [3, 12], and norm-constrained linear minimization oracle (LMO) methods make the same principle explicit at the level of unit balls [18]. That premise exposes a gap, since some neural routers and residual-stream mixers are constrained to the Birkhoff polytope

$$\mathcal{B}_n = \{B \in \mathbb{R}^{n\times n} : B \geq 0,\ B\mathbf{1} = \mathbf{1},\ B^\top \mathbf{1} = \mathbf{1}\} = \mathrm{conv}(\mathcal{P}_n), \tag{1}$$

where $\mathcal{P}_n$ is the set of permutation matrices [4]. Such blocks appear in manifold-constrained hyper-connections and related architectures [22]. Their atoms are assignments, not coordinates and not orthogonal matrices, yet the common implementation forms $B = \mathrm{Sinkhorn}(S/T)$ [8, 20] and applies AdamW to the score S. This is geometrically indirect twice over. Row–column gauge components of S do not change B at all, and the Sinkhorn Jacobian contracts the useful primal signal as routes sharpen.

Our contributions are a *matching norm* for Birkhoff blocks with its dual norm, hard LMO, and stationarity guarantee, together with an exact Sinkhorn–permanent decomposition identifying what Sinkhorn geometry omits (Sections 2 and 3); a uniquely curvature-matched coefficient $\gamma_n = -(n-1)/n$, derived from exact permutation covariance and, independently, from the local Dittert

metric (Section 4); accuracy theory for the deployed map, quadratic for raw marginals and cubic for the odd signed map, with a Lipschitz bound the hard LMO cannot satisfy across ties (Section 5); and a homotopy–Newton solver over gauge-fixed $(2n-1)$ systems with a primal-gradient router interface, evaluated from fidelity against the exact map up to a 5M-parameter transformer (Sections 6 and 7). The scope is deliberately narrow. A block should receive Rook only when its forward value is doubly stochastic by construction, permutations are meaningful task atoms, and the primal gradient is available.

2. A matching norm for routing blocks

The affine tangent space of $\mathcal{B}_n$ is $\mathcal{T}_n = \{D : D\mathbf{1} = 0,\ D^\top \mathbf{1} = 0\}$, with orthogonal projection $\Pi_{\mathcal{T}_n} G = HGH$, $H = I - \mathbf{1}\mathbf{1}^\top/n$. A row–column gauge $a\mathbf{1}^\top + \mathbf{1}b^\top$ scores every permutation equally, so assignments, Gibbs marginals, and Sinkhorn matrices depend only on $\Pi_{\mathcal{T}_n} G$.

$\mathcal{B}_n$ is not centrally symmetric, so it cannot serve as a signed unit ball. Its central symmetrization $\mathcal{K}_n = \frac{1}{2}(\mathcal{B}_n - \mathcal{B}_n) \subset \mathcal{T}_n$ contains a neighborhood of zero in $\mathcal{T}_n$ and is therefore the unit ball of a norm on the tangent space, the *matching norm*. Its dual norm is the assignment spread

$$\|G\|_{\mathrm{match},*} = \tfrac{1}{2}\Big[\max_{P\in\mathcal{P}_n} \langle G, P\rangle - \min_{P\in\mathcal{P}_n} \langle G, P\rangle\Big], \tag{2}$$

computable by two assignment solves [14], and its linear oracle is the signed difference of extreme assignments, which we call the *hard LMO*.

For compact convex C and $K = (C - C)/2$ the support function obeys $h_K(G) = \frac{1}{2}(h_C(G) + h_C(-G))$, and $(c_+ - c_-)/2$ maximizes $\langle G, \cdot\rangle$ over K whenever $c_\pm$ extremize $\langle G, \cdot\rangle$ over C (Theorem 9). At $C = \mathcal{B}_n$ this gives the zero-temperature direction $A_0(G) = \frac{1}{2}(P_+(G) - P_-(G))$, which is tangent, gauge- and scale-invariant, aligned ($\langle G, A_0(G)\rangle = h_{\mathcal{K}_n}(G) \geq 0$), and bounded ($\|A_0(G)\|_F^2 \leq n/2$). For f bounded below and L-smooth in the matching norm, steepest descent along it with step $\eta_t = \|\nabla f(x_t)\|_{\mathrm{match},*}/L$ gives $\min_{t<T} \|\nabla f(x_t)\|_{\mathrm{match},*} \leq \sqrt{2L(f(x_0) - f_\star)/T}$ (Proposition 10).

This is the polytope analogue of choosing updates from the norm ball induced by a parameter class's own atoms. Orthogonal matrices give the spectral ball and the polar factor (Muon), and permutations give $\mathcal{B}_n$ and a difference of assignments [5, 19].

3. Exact geometry and the Sinkhorn–permanent decomposition

For $\tau > 0$ the smooth matching potential and its marginal are

$$\Psi_\tau(G) = \tau \log \sum_{P\in\mathcal{P}_n} e^{\langle G,P\rangle/\tau} = \tau \log \mathrm{per}(e^{G/\tau}), \qquad M_\tau(G) = \nabla\Psi_\tau(G) = \mathbb{E}_{q_G}[P] \in \mathcal{B}_n, \tag{3}$$

with Hessian $\frac{1}{\tau}\mathrm{Var}_{q_G}(\langle\cdot, P\rangle)$: the geometry of the block is the covariance of a Gibbs law over assignments. Exact evaluation is exponential in general.

Proposition 1 (Exact high-temperature metric) *For every tangent G, $M_\tau(G) = U + \frac{1}{\tau(n-1)}G + O(\tau^{-2})$ with $U = \mathbf{1}\mathbf{1}^\top/n$; the covariance operator of the uniform assignment law acts on $\mathcal{T}_n$ as $Z \mapsto Z/(n-1)$.*

The entropic relaxation F_τ with entrywise entropy has optimizer $B_\tau(G) = \mathrm{Sinkhorn}(G/\tau)$ and tangent response $G/(n\tau)$, so Sinkhorn under-responds by $(n-1)/n$; an independent-entry entropy is not the entropy of a distribution over perfect matchings. The omitted term is exact:

Theorem 2 (Sinkhorn–permanent decomposition) *For every G and $\tau > 0$,*

$$\boxed{\Psi_\tau(G) = F_\tau(G) + \tau \log \mathrm{per}\big(B_\tau(G)\big), \qquad F_\tau(G) = \tau \log \mathrm{cap}\big(e^{G/\tau}\big).} \tag{4}$$

Consequently $\tau \log \nu_n \le \Psi_\tau(G) - F_\tau(G) \le 0$ with $\nu_n = n!/n^n$, by the Van der Waerden bound [9, 10]. On $\mathcal{B}_n$, Dittert's functional [6] satisfies $\Phi_D(B) = 2 - \mathrm{per}(B)$, so the omitted free energy is exactly Dittert structure of the scaled matrix; and since a doubly stochastic matrix is its own scaling, $\mathrm{cap}(B) = 1$ for every $B \in \mathcal{B}_n$, so capacity captures the Sinkhorn part of matching geometry and nothing else.

Differentiating Equation (4) gives the exact marginal correction, and each of its entries depends on a global $(n-1) \times (n-1)$ subpermanent (Section B), which is why coordinatewise corrections are structurally wrong.

4. One curvature-matched global map

For $\gamma \in [-1, 1]$ define the fractional matching entropy [7]

$$H_\gamma(B) = -\sum_{ij} B_{ij} \log B_{ij} - \gamma \sum_{ij} (1 - B_{ij}) \log(1 - B_{ij}), \tag{5}$$

and the variational map $R_{\tau,\gamma}(G) = \arg\max_{B \in \mathcal{B}_n} \{\langle G, B\rangle + \tau H_\gamma(B)\}$. The endpoints are familiar. Setting $\gamma = 0$ recovers Sinkhorn and $\gamma = -1$ the classical Bethe approximation, a certified lower bound on the permanent, while $\gamma = -1/2$ upper-bounds it, and the fractional value is monotone in γ [1, 7, 21].

Theorem 3 (Unique curvature match) *For tangent Z, $DR_{\tau,\gamma}(0)[Z] = \frac{n-1}{\tau n(n-1+\gamma)} Z$. The unique coefficient matching the exact response $DM_\tau(0)[Z] = Z/[\tau(n-1)]$ is*

$$\boxed{\gamma_n = -\frac{n-1}{n}.} \tag{6}$$

The same coefficient is implied by Theorem 12 (appendix), where adding the missing quadratic curvature of $\log \mathrm{per}$ to the Sinkhorn entropy changes the inverse tangent metric from n to $n-1$, achieved globally at γ_n. Raw Bethe over-responds by $(n-1)^2/[n(n-2)]$; Rook removes that finite-width error, and since $\gamma_n \in [-1, -\frac{1}{2}]$ its potential stays inside the certified bracket containing the true permanent. From here on $R_\tau = R_{\tau,\gamma_n}$.

5. The signed map and its guarantees

For nonzero tangent momentum, Rook uses the odd, RMS-normalized direction

$$\bar{G} = \frac{\Pi_{\mathcal{T}_n} G}{\mathrm{RMS}(\Pi_{\mathcal{T}_n} G)}, \qquad A_\rho(G) = \tfrac{1}{2}\{R_\rho(\bar{G}) - R_\rho(-\bar{G})\}, \qquad d_\rho(G) = \mathrm{RMSNorm}\big(A_\rho(G)\big). \tag{7}$$

The map A_ρ is a subgradient of an even convex potential, hence odd, tangent, monotone, aligned, and bounded by $\sqrt{n}$, and temperature interpolates between tangent-gradient geometry ($\rho \to \infty$) and the assignment oracle ($\rho \downarrow 0$) inside one optimizer (Proposition 11, appendix).

Theorem 4 (Local accuracy, three orders) *Let $S = \Pi_{\mathcal{T}_n} G/\tau$. There are constants μ_n, L_n, C_n depending only on n, given in Section E, such that if $\|S\|_F \le \mu_n/(2n)$,*

$$\textit{(raw marginal)} \quad \|M(S) - R(S)\|_F \;\le\; \left(4n^{3/2} + \tfrac{L_n}{2\mu_n^3}\right)\|S\|_F^2; \tag{8}$$

$$\textit{(odd signed map)} \quad \left\|\tfrac{1}{2}\{M(S) - M(-S)\} - A_R(S)\right\|_F \;\le\; C_n\,\|S\|_F^3, \tag{9}$$

After RMS normalization the direction error is again $O(\|S\|_F^2)$.

Theorem 4 is a local statement. Its sufficient radius is $\mu_n/(2n) \approx 0.183$ at $n = 4$, while the deployed optimizer applies the map to an RMS-normalized momentum at relative temperature 1, outside that radius. The theorem certifies the *order* of agreement near the uniform matrix, and the fidelity study of Section 7 establishes accuracy in the regime training visits. Odd symmetrization cancels the quadratic terms of both expansions, and normalization restores a second-order term through the $O(\|S\|_F^2)$ relative perturbation of $\mathrm{RMS}(A)$. Measured over widths 4–6, the log–log error slopes are 2.97 for the signed map and 2.00 for the normalized direction, matching both predicted exponents.

Proposition 5 (Lipschitz response versus tie discontinuity) *Where the restricted fractional Hessian has curvature at least $\mu > 0$, the Rook marginal is $1/\mu$-Lipschitz in the dimensionless score, $\|R(S_1) - R(S_2)\|_F \le \mu^{-1}\|S_1 - S_2\|_F$, and the odd signed map inherits the bound. The hard LMO is piecewise constant and admits no finite local Lipschitz constant across a hyperplane where two assignments tie.*

This contrast is the mechanism behind the gradient-noise experiment of Section 7.

6. Solver and primal-gradient execution

The map is solved in primal Birkhoff coordinates. Eliminating the equality-constrained Newton step and fixing one gauge leaves a dense symmetric system of order $2n - 1$ (7×7 at $n = 4$), globalized by a cosine score homotopy at fixed budget $O(KIn^3)$. Entries of the coordinate Hessian can be negative near a vertex even though the constrained free energy is convex, so the solver pivots, checks ascent, and falls back to projected gradient.

The router interface contributes as much as the solver. `BirkhoffRouter` computes the same log-domain Sinkhorn forward as a conventional implementation but detaches it and re-exposes it as a leaf, so `backward()` deposits $\nabla_B \ell$ without constructing the Jacobian graph of the normalization, which no optimizer reading the primal gradient consumes. Removing it roughly halves the isolated router forward+backward cost and makes the full Rook stack 7.1% faster per step than Muon+AdamW despite the added solves. Batched GPU execution of the solves is outside this paper's scope, and no optimizer-level claim depends on it.

Table 1: Held-out rewiring from a shared checkpoint, mean ± s.e. over 10 seeds; CE averages eight adaptation epochs. Each arm isolates one explanation.

Stack	Post-shift CE ↓	Recovery (ep.) ↓	Accuracy (%) ↑	ms/step
AdamW	$.406 \pm .009$	$6.9 \pm .5$	$94.44 \pm .36$	2.882
Muon+AdamW	$.415 \pm .015$	$7.1 \pm .5$	$95.91 \pm .20$	3.915
+ score cap	$.270 \pm .011$	$5.3 \pm .3$	$96.42 \pm .19$	4.056
+ hard LMO	$.174 \pm .008$	$2.6 \pm .2$	$97.42 \pm .13$	3.640
+ **Rook**	$\mathbf{.156 \pm .006}$	$\mathbf{2.4 \pm .3}$	$97.16 \pm .24$	**3.639**
+ exact Gibbs	$.156 \pm .006$	$2.2 \pm .2$	$97.09 \pm .23$	3.941

7. Experiments

Hyperparameters are selected on development seeds and frozen before held-out evaluation; exact permanents and Gibbs marginals appear only in validation oracles, never in training. Every comparison is paired. Arms are cloned from one checkpoint and receive identical minibatches, so only the optimizer assignment varies.

Fidelity to the exact map For $n \in \{4, 6, 8\}$, five temperatures spanning the deployed range, and 50 held-out tangent matrices per cell, we compare signed directions against exact permutation enumeration (Table 3, appendix). Rook wins 748/750 paired cases, cutting Sinkhorn's mean relative error from 0.2169 to 0.0395 and raw Bethe's by 72.8%, which covers the regime Theorem 4 does not.

Adaptation after a rewiring Four ordered image streams feed a learned 4×4 Birkhoff router, a residual MLP, and a classifier; one Muon+AdamW checkpoint per seed is cloned into every arm and the stream order is rewired (Table 1). Rook lowers mean post-shift CE by 62.5% and recovers $2.96\times$ sooner, winning loss and recovery on 10/10 paired seeds at 7.1% lower step time, and matches the exact Gibbs reference to four decimals. The score-cap control rules out clipping. The hard LMO, identical but for the duality map, reaches the highest terminal accuracy in this batch-128 run while Rook holds lower trajectory CE in 9/10 seeds.

Hard versus smooth geometry under gradient noise What separates the two matching-norm maps is continuity (Proposition 5). We clone the same checkpoints, feed identical minibatches to both, and reduce the post-shift batch from 128 to 32 without retuning; chatter is the mean of $1 - \cos(d_t, d_{t-1})$ over consecutive directions (Table 5, appendix). At batch 128 Rook reduces chatter by 80.1% with lower trajectory CE in 9/10 seeds; at batch 32 it reduces chatter by 62.4%, halves final test loss, and improves accuracy by 2.42 points in all ten paired seeds, while the hard LMO reaches the recovery threshold 0.3 epoch sooner. The two maps therefore sit at different points of a speed and stability tradeoff, the smooth map trading a little threshold speed for continuity, steadier directions, and better final quality once gradients are noisy.

Width sweep and stationary control On a latent-permutation switch, Rook has the best post-switch trajectory error at widths $4, 8, 12, 16$ (39/40 paired runs against AdamW), with hard LMO and Rook alternating. On a stationary 12×12 assignment, Adam reaches MSE 4.7×10^{-11} against 4.5×10^{-5}; Rook targets blocks that must stay sharp while adapting, not stationary polishing.

Table 2: A 5M-parameter transformer, mean over three seeds. Left: one rewiring, 8,000 steps, at each arm's tuned step size. Right: six recurring rewirings of 750 steps. Dashes mark arms not run at that horizon. Sharpness is the normalized permanent $(\mathrm{per}(B) - \nu_n)/(1 - \nu_n)$ averaged over routes, equal to 1 at a permutation and 0 at the uniform block.

	One rewiring, 8,000 steps		Six recurring rewirings	
Stack	Final CE ↓	Sharpness	Cumulative CE ↓	Last cycle ↓
Muon+AdamW	1.0890	.443	1.6620	1.5388
+ score cap	1.0851	.321	–	–
+ hard LMO	–	–	**1.5505**	1.3752
+ Rook ($\eta = 0.75$)	1.1353	.206	1.5573	**1.3404**
+ **Rook** ($\eta = 0.25$)	**1.0837**	.318	1.5899	1.4503

A 5M-parameter transformer We repeated the protocol on a 5M-parameter byte-level model on enwik8, six layers of width 256 split into four 64-wide streams, each mixing through a learned 4×4 Birkhoff block, over three seeds (Table 2). At the small-model step size over 1,500 steps, Muon+AdamW reaches lower final CE, while the two optimizer classes settle at persistently different route sharpness, 0.23–0.25 for the matching-norm arms against 0.45–0.49 (Table 4 and Figure 2, appendix). Muon's learning rate on the scores is flat across a $30\times$ grid, so that baseline is not undertuned; Rook's tuned step size here is 0.25, and at that value it reaches the best 8,000-step endpoint of any arm. Under six recurring rewirings (Figure 1, appendix) it pays a one-time cost in the first cycle, then leads in every later cycle by $+0.16$ to $+0.20$ mean CE. The hard LMO slightly edges Rook on cumulative CE while Rook reaches the best final-cycle CE, so recurring shifts separate the optimizer class while the noise test separates the two maps.

8. Related work

Gumbel–Sinkhorn and Sinkhorn-based mixture-of-experts (MoE) routing relax assignments in the forward model [13, 16, 17]; Rook changes the optimizer's duality map for an already-explicit block [4, 8, 20]. Bethe and fractional belief propagation supply certified variational permanent approximations [1, 7, 21]. By Theorem 2, Gurvits capacity [11] is exactly the Sinkhorn part of matching free energy, and Dittert structure ties the permanent to row and column mass [6]. Muon, modular duality, and norm-constrained LMO methods all match the update geometry to the parameter block [3, 12, 18]. Closest to our setting, symmetry-compatible rules derive updates from the invariance group of a block, including permutation-equivariant updates for MoE router scores [2, 15]. Rook starts from the feasible set rather than the symmetry group, so the Birkhoff atoms fix the duality map and its coefficient, not only the equivariance class.

9. Limitations

The solver targets widths up to a few dozen, the evidence isolates optimizer behavior under rewiring and noise, and the transformer study is a single 5M-parameter model, not a pretraining claim. Rook's step size does not transfer across scale, and stationary runs favor AdamW. We do not compare against projected gradient or mirror descent on $\mathcal{B}_n$, so the advantage cannot yet be separated from primal updating in general; rectangular transportation polytopes also remain open.

References

[1] N. Anari and A. Rezaei. A tight analysis of Bethe approximation for permanent. In *Proc. FOCS*, 2019.

[2] A. Artemev et al. Exploiting weight-space symmetries for approximating curvature. In *Proc. ICML*, 2026.

[3] J. Bernstein and L. Newhouse. Modular duality in deep learning. In *Proc. ICML*, 2025.

[4] G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucumán Rev. Ser. A*, 5:147–151, 1946.

[5] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12, 2012.

[6] G.-S. Cheon and I. M. Wanless. Some results towards the Dittert conjecture on permanents. *Linear Algebra Appl.*, 436(3):791–801, 2012.

[7] M. Chertkov and A. B. Yedidia. Approximating the permanent with fractional belief propagation. *J. Mach. Learn. Res.*, 14:2029–2066, 2013.

[8] M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *NeurIPS*, 2013.

[9] G. P. Egorychev. The solution of van der Waerden's problem for permanents. *Adv. Math.*, 42, 1981.

[10] D. I. Falikman. Proof of the van der Waerden conjecture. *Math. Notes*, 29, 1981.

[11] L. Gurvits. Van der Waerden/Schrijver–Valiant like conjectures and stable (aka hyperbolic) homogeneous polynomials: one theorem for all. *Electron. J. Comb.*, 15:R66, 2008.

[12] K. Jordan et al. Muon: an optimizer for hidden layers in neural networks. Technical blog post, 2024.

[13] W. Kool, H. van Hoof, and M. Welling. Unbiased gradient estimation with balanced assignments for mixtures of experts. arXiv:2109.11817, 2021.

[14] H. W. Kuhn. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.*, 2:83–97, 1955.

[15] T. T.-K. Lau and W. Su. Symmetry-compatible principle for optimizer design: embeddings, LM heads, SwiGLU MLPs, and MoE routers. arXiv:2605.18106, 2026.

[16] G. Mena, D. Belanger, S. Linderman, and J. Snoek. Learning latent permutations with Gumbel–Sinkhorn networks. In *ICLR*, 2018.

[17] D. A. Nguyen et al. Selective Sinkhorn routing for improved sparse mixture of experts. arXiv:2511.08972, 2025.

[18] T. Pethick et al. Training deep learning models with norm-constrained LMOs. arXiv:2502.07529, 2025.

[19] R. Sanyal, F. Sottile, and B. Sturmfels. Orbitopes. *Mathematika*, 57, 2011.

[20] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21:343–348, 1967.

[21] P. O. Vontobel. The Bethe permanent of a nonnegative matrix. *IEEE Trans. Inf. Theory*, 59(3):1866–1901, 2013.

[22] Z. Xie et al. mHC: manifold-constrained hyper-connections. arXiv:2512.24880, 2025.

Appendix A. Recurring-rewiring dynamics

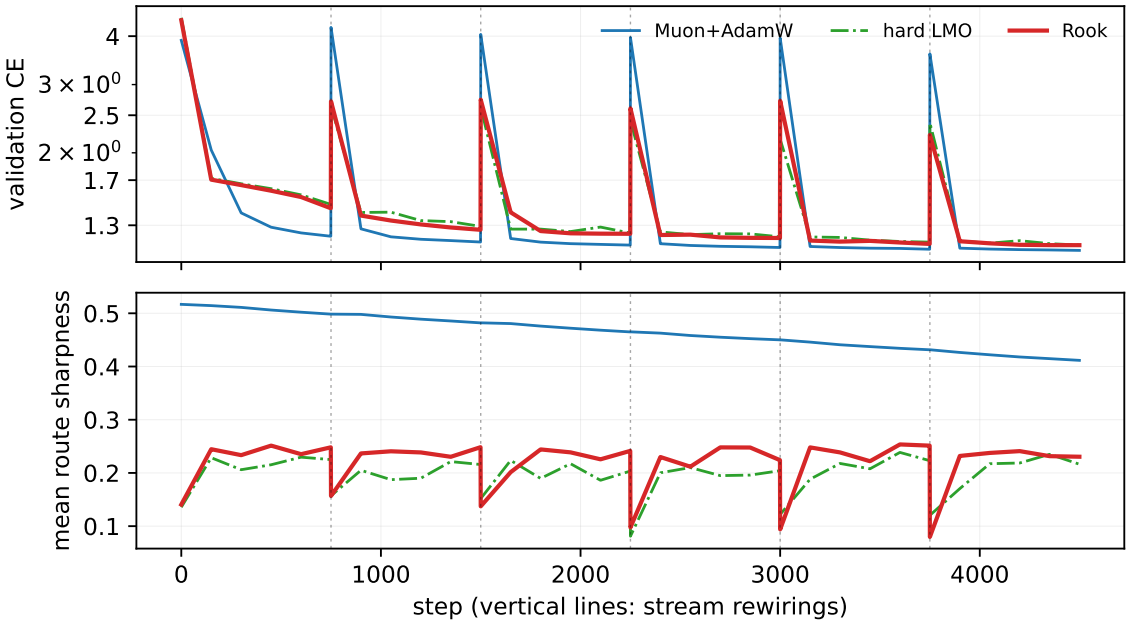

Figure 1: Six recurring rewirings of 750 steps, mean over three seeds; vertical lines mark the rewirings. Top: validation CE. Bottom: route sharpness (Table 2), averaged over routes. The matching-norm arms re-adapt more cheaply at each later rewiring, while the arm trained on the score gradient pays a larger recovery every cycle. The two classes hold persistently different route sharpness.

Appendix B. Subpermanent expansion and the exact cubic coefficient

For $E \in \mathbb{R}^{n\times n}$ and $0 \le k \le n$ let $\Sigma_k(E) = \sum_{|I|=|J|=k} \mathrm{per}(E_{I,J})$ and $(n)_k = n(n-1)\cdots(n-k+1)$.

Theorem 6 (Exact uniform subpermanent expansion) $\mathrm{per}(U+E)/\nu_n = \sum_{k=0}^{n} \frac{n^k}{(n)_k}\Sigma_k(E)$.

Proof Expand each permutation monomial $\prod_i (1/n + E_{i,\pi(i)})$ by the subset I of rows taking the E factor; a k-subset contributes $\mathrm{per}(E_{I,J})$ over its column set with $(n-k)!$ completions, giving $\mathrm{per}(U+E) = \sum_k (n-k)!\, n^{k-n}\, \Sigma_k(E)$; divide by ν_n. ■

Lemma 7 (First three invariants) *With $s = \sum_{ij} E_{ij}$, $x = E\mathbf{1}$, $y = E^\top \mathbf{1}$: $\Sigma_1 = s$ and $2\Sigma_2 = s^2 - \|x\|^2 - \|y\|^2 + \|E\|_F^2$. If $E \in \mathcal{T}_n$, additionally $\Sigma_3(E) = \frac{2}{3}\sum_{ij} E_{ij}^3$.*

Proof The first two are direct sums over nonattacking pairs. For the third, let N_3 be the ordered nonattacking-triple sum, so $N_3 = 6\Sigma_3$. Fixing the first entry (i,j), the ordered nonattacking-pair sum of the deleted submatrix equals $4E_{ij}^2 + \|E\|_F^2 - 2\sum_\ell E_{i\ell}^2 - 2\sum_k E_{kj}^2$ by the Σ_2 identity and the tangent constraints; multiplying by E_{ij} and summing leaves only $4\sum E_{ij}^3$, so $N_3 = 4\sum_{ij} E_{ij}^3$. ■

Corollary 8 (Cubic log-permanent expansion) *For tangent E small and $n \geq 3$,*

$$\log \mathrm{per}(U+E) = \log \nu_n + \frac{n}{2(n-1)}\|E\|_F^2 + \frac{2n^2}{3(n-1)(n-2)} \sum_{ij} E_{ij}^3 + O(\|E\|_F^4). \tag{10}$$

The quadratic term is the local Dittert metric of Theorem 12; the cubic term is the first orientation-sensitive skew and the source of the residual error in Table 3.

Appendix C. Audits, reproducibility, and additional tables

Solver and kernel checks Against a float64 reference with $2.5\times$ the homotopy/Newton budget over 180 matrices ($n \in \{4, 8, 16\}$), the largest objective gap is 5.7×10^{-14} and row/column residuals stay below 9×10^{-15}; the one relative marginal outlier (1.37×10^{-3}) occurs at a nearly resolved face where relative matrix error is ill-conditioned, and the corresponding signed directions still agree to 2.5×10^{-7}. The fused CPU kernel evaluates a signed direction in 0.088 ms at $n = 4$ and 1.10 ms at $n = 16$, versus 1.46 s for the exact $O(n^2 2^n)$ oracle at $n = 16$.

Reproducibility The archive contains the reference and fused solvers, exact enumeration and subset dynamic-programming oracles, all raw held-out runs and tuning logs for the small-model tables, the transformer shared-checkpoint harness with raw logs for the single-rewiring, recurring-rewiring, and long-horizon runs and for the step-size grid, the 180-case solver check, environment metadata, a SHA-256 manifest, and 22 unit tests; no held-out seed influenced any hyperparameter.

Table 3: Fidelity to the exact signed Gibbs map over 750 held-out cases.

Map	Mean cosine ↑	Minimum cosine ↑	Mean rel. error ↓
Sinkhorn ($\gamma = 0$)	0.987724	0.928120	0.216913
Classical Bethe ($\gamma = -1$)	0.990092	0.859468	0.145470
Rook (γ_n)	**0.998802**	**0.983241**	**0.039506**

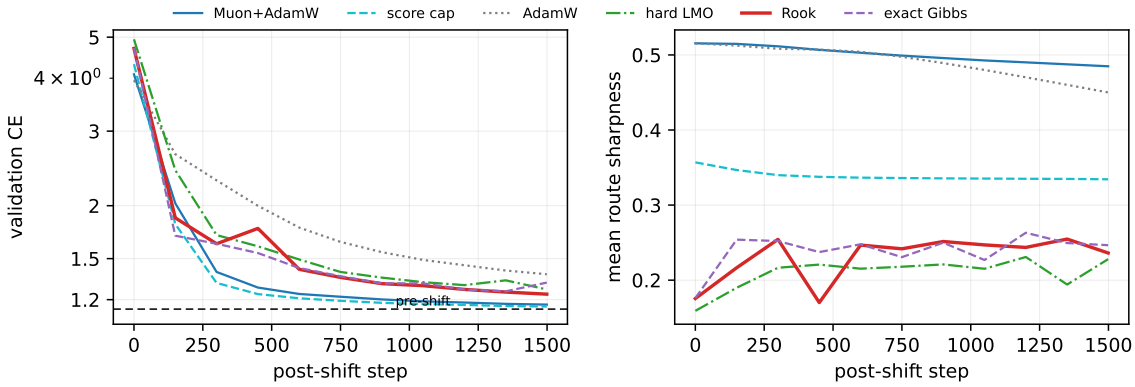

Figure 2: Transformer-scale recovery after one stream permutation, mean over three seeds. Left: validation CE; the dashed line is the pre-shift 1.140. Right: mean route sharpness, which separates the two optimizer classes throughout recovery.

Table 4: Transformer-scale rewiring from a shared checkpoint, mean over three seeds. Pre-shift validation CE is 1.140; Mean CE averages the 1,500-step recovery; sharpness is as in Table 2, averaged over the six routes at the end.

Stack	Final CE ↓	Mean CE ↓	Sharpness
AdamW	1.378	1.963	.450
Muon+AdamW	**1.168**	1.560	.485
+ score cap	1.156	1.538	.334
+ hard LMO	1.269	1.831	.228
+ **Rook**	1.236	1.736	.236
+ exact Gibbs	1.317	1.713	.247

Appendix D. Deferred statements

Theorem 9 (Difference-body oracle) *Let C be a nonempty compact convex subset of a finite-dimensional inner-product space and $K = (C - C)/2$. Then $h_K(G) = \frac{1}{2}(h_C(G) + h_C(-G))$, and if $c_+ \in \arg\max_{c\in C} \langle G, c\rangle$ and $c_- \in \arg\min_{c\in C} \langle G, c\rangle$, then $(c_+ - c_-)/2$ maximizes $\langle G, \cdot\rangle$ over K.*

Proposition 10 (Matching-norm stationarity) *Let f be bounded below and L-smooth with respect to the matching norm. Steepest descent with the oracle direction and step $\eta_t = \|\nabla f(x_t)\|_{\mathrm{match},*}/L$ satisfies $\min_{0\le t<T} \|\nabla f(x_t)\|_{\mathrm{match},*} \le \sqrt{2L(f(x_0) - f_\star)/T}$.*

Proposition 11 (Convex potential, monotonicity, limits) *A_ρ is a subgradient of the even convex potential $\frac{1}{2}\{\Phi_{\rho,\gamma_n}(G) + \Phi_{\rho,\gamma_n}(-G)\}$; hence it is odd, tangent, monotone, aligned, and bounded by $\sqrt{n}$. As $\rho \to \infty$, $A_\rho(G) = G/[\rho(n-1)] + O(\rho^{-3})$; as $\rho \downarrow 0$ (unique extremes), $A_\rho(G) \to \frac{1}{2}(P_{\max} - P_{\min})$; temperature is thus a homotopy between tangent-gradient and assignment-oracle geometry inside one optimizer.*

Table 5: Hard versus smooth map under gradient noise, mean ± standard error over ten seeds.

Batch	Map	Chatter ↓	Test CE ↓	Accuracy (%) ↑
128	Hard LMO	$.178 \pm .013$	$.110 \pm .013$	$96.98 \pm .59$
128	**Rook**	$\mathbf{.035 \pm .002}$	$\mathbf{.098 \pm .005}$	$\mathbf{97.47 \pm .24}$
32	Hard LMO	$.182 \pm .004$	$.149 \pm .034$	95.42 ± 1.06
32	**Rook**	$\mathbf{.069 \pm .001}$	$\mathbf{.073 \pm .008}$	$\mathbf{97.84 \pm .29}$

Theorem 12 (Local Dittert metric) *For E with entries summing to zero and $U + E \geq 0$, writing $x = E\mathbf{1}$, $y = E^\top \mathbf{1}$, $T = \Pi_{\mathcal{T}_n} E$,*

$$(2 - \nu_n) - \Phi_D(U + E) = \frac{\nu_n n}{2(n-1)} \|T\|_F^2 + \frac{1 - \nu_n}{2} \big(\|x\|_2^2 + \|y\|_2^2\big) + O(\|E\|_F^3). \tag{11}$$

On the Birkhoff slice ($x = y = 0$): $\log \operatorname{per}(U + E) = \log \nu_n + \frac{n}{2(n-1)} \|E\|_F^2 + O(\|E\|_F^3)$.

Proposition 13 (Stability under inexact solves) *If each numerical marginal differs from its exact fractional optimum by at most ε, the unnormalized signed map differs by at most ε, and if $\|A_\rho(G)\|_F \geq 2\varepsilon$ then the normalized direction differs by at most $2n\varepsilon / \|A_\rho(G)\|_F$.*

Appendix E. Proof sketches

Difference body (Theorem 9). Support functions add under Minkowski combination and satisfy $h_{-C}(G) = h_C(-G)$, so $h_{(C-C)/2}(G) = \frac{1}{2}\big(h_C(G) + h_C(-G)\big)$. The point $(c_+ - c_-)/2$ lies in K and attains this value, so it maximizes $\langle G, \cdot \rangle$ over K.

Stationarity (Proposition 10). Write $g_t = \|\nabla f(x_t)\|_{\mathrm{match},*}$. The oracle direction d_t has unit matching norm and satisfies $\langle \nabla f(x_t), d_t \rangle = g_t$, and smoothness with respect to the norm pair gives $f(x_{t+1}) \leq f(x_t) - \eta_t g_t + L\eta_t^2/2$, so the stated step decreases f by at least $g_t^2/(2L)$; summing over $t < T$ and telescoping gives the bound.

Covariance (Proposition 1). At $G = 0$ the Gibbs law is uniform. For tangent R, $\mathbb{E}[P_{ij} \langle R, P \rangle] = \frac{1}{n}(R_{ij} + \frac{1}{n-1} \sum_{k \neq i, \ell \neq j} R_{k\ell}) = R_{ij}/(n-1)$, since the deleted-submatrix sum equals R_{ij} under zero row/column sums.

Factorization (Theorem 2). With $B = \operatorname{diag}(a) e^{G/\tau} \operatorname{diag}(b)$ doubly stochastic, the unit row and column sums give

$$H(B) = -\langle G, B \rangle / \tau - \textstyle\sum_i \log a_i - \sum_j \log b_j, \qquad F_\tau(G) = -\tau\big(\sum_i \log a_i + \sum_j \log b_j\big).$$

Permanent homogeneity under diagonal scaling gives $\operatorname{per}(B) = (\prod_i a_i)(\prod_j b_j) \operatorname{per}(e^{G/\tau})$. The capacity identity is the matrix-scaling characterization of cap; on $\mathcal{B}_n$, weighted AM–GM gives $\prod_i (Bx)_i \geq \prod_j x_j$ with equality at $x = \mathbf{1}$.

Curvature match (Theorem 3). The coordinate Hessian of $-H_\gamma$ is $1/B_{ij} + \gamma/(1 - B_{ij})$, equal at U to $n(n - 1 + \gamma)/(n - 1)$ times the identity on $\mathcal{T}_n$; equate its inverse with $1/(n-1)$.

Accuracy (Theorem 4). The quadratic bound combines the exact-map remainder $\|M(S) - U - S/(n-1)\|_F \le 4n^{3/2}\|S\|_F^2$ (third cumulant of the assignment statistic, $\|P - M\|_F \le 2\sqrt{n}$, $|X - \bar{X}| \le 2\sqrt{n}\|S\|_F$) with the fractional inverse-map remainder $L_n\|S\|_F^2/(2\mu_n^3)$, obtained from μ_n-strong tangent curvature in the cube $\|B - U\|_\infty \le 1/(2n)$ (where $1/b \ge 2n/3$ dominates γ_n's negative term) and the coordinate third-derivative bound L_n, via a first-exit argument along the solution path. For the cubic statement, write both maps' expansions $U + S/(n-1) + Q_\bullet(S,S) + C_\bullet(S,S,S) + \ldots$; odd symmetrization cancels U and both quadratic terms, and the linear terms coincide by Theorem 3, leaving a cubic-order difference with constant C_n. Here $\mu_n = \frac{2n}{3} - \frac{2(n-1)}{2n-3}$ and $L_n = 4n^2 + \frac{4n(n-1)}{(2n-3)^2}$, and C_n is assembled from the third-cumulant bound $4n^{3/2}$, the coordinate third-derivative bound L_n, and the curvature floor μ_n; we do not optimize its numerical value, and the measured log–log slopes confirm the exponent it certifies. Normalization divides by $\mathrm{RMS}(A)$, which itself carries an $O(\|S\|^2)$ relative perturbation, restoring second order.

Lipschitz (Proposition 5). Strong monotonicity of the restricted gradient field gives $\mu\|R(S_1) - R(S_2)\|^2 \le \langle S_1 - S_2, R(S_1) - R(S_2)\rangle$, and Cauchy–Schwarz completes the bound. The statement about ties follows because on a hyperplane where two assignments tie, arbitrarily small tangent perturbations change the selected maximizer by a fixed nonzero difference of permutations.

Appendix F. Exact oracle and protocol

The exact validation oracle computes assignment marginals by forward/backward subset dynamic programs over column subsets in $O(n^2 2^n)$; it is used for the 750-case study and the width ceiling only. Primary settings: Rook learning rate 0.75, momentum 0.9, relative temperature 1.0, score radius 0.75, $K = 8$ homotopy stages, $I = 6$ Newton iterations; router forward: log-domain Sinkhorn, temperature 0.7, 20 iterations; Muon on hidden 2D weights, with its learning rate swept on the development seeds; AdamW on biases, norms, embeddings, and head. Chatter is $\mathrm{mean}_t\,[1 - \cos(d_t, d_{t-1})]$. The tuning CSVs contain the full sweeps rather than selected points; the arms receive identical minibatches after cloning; every table reports held-out seeds disjoint from tuning seeds.