THE DYNAMICS OF ROUTING IN DEEP LINEAR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how architecture and initialization shape the routing of computation in linear networks on directed acyclic graphs. Hidden basis changes yield conserved matrix imbalances that constrain the allocations accessible to learning. For positive scalar networks, we derive an exact capacity-flow equation: each edge's capacity changes in proportion to its usage. In balanced networks of uniform path depth, this separates overall growth from autonomous routing dynamics. Stable limiting routing states are equivalent to chains. This separation connects routing instability to path memory, kernel evolution to conductance depth, and recruitment of dormant computations to the number and compatibility of edges that must activate together. Selected matrix and ReLU experiments illustrate related learning phenomena beyond the scalar analysis.

1 INTRODUCTION

Networks often offer several routes for the same task, yet training need not use them equally. Mixtures of experts can concentrate computation on a few experts (Shazeer et al., 2017; Fedus et al., 2022), and multimodal networks can neglect informative streams (Huang et al., 2022). We ask when learning sustains competing routes, concentrates onto a subgraph, or recruits a weak computation. We study linear networks on directed acyclic graphs (DAGs), where paths branch, merge, and share trainable edges.

We show that conserved matrix imbalances restrict learning to spaces determined by architecture and initialization. For positive scalar networks, squared weights are exact capacity coordinates, and gradient flow changes each capacity in proportion to edge usage. In balanced networks of equal path depth, we separate overall growth from routing: the graph and initial allocation determine the routing trajectory, while the loss determines its timing and stopping point.

We show that stable limiting routing states in this regime are equivalent to chains, and express local routing stability through path covariance. Building on Kunin et al. (2024), we derive a concrete conductance formula for instantaneous richness. By analyzing boundary fixed points, we show when a small seed grows into a new subnetwork through cooperation among dormant edges. We extend the routing analysis to mixed depths and nonzero imbalances, and illustrate selected phenomena in matrix and ReLU networks.

2 DAG NETWORKS

Let $\Gamma = (\mathcal{V}, \mathcal{E})$ be a finite directed acyclic graph (DAG), with inputs $\mathcal{V}_{\text{in}}$ at sources, outputs $\mathcal{V}_{\text{out}}$ at sinks, and hidden vertices $\mathcal{V}^\circ = \mathcal{V} \setminus (\mathcal{V}_{\text{in}} \cup \mathcal{V}_{\text{out}})$. Every edge lies on an input-to-output path. A vertex represents a real inner-product space V_v of dimension d_v: a collection of neurons, rather than an individual neuron. An edge e runs from its input vertex v_e^{in} to its output vertex v_e^{out} and carries a trainable matrix $w_e : V_{v_e^{\text{in}}} \to V_{v_e^{\text{out}}}$. Let W be the space of all weight collections $w = (w_e)_{e \in \mathcal{E}}$.

Hidden and output vertices sum transformed incoming activations. Let $\mathcal{P}$ be the set of complete input-to-output paths and $\mathcal{P}_{io}$ those from input i to output o. Each pair defines a map $F_w^{oi} : V_i \to V_o$:

$$F_w^{oi} = \sum_{P \in \mathcal{P}_{io}} w_P, \qquad w_P = w_{e_m} \cdots w_{e_1} \quad \text{for } P = (e_1, \ldots, e_m). \tag{1}$$

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.

Given inputs (x_i), output o is $\sum_i F_w^{oi} x_i$; write F_w for this full predictor. We train a smooth loss $L(w) = \ell(F_w)$ by Euclidean gradient flow, $\dot{w}_e = -\nabla_{w_e} L$, with squared parameter norm $\|w\|_F^2 = \sum_e \|w_e\|_F^2$. A common learning rate is absorbed into time; there is no weight decay.

Twin collapse. Vertices with the same predecessors, successors, and role are twins. Concatenating their activations and associated weight matrices merges them while preserving the predictor and gradient flow. Figure 1, chosen for its simplicity, becomes a chain by merging its two scalar outputs into one two-dimensional output.

3 From conserved quantities to edge capacities

Loss-preserving transformations yield conserved squared-weight imbalances in linear and ReLU networks (Du et al., 2018; Zhao et al., 2023; Marcotte et al., 2023). Their geometry has recently been studied for ReLU DAGs (Nurisso et al., 2026). For linear DAGs, hidden changes of basis give a matrix conservation law (Appendix D):

Proposition 1 (Conserved matrix imbalance)**.** *At every hidden vertex, gradient flow preserves the symmetric matrix*

$$\boxed{c_v(w) = \sum_{e: v_e^{\text{in}} = v} w_e^T w_e - \sum_{e: v_e^{\text{out}} = v} w_e w_e^T.} \tag{2}$$

The conserved matrices, or *charges*, restrict learning to the level set

$$W_c = \{w : c_v(w) = c_v \text{ for all } v \in \mathcal{V}^\circ\}.$$

Taking squared edge norms maps this set into a capacity polyhedron:

$$\rho : W_c \longrightarrow \mathcal{P}_{\text{tr}\, c} := \{r \in \mathbb{R}^{\mathcal{E}}_{\geq 0} : Br = (\text{tr}\, c_v)_v\}, \qquad \rho(w)_e = r_e := \|w_e\|_F^2. \tag{3}$$

Here B is the hidden-vertex incidence matrix, with $+1$ at an edge's input vertex and -1 at its output vertex; the constraint follows by taking traces of (2). Learning first restricts the full weight space W to W_c. Quotienting by the orthogonal hidden basis changes K_c that preserve c then gives $\mathcal{M}_c = W_c / K_c$ (Figure 1). Both steps can reduce dimension; for scalars, K_c is finite, so conservation supplies the entire dimension reduction. Capacities are unchanged by K_c, so ρ factors through $\mathcal{M}_c$, although capacities need not determine a matrix network's remaining state.

For *positive scalar weights*, capacities do determine the state: $w_e = \sqrt{r_e}$ identifies the positive sector of W_c with $\mathcal{P}_c \cap \mathbb{R}^{\mathcal{E}}_{>0}$. Allowing zero weights gives the whole capacity polyhedron $\mathcal{P}_c$. It is generally unbounded; fixing total input capacity gives a bounded polytope.

For a connected scalar DAG with one input and output, conservation can greatly reduce the $|\mathcal{E}|$-dimensional weight space: at positive states, $\dim \mathcal{M}_c = |\mathcal{E}| - |\mathcal{V}^\circ|$, the number of independent undirected loops plus one. Fixing total input capacity removes the extra scale coordinate; a chain has no routing freedom (Figure 2; Appendix B).

Capacity coordinates carry a canonical flat metric in the edge basis, different from the metric inherited from weight space. The inherited metric is Hessian, and the path sum defines an exponential family. Appendix C develops their role in the gradient and covariance identities.

4 Capacity flow and routing dynamics

For positive scalar weights with one input and output, write $r_e = w_e^2$, $z_P = \prod_{e \in P} w_e$, and $f = \sum_P z_P$. Define path usage, edge usage, and total input capacity by

$$\pi(P) = z_P / f, \qquad u_e = \sum_{P \ni e} \pi(P), \qquad Q = \sum_{e \in C_{\text{in}}} r_e,$$

where C_{in} contains the edges leaving the source. On a growth interval, Appendix E derives

$$\boxed{\frac{dr}{dQ} = u}, \qquad \dot{Q} = -2\ell'(f) f > 0.$$

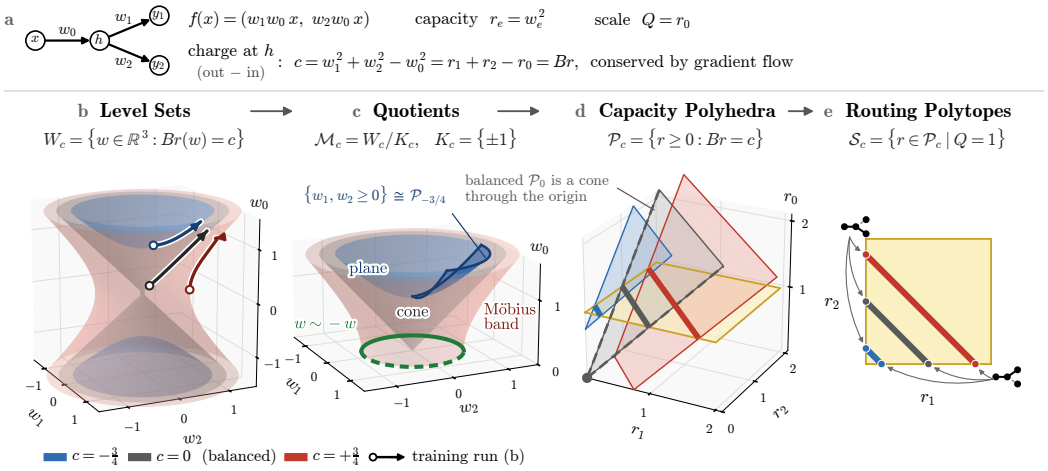

Figure 1: **Conservation reduces learning to a smaller geometric problem.** On the Y network, gradient flow stays on a surface of fixed imbalance $c = r_1 + r_2 - r_0$, where $r_i = w_i^2$. Panels (b)–(e) follow the reduction from weights to the quotient $\mathcal{M}_c$, then to capacities and a fixed-scale routing segment. The quotient identifies w with $-w$; squaring weights gives exact coordinates on its positive sector. At $Q = r_0 = 1$, the segment's endpoints use only one output edge.

Each edge gains capacity in proportion to its current usage. Since $Bu = 0$, this preserves $Br = c$. Any cut meeting every path exactly once gives the same clock up to a conserved additive shift (Appendix E).

To separate routing from overall growth, we first assume *balance* ($c = 0$) and *equal depth*: every input-to-output path has m edges. Write $r = Qs$, where s lies in the routing polytope

$$\mathcal{S}_0 = \left\{ s \geq 0 : Bs = 0,\ \sum_{e \in C_{\text{in}}} s_e = 1 \right\}.$$

Boundary limits $s_e \to 0$ describe relative pruning, $r_e/Q \to 0$. Differentiation gives $ds/d\log Q = u(Q, s) - s$. Equal depth gives

$$f(Q, s) = Q^{m/2}\Phi(s), \qquad \Phi(s) = \sum_P \prod_{e \in P} \sqrt{s_e},$$
$$\boxed{\frac{ds}{d\log Q} = u(s) - s}. \tag{4}$$

Here Φ is productivity: output relative to capacity scale. The common scale factor cancels from path usage, making routing autonomous; loss and target determine the clock and stopping scale. Section 6 treats charge and unequal depth.

The fixed points of (4) satisfy $u(s_\star) = s_\star$: each edge's share of computation equals its share of capacity, and learning follows the ray $r(Q) = Qs_\star$. Routing is gradient ascent of $\log \Phi$ in the inherited metric, increasing productivity and decreasing $\|w\|^2/|f|^{2/m}$. Appendix G proves the strict Lyapunov identity, convergence, and global norm minimality of stable states. Figure 2 compares the attracting sets of different graphs.

5 Collapse, Routing Memory, and Coherence Depth

We study routing stability in balanced, equal-depth networks.

The depth-two transition for parallel routes. For three balanced, disjoint routes of depth d, capacity is constant along each route, with shares $s_1 + s_2 + s_3 = 1$, and

$$\frac{ds_i}{d\log Q} = \frac{s_i^{d/2}}{\sum_j s_j^{d/2}} - s_i.$$

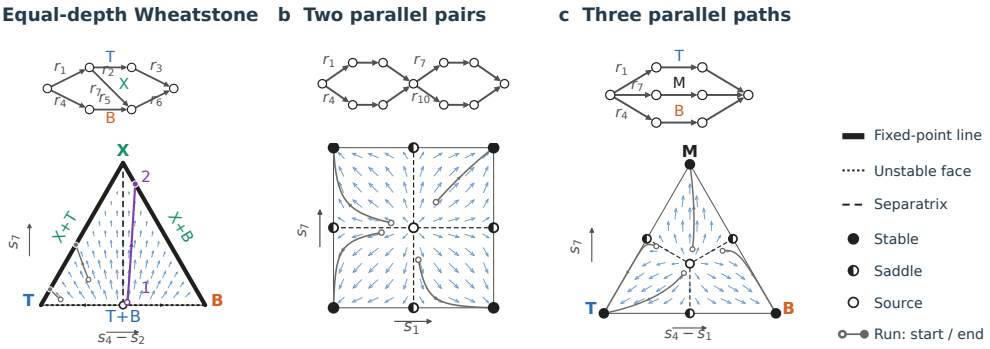

Figure 2: **Routing polytopes and their dynamics.** Each balanced, equal-depth scalar graph has a two-dimensional routing polytope; trajectories show how learning redistributes capacity. (a) In the Wheatstone graph, the bridge route X survives together with T or B; the thick edges are continuous families of fixed points. Purple trajectories begin with 2% and 85% bridge allocation. (b,c) The series composition and three disjoint paths instead approach isolated corners. Arrows indicate routing direction; dashed curves separate basins.

The symmetric state $s_i = 1/3$ is fixed, with linear growth rate $d/2 - 1$ for routing perturbations. Thus $d = 1$ equalizes the routes, $d = 2$ leaves every routing state fixed, and $d = 3$ amplifies route differences, favoring specialization (Figure 3).[1] For $d \neq 2$, this is the only interior fixed point. For disjoint routes, depth alone determines this stability threshold. To account for shared edges, we now express the criterion in terms of path statistics.

Local stability: coherence depth and the $\lambda = 2$ threshold. At an interior fixed point $s_\star$ with nonzero routing dimension, the largest routing-Jacobian eigenvalue is $\lambda_{\max}/2 - 1$. Appendix F expresses $\lambda_{\max}$ through path usage $\pi = \pi_{s_\star}$. Grade the equal-depth DAG into vertex layers $0, \ldots, d$, with edges joining consecutive layers. Let $e_\ell(P)$ be the edge traversed between layers $\ell - 1$ and ℓ, and $F_h(P) = \sum_{e \in P} h_e$ for edge values h_e. Then

$$\lambda_{\max} = \max_{h:\, \mathrm{Var}_\pi(F_h) > 0} \sum_{\ell=1}^{d} \frac{\mathrm{Var}_\pi(\mathbb{E}_\pi[F_h(P) \mid e_\ell(P)])}{\mathrm{Var}_\pi(F_h)}.$$

Perturbations decay for $\lambda_{\max} < 2$ and grow for $\lambda_{\max} > 2$; equality gives a marginal linear mode. Parallel routes recover $\lambda_{\max} = d$.

Each summand measures how informative the edge traversed in a given layer is about the path as a whole, quantified by the fraction of variance in $F_h(P)$ it explains. Intuitively, an edge reveals more about the route when paths passing through different edges in its layer also tend to use different edges in other layers. Thus this *coherence depth* measures routing memory rather than architectural depth alone. Disjoint parallel paths retain route identity throughout, whereas shared connections allow paths to mix. In Figure 3, we compare routing memory and specialization across equal-depth ReLU architectures.

Global structure of stable routing states. We characterize stable supports using a free energy that measures dependence between adjacent layers. Given a unit routing flow s, sample a path by starting at the source and choosing each outgoing edge $u \to v$ with probability $s_{uv} / \sum_{u \to w} s_{uw}$. This capacity-based distribution need not equal the path-usage distribution π. Let X_ℓ be the vertex visited in layer ℓ, with vertex measure $m_\ell(v) = \Pr(X_\ell = v)$. Appendix G shows that routing decreases the free energy

$$\mathcal{F}(s) = \frac{1}{2} \sum_{\ell=1}^{d} I(X_{\ell-1}; X_\ell),$$

where I denotes mutual information. Thus $\mathcal{F} \geq 0$, with equality exactly when adjacent layer variables are independent. Together with the stability analysis, this characterizes stable fixed points: their active supports are layerwise complete, with product capacities $s_{uv} = m_{\ell-1}(u) m_\ell(v)$ and $\mathcal{F} = 0$; other

[1] The $d = 1$ example permits parallel edges between the same two vertices.

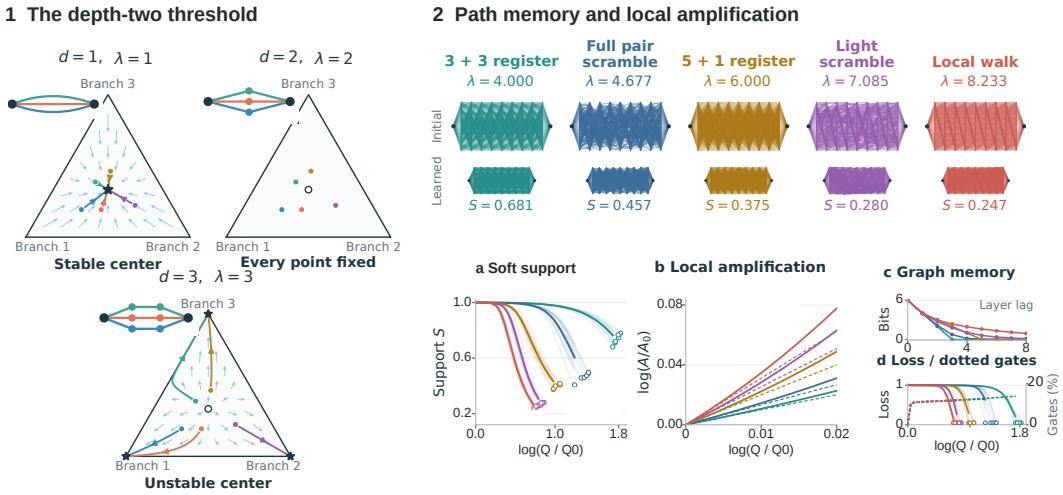

Figure 3: **Path memory controls whether routing differences grow.** (1) Parallel scalar routes show the depth-two transition. (2) Five equal-depth ReLU architectures show why depth alone does not determine specialization. In (a), architectures with larger local amplification rates leave the broadly distributed initialization sooner and finish with capacity concentrated on fewer edges, reflected in lower soft support. (b) Measured amplification follows the ordering of the scalar rates $\lambda/2 - 1$, while (c) relates this ordering to how long the graph retains route information. (d) shows loss and gate changes. Support curves summarize six seeds. The horizontal learning coordinate is $\log(Q/Q_0)$. Scalar reference eigenvalues label the architectures; experimental definitions are in Appendix K.

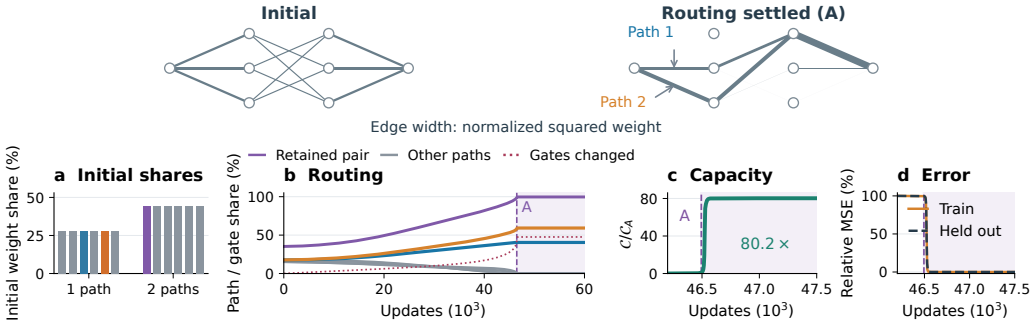

Figure 4: **Learning first selects routes, then improves the selected computation.** Despite equal initial squared-weight allocations among candidate supports of the same size (a), a width-32 ReLU network selects a two-path support that becomes a chain after twin aggregation. By A, their functional path shares have nearly settled (b); capacity then grows 80.2-fold (c) while training and held-out errors fall (d). Edge widths show normalized squared weights, and the dotted curve counts gates that have changed. Individual weights and gates remain free to adapt.

fixed points are unstable. **Every stable limiting routing state therefore collapses to a chain up to twin aggregation (Section 2).** Learning can simplify a branched computation to a chain without retaining just one scalar path: several vertices may survive in a layer and combine into a larger feature space (Figure 9). The statement concerns normalized routing as $Q \to \infty$; fitting a finite target can stop learning earlier.

In Figure 4, we show routing selection followed by further growth and fitting in a ReLU network. At a scalar routing fixed point, only capacity scale changes; Figure 12 shows a scalar counterpart.

Varying the positive vertex measures gives continuous families of fixed points, so chain collapse can retain memory of the initial capacity allocation.

6 GENERALIZATION TO MIXED DEPTH AND GENERAL CHARGE

Retaining the input-cut scale Q and normalized capacities $s = r/Q$, the routing equation becomes

$$\frac{ds}{d\log Q} = u(Q, s) - s, \qquad Bs = \frac{c}{Q}.$$

Here scale enters in two ways: unequal path lengths make usage depend on Q, while conserved charge changes the feasible routing states.

Depth races. Write $d_P = |P|$ and $\psi_P(s) = \prod_{e\in P} \sqrt{s_e}$, so that $z_P = Q^{d_P/2}\psi_P(s)$. The ratio

$$\frac{z_P}{z_R} = Q^{(d_P - d_R)/2}\frac{\psi_P(s)}{\psi_R(s)}$$

implies a depth race: at fixed routing, growth favors longer paths. For the usage-weighted effective depth $D_{\text{eff}} = \mathbb{E}_\pi[d_P]$, Appendix H gives

$$\frac{dD_{\text{eff}}}{d\log Q} = \frac{1}{2}\operatorname{Var}_\pi(d_P) + \operatorname{Cov}_\pi\left(d_P, \frac{d\log\psi_P(s)}{d\log Q}\right).$$

At fixed s, only the nonnegative variance term contributes: growth favors depth more strongly when used path lengths vary more. The covariance records changes in routing, which can reweight paths of different lengths upward or downward; effective depth can therefore decrease. In Figure 5a, the same initial routing follows different trajectories at different initial scales.

Conserved charge as a transient routing bias. The normalized state lies in the moving family

$$\mathcal{S}_{c,Q} = \left\{ s \geq 0 : Bs = c/Q,\ \sum_{e\in C_{\text{in}}} s_e = 1 \right\},$$

where C_{in} is the input cut. Charge can exclude pruning faces at finite scale, favor one branch over another, delay recruitment, and redirect the trajectory toward a different limiting face. In Figure 5b, a transient charge changes which pair of routes survives.

Asymptotic recovery of the balanced homogeneous theory. At large capacity,

$$Bs = \frac{c}{Q} \longrightarrow 0.$$

Appendix H proves convergence to balanced homogeneous routing on the surviving support. If H is the edge support of the selected limiting face, the surviving depth is $d_H = \max\{d_P : P \subset H\}$, the largest depth remaining on that face. This can be smaller than the original graph's maximum depth: transient routing can suppress the globally deepest alternatives.

The limit need not be stable. Stable endpoints are balanced, homogeneous product states and hence also twin-collapse to chains.

7 RICH AND LAZY LEARNING: A CONDUCTANCE LAW

Building on the analysis of initialization imbalance by Kunin et al. (2024), we derive an explicit formula for richness in positive scalar DAGs. In *lazy* training, the model linearized at initialization approximates learning; *rich* training involves substantial changes in tangent features, the output derivatives with respect to parameters (Chizat et al., 2019; Woodworth et al., 2020). These derivatives define the scalar neural tangent kernel $\mathcal{K} = \|\nabla_w f\|^2$. We measure its local evolution by *instantaneous richness* $\mathcal{R} = d\log\mathcal{K}/d\log f$: small $\mathcal{R}$ means little kernel change per relative predictor change. Laziness over a whole task also depends on how far f travels.

To express the kernel in terms of capacities and usage, define edge drops $\gamma_e = fu_e/r_e$ and path drops $\sigma_P = \sum_{e\in P}\gamma_e$. With capacities as conductances and fu_e as current, the kernel is dissipated power:

$$\mathcal{K} = \sum_e \frac{(fu_e)^2}{r_e} = f\mathbb{E}_\pi[\sigma_P], \qquad \dot{f} = -\ell'(f)\mathcal{K}. \tag{5}$$

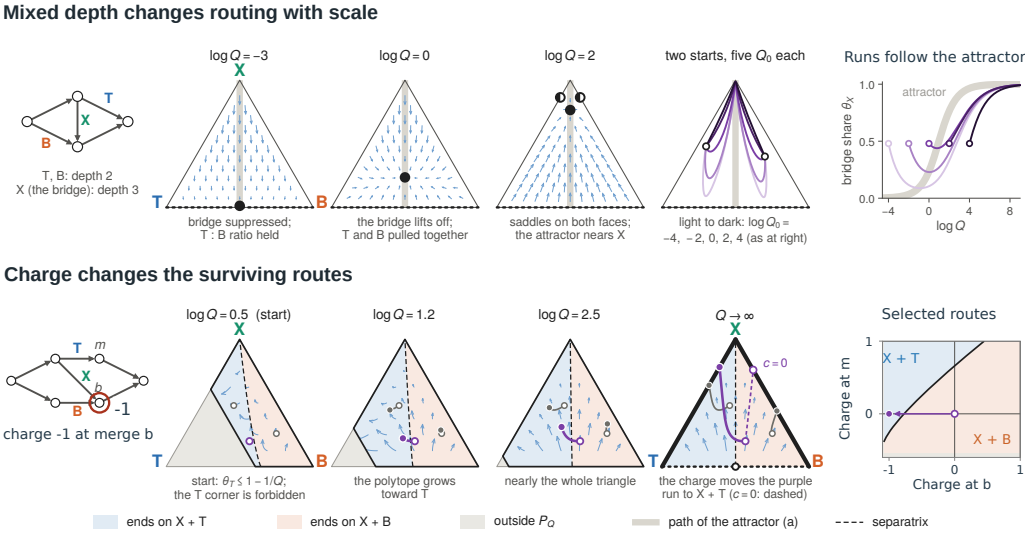

Figure 5: **Routing dynamics with mixed depth and nonzero charge.** (a) The balanced graph combines depth-two routes T and B with a depth-three bridge route X. Increasing Q moves the attractor toward X; runs from the same routing state but different initial scales follow different trajectories. (b) In the equal-depth graph, incoming excess at the merge initially excludes the pure-T corner. The feasible region expands toward the balanced triangle, yet the resulting routing history can change the limiting support from X+B to X+T. The rightmost panel maps the surviving route pair as the charges at m and b vary: constraints that vanish after normalization can still select different computations. T, B, and X denote the complete routes; the dashed comparison has zero charge.

Conductance depth and path dispersion. Define effective conductance depth and path dispersion (squared coefficient of variation), measuring how drops are shared within paths and how total drops vary between paths:

$$m_{\text{eff}} = \frac{\mathbb{E}_\pi[(\sum_{e\in P}\gamma_e)^2]}{\mathbb{E}_\pi[\sum_{e\in P}\gamma_e^2]}, \qquad \text{CV}_\pi^2(\sigma) = \frac{\text{Var}_\pi(\sigma_P)}{\mathbb{E}_\pi[\sigma_P]^2}. \tag{6}$$

Conductance depth counts how many edges share the voltage drop along a path: one dominant drop gives depth near one, while equal drops on a length-m path give depth m. It differs from mean path length and routing coherence depth. Appendix I derives

$$\boxed{\mathcal{R} = 2\left(1 + \text{CV}_\pi^2(\sigma)\right)\left(1 - \frac{1}{m_{\text{eff}}}\right).} \tag{7}$$

A richness formula for chains. A chain has one path and no path dispersion, giving

$$m_{\text{eff}} = \frac{(\sum_e r_e^{-1})^2}{\sum_e r_e^{-2}}, \qquad \mathcal{R} = 2(1 - m_{\text{eff}}^{-1}). \tag{8}$$

One edge much smaller than the others gives $m_{\text{eff}} \approx 1$ and small instantaneous richness; two comparably small edges give $m_{\text{eff}} \approx 2$ and $\mathcal{R} \approx 1$. One large edge still leaves several others sharing the drop, so imbalance alone does not imply laziness. Figure 6A compares the resulting learning curves and kernel evolution. Panel B tests related predictions through function-preserving rescalings of ReLU networks, illustrating possible extensions beyond the exact scalar identities. Figure 10 explores a consequence for transfer after training.

8 Nucleation: Recruiting Dormant Computations

We next analyze routing fixed points on the boundary, where some edges are dormant. A small perturbation can grow into a new subnetwork when dormant edges reinforce one another through active connections (Figure 7).

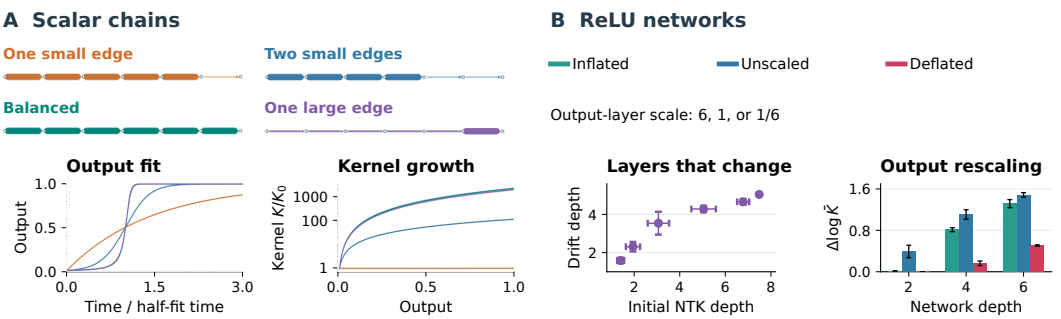

Figure 6: **A conductance law predicts rich and lazy learning in chains.** (A) Scalar chains start with the same output. With one very small edge, the network fits the target while the kernel stays near its initial value; several comparably small edges allow much stronger kernel change. A single large edge leaves the other edges able to learn. Time is divided by each run's half-fit time. (B) Function-preserving rescalings in two ReLU studies show corresponding patterns: initial NTK depth tracks drift depth (the effective numbers of layers contributing to the initial kernel and to representation change), and shrinking the output layer suppresses NTK growth more than enlarging it in deeper networks. Points and bars are two-seed means; whiskers span the two runs. Definitions and fixed training budgets are in Appendix K.

Activation order. In a balanced equal-depth network, let $s^\star$ be a routing fixed point on active support H, with dormant edges $Z = \mathcal{E} \setminus H$. The activation order $k = \min_{P \not\subset H} |P \cap Z|$ counts the fewest dormant edges needed to complete a new path. The boundary expansion in Appendix J gives three regimes: $k = 1$ permits immediate recruitment; $k = 2$ has a collective threshold; $k > 2$ makes sufficiently small routing seeds decay locally, so growth requires a finite perturbation. The relevant order counts dormant edges, not total path depth: reusing active connections can make a long route easier to recruit.

Collective recruitment. To determine the $k = 2$ threshold, we collect the contributions of active connections to each dormant edge pair. Set $\phi_H = \Phi(s^\star)$ and define the symmetric coactivation matrix on dormant edges by $A_{ee} = 0$ and

$$A_{ef} = \sum_{P:\, P \cap Z = \{e,f\}} \prod_{a \in P \cap H} \sqrt{s_a^\star} \quad (e \neq f). \tag{9}$$

Each entry measures how strongly active connections join a dormant pair into complete paths. The threshold $\lambda_{\max}(A) > \phi_H$ gives a growing linear mode if its direction can be extended to a perturbation satisfying balance; equality is marginal. When this mode grows and the coactivation graph (edges with $A_{ef} > 0$) is connected, its positive leading eigenvector gives the relative weight amplitudes of the growing seed.

In Figure 7a, one recruited route supplies a connection used by the next, changing recruitment thresholds without changing the architecture. Panel b tests the $k = 2$ threshold through junction sharing.

9 Related Work

Conservation and learning geometry. Deep linear networks reveal nonlinear learning dynamics and parameterization-dependent implicit bias (Saxe et al., 2014; Arora et al., 2019). We build on conserved imbalances and symmetry reductions (Du et al., 2018; Kunin et al., 2021; Zhao et al., 2023; Marcotte et al., 2023), and quiver formulations (Ganev and Walters, 2022; Jeffreys and Lau, 2021). In closely related work, Nurisso et al. (2026) formulate squared-weight conservation on ReLU DAGs and analyze invariant-set connectivity, singularities, and pruning. We study capacity evolution within these conserved levels.

Pathway competition and specialization. Shi et al. (2022) analyze specialization in parallel linear pathways. The *neural race reduction* treats shared trainable modules on architecture graphs and

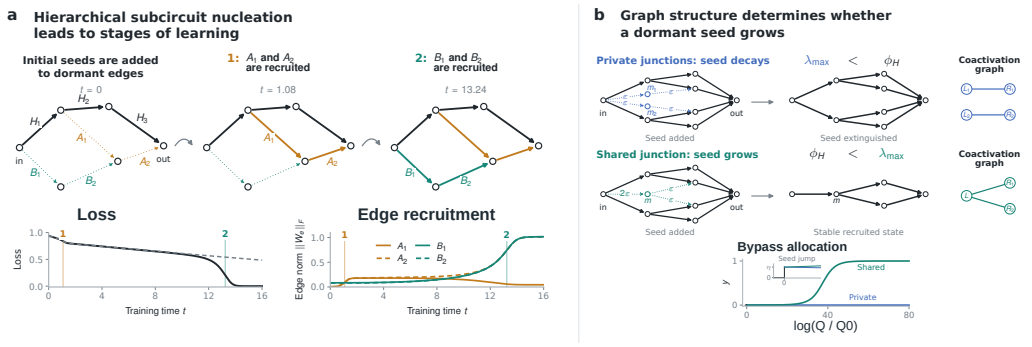

Figure 7: **Small perturbations can recruit dormant subnetworks.** (a) We add small seeds to dormant matrix edges and follow gradient descent. The network first recruits A_1, A_2, then reuses A_2 to recruit B_1, B_2; loss falls in stages. Replacing the shared edge by an independently trained copy delays the second stage (dashed). (b) Each new scalar path needs two dormant edges ($k = 2$). Sharing the junction raises the leading coactivation eigenvalue above ϕ_H, turning decay into growth for the same total seed; all paths have depth three. The curves follow the equal-arm initialization described in the appendix.

explains learning advantages from reuse (Saxe et al., 2022). Jarvis et al. (2025) connect imbalance and readout initialization to specialization: one pathway can nearly finish learning before a competitor starts. We classify stable supports and relate instability to path covariance in balanced, equal-depth positive scalar DAGs, where multiple paths can survive.

Path representations and reduced dynamics. Path-based optimization, complexity measures, and kernels describe architecture (Neyshabur et al., 2015; Gonon et al., 2024; Gebhart et al., 2021). Marcotte et al. (2026) derive intrinsic dynamics through path lifting for ReLU DAGs. SynFlow connects conserved saliency to squared-weight dynamics (Tanaka et al., 2020); its normalized saliency equals edge usage in our positive scalar setting. We study the resulting scale–routing separation, stability, memory, and recruitment.

Effective depth and recruitment. Kernel and lazy-training theory (Jacot et al., 2018; Chizat et al., 2019) and analyses of initialization scale, shape, and imbalance (Woodworth et al., 2020; Azulay et al., 2021; Kunin et al., 2024; Dominé et al., 2025) explain how parameterization governs kernel evolution. Residual-network studies distinguish architectural depth from contributing path lengths (Veit et al., 2016; Littwin and Wolf, 2016); we distinguish coherence depth governing routing stability and conductance depth entering kernel evolution. Stochastic collapse links gradient noise to simpler subnetworks (Chen et al., 2024); our deterministic boundary analysis concerns recruitment through cooperating weak edges. Normalized suppression does not imply finite-time weight deletion.

10 Conclusion

Conserved quantities constrain how learning distributes computation across a network. For positive scalar DAGs, we make these dynamics explicit as capacity flow and, under balance and equal depth, we show that growth can be separated from routing. This yields chain-equivalent stable routing states, a path-memory criterion for instability, and thresholds for recruiting dormant subnetworks. Building on rich/lazy theory, we derive a formula for richness. Together, these results connect graph structure to which computations survive and how new ones become accessible. The matrix and ReLU experiments motivate extending this analysis beyond the scalar setting.

REFERENCES

P.-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005. doi: 10.1137/040605266.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit Regularization in Deep Matrix Factorization. *arXiv preprint arXiv:1905.13655*, 2019. URL https://arxiv.org/abs/1905.13655.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1NvflqAdoom.

Shahar Azulay, Edward Moroshko, Mor Shpigel Nacson, Blake E. Woodworth, Nathan Srebro, Amir Globerson, and Daniel Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 468–477, 2021. URL https://proceedings.mlr.press/v139/azulay21a.html.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025. URL https://jmlr.org/papers/v26/24-1523.html.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pages 20105–20118, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7eeb9af3eb1f48e29c05e8dd3342b286-Abstract-Conference.html.

Ana Cannas da Silva. *Lectures on Symplectic Geometry*, volume 1764 of *Lecture Notes in Mathematics*. Springer, 2001. doi: 10.1007/978-3-540-45330-7.

Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. Stochastic Collapse: How Gradient Noise Attracts SGD Dynamics Towards Simpler Subnetworks. *Journal of Statistical Mechanics: Theory and Experiment*, 2024:104023, 2024. URL https://arxiv.org/abs/2306.04251.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Clémentine C. J. Dominé, Nicolas Anguita, Alexandra M. Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.14623.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

Iordan Ganev and Robin Walters. Quiver neural networks. *arXiv preprint arXiv:2207.12773*, 2022.

Thomas Gebhart, Udit Saxena, and Paul Schrater. A unified paths perspective for pruning at initialization. *arXiv preprint arXiv:2101.10552*, 2021. URL https://arxiv.org/abs/2101.10552.

Antoine Gonon, Nicolas Brisebarre, Elisa Riccietti, and Rémi Gribonval. A path-norm toolkit for modern networks: consequences, promises and challenges. *arXiv preprint arXiv:2310.01225*, 2024. URL https://arxiv.org/abs/2310.01225.

Vincent Grandjean. On the limit set at infinity of a gradient trajectory of a semialgebraic function. *Journal of Differential Equations*, 233(1):22–41, 2007. doi: 10.1016/j.jde.2006.10.009.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 9226–9259, 2022. URL https://proceedings.mlr.press/v162/huang22e.html.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Devon Jarvis, Sebastian Lee, Clémentine Carla Juliette Dominé, Andrew M. Saxe, and Stefano Sarao Mannelli. A theory of initialisation's impact on specialisation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2503.02526.

George Jeffreys and Siu-Cheong Lau. Kähler geometry of quiver varieties and machine learning. *arXiv preprint arXiv:2101.11487*, 2021. URL https://arxiv.org/abs/2101.11487.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: Exact solutions reveal how unbalanced initializations promote rapid feature learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/94074dd5a072d28ff75a76dabed43767-Abstract-Conference.html.

Etai Littwin and Lior Wolf. The Loss Surface of Residual Networks: Ensembles and the Role of Batch Normalization. *arXiv preprint arXiv:1611.02525*, 2016. URL https://arxiv.org/abs/1611.02525.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. *arXiv preprint arXiv:2307.00144*, 2023.

Sibylle Marcotte, Gabriel Peyré, and Rémi Gribonval. Intrinsic training dynamics of deep neural networks. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.07370.

Hancheng Min, Enrique Mallada, and René Vidal. Early neuron alignment in two-layer ReLU networks with small initialization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QibPzdVrRu.

Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Marco Nurisso, Pierrick Leroy, Giovanni Petri, and Francesco Vaccarino. Topology and geometry of the learning space of relu networks: Connectivity and singularities. *arXiv preprint arXiv:2602.00693*, 2026.

Andrew Saxe, Shagun Sodhani, and Sam Jay Lewallen. The neural race reduction: Dynamics of abstraction in gated networks. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19287–19309, 2022.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1701.06538.

Jianghong Shi, Eric Shea-Brown, and Michael A. Buice. Learning dynamics of deep linear networks with multiple pathways. In *Advances in Neural Information Processing Systems*, volume 35, pages 34064–34076, 2022.

Hidenori Tanaka, Daniel Kunin, Daniel L. K. Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Andreas Veit, Michael J. Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparameterized models. In *Proceedings of Thirty Third Conference on Learning Theory*, pages 3635–3673, 2020.

Bo Zhao, Iordan Ganev, Robin Walters, Rose Yu, and Nima Dehmamy. Symmetries, flat minima, and the conserved quantities of gradient flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9ZpciCOunFb.

AI DISCLOSURE

We used OpenAI's ChatGPT and Codex to discuss and evaluate theoretical ideas and hypotheses, which included formulating mathematical claims, as well as developing proof arguments. We also used these tools to help design numerical experiments, implement simulation and analysis code, generate synthetic experimental data through code, process numerical outputs, and interpret results.

Additional uses included literature searches and synthesis, identifying relevant references, creating and revising scientific figures, drafting and editing manuscript text. Final use included using AI to shorten the paper, (in all seriousness, this did not go very well and the compressed writing was much worse than the authors expected).

AI-assisted contributions were developed through iterative discussion, review, and revision under the authors' direction. Checks included careful examination of mathematical arguments, checking of figure code, and verification of reference claims against source publications. The authors take responsibility for the final content of this work, including its mathematical claims, proofs, experimental results, text, code, and figures.

A SUPPLEMENTARY FIGURES

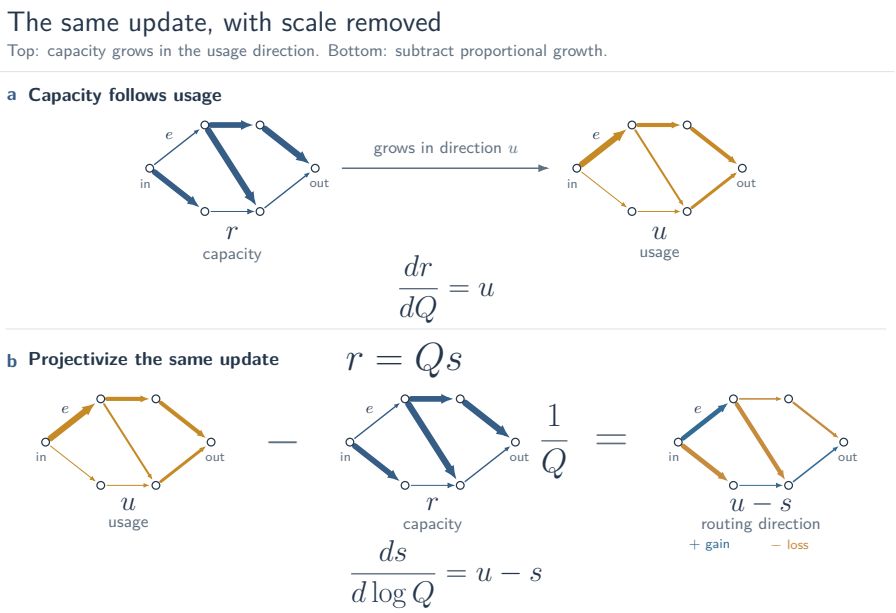

Figure 8: **Learning changes both network size and the allocation of capacity.** The top row depicts the capacity increment $dr = u\,dQ$. Writing $r = Qs$ separates proportional growth from redistribution: $Q\,ds = (u - s)\,dQ$. Red and blue mark gains and losses in normalized capacity share. This is a schematic of the exact scalar flow law.

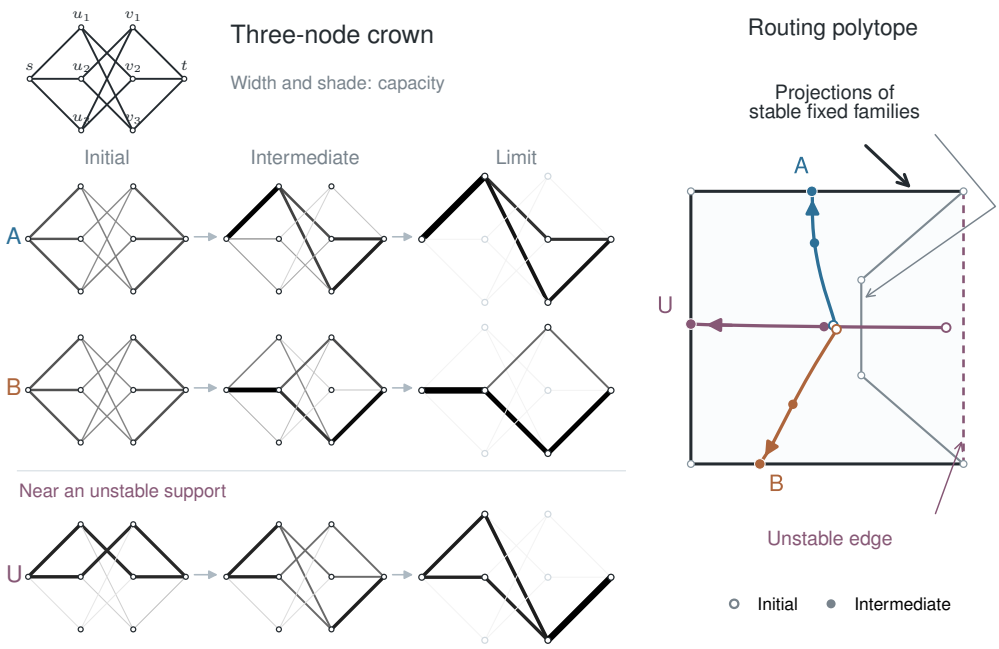

Figure 9: **Different routing histories can leave different chains.** The graph connects u_i to v_j for $i \neq j$, between a single source and sink. Two trajectories approach different layerwise-complete supports. Each becomes a chain after aggregating twins, although several scalar paths survive. The lower sequence passes near an unstable support. Edge width and shading encode normalized capacity. The routing polytope has dimension five, so the three-dimensional display is a projection, as labeled.

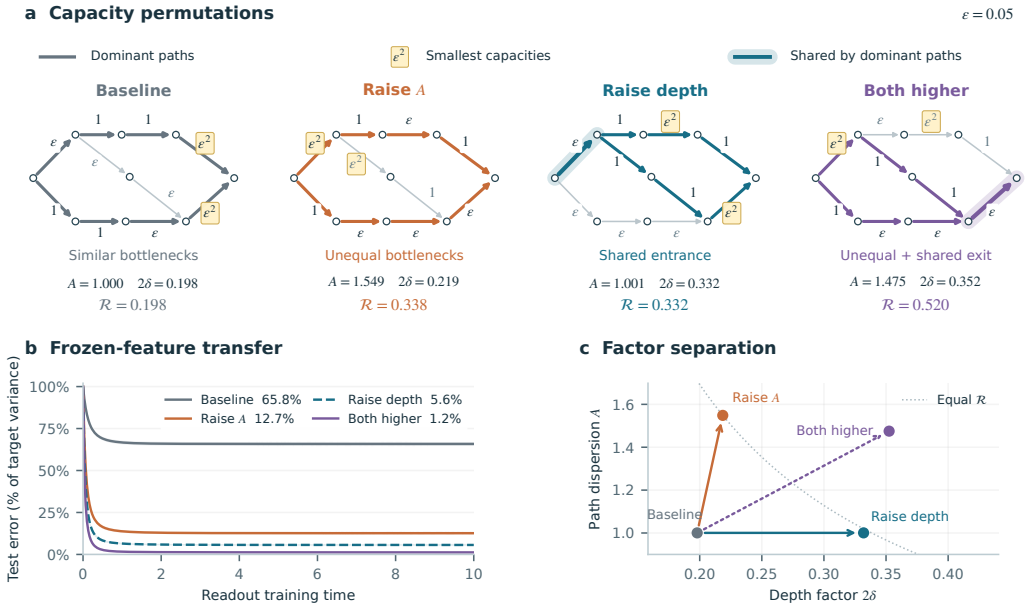

Figure 10: **Moving the same capacities changes how a network learns.** (a) Four permutations of small capacities, with $\varepsilon = 0.05$, change which bottlenecks dominant paths encounter. The figure's $A = 1 + \mathrm{CV}_\pi^2(\sigma)$ is its path-dispersion factor and $\delta = 1 - 1/m_{\mathrm{eff}}$ its depth factor, giving richness $\mathcal{R} = 2A\delta$. (b) Frozen-feature transfer compares readout learning across these constructions. (c) The two factors can vary separately in the illustrated examples; dashed curves have equal richness. These finite comparisons illustrate a possible learning consequence, not a general transfer guarantee.

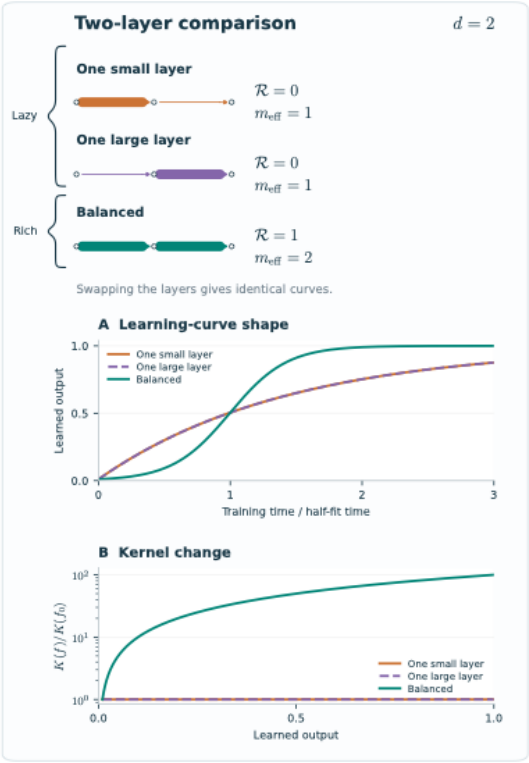

Figure 11: **The depth-two chain control.** Interchanging the two edge capacities preserves the scalar learning curve. Either strongly unequal allocation makes one edge dominate the conductance depth, approaching $m_{\text{eff}} = 1$, whereas balance gives $m_{\text{eff}} = 2$ and richness one. This inset accompanies the depth-six comparison in Figure 6.

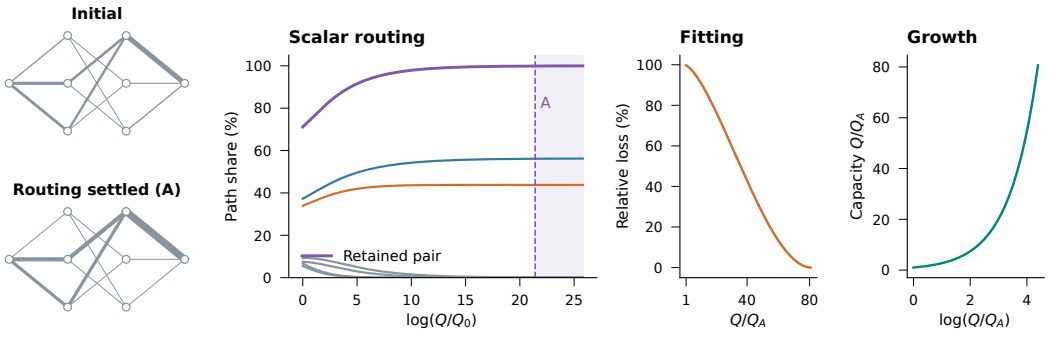

Figure 12: **A scalar counterpart of the two-stage ReLU example.** On the same crown graph as Figure 4, balanced positive scalar dynamics select two paths that merge into a chain under twin aggregation. A marks 99.9% of path weight on this pair. Capacity subsequently grows 80.6-fold while the path-share vector changes by at most 0.081 percentage points in total variation and relative squared loss falls to 10^{-4}. The graph and ordering of stages match the ReLU example; the initialization and target are chosen separately.

B CAPACITY COORDINATES AND THE SCALAR DIMENSION COUNT

Taking traces of the conserved matrix equations gives $Br = (\operatorname{tr} c_v)_v$, with $r_e = \|w_e\|_F^2$. Thus the matrix norm map lands in $\mathcal{P}_{\operatorname{tr} c}$, but this trace constraint alone does not assert that every capacity vector is realizable at a specified matrix charge. For nonnegative scalars the inverse map is simply $w_e = \sqrt{r_e}$.

Suppose the underlying undirected graph is connected and there is one source and one sink. To see that the hidden incidence matrix has full row rank, extend a vector in its left kernel by zero at the source and sink. The incidence relation makes this extension constant across every edge, hence constant on the connected graph, and therefore zero. Consequently $\operatorname{rank} B = |\mathcal{V}^\circ|$. A level containing a strictly positive state has dimension $|\mathcal{E}| - |\mathcal{V}^\circ|$. The scalar orthogonal gauge group is finite, so quotienting by it does not change dimension.

The source-cut functional is independent of the hidden constraints: the unit flow along any source-to-sink path lies in $\ker B$ and has source-cut value one. A fixed-cut slice meeting the positive orthant therefore has dimension $|\mathcal{E}| - |\mathcal{V}^\circ| - 1 = |\mathcal{E}| - |\mathcal{V}| + 1 = \beta_1$. It is bounded because a nonnegative flow on an acyclic graph cannot contain a circulation; fixed source inflow and fixed hidden imbalances bound every edge recursively in topological order. The formula concerns the full positive level or slice; a boundary face can have smaller dimension. With multiple boundary vertices, the unnormalized count remains $|\mathcal{E}| - |\mathcal{V}^\circ|$, but the single-source/sink formula in terms of $\beta_1 + 1$ no longer applies.

C CAPACITY GEOMETRY AND THE PATH EXPONENTIAL FAMILY

Two metrics in capacity coordinates. On the positive scalar sector, $r_e = w_e^2$ identifies the conserved level with the interior of $\mathcal{P}_c$. The edge basis gives capacity space the canonical flat metric $\sum_e dr_e^2$, restricted to $Br = c$. This is distinct from the Euclidean weight metric: since $dw_e = dr_e/(2\sqrt{r_e})$, its expression in capacity coordinates is

$$g_r(\xi, \zeta) = \frac{1}{4} \sum_e \frac{\xi_e \zeta_e}{r_e}, \qquad B\xi = B\zeta = 0.$$

It is a *Hessian metric*: in any affine coordinates on $Br = c$, its matrix is the Hessian of the restricted potential

$$\Psi(r) = \frac{1}{4} \sum_e r_e(\log r_e - 1).$$

Thus equal displacements in a drawn capacity polytope need not represent equal distances in weight space. The formulas hold on the positive interior; at a boundary, they apply within the relative interior of the active face.

For balanced networks of common depth m, write $r = Qs$ with $s \in \mathcal{S}_0$. Every layer carries unit flow, so $\sum_e s_e = m$ and $\sum_e ds_e = 0$. Substitution gives the orthogonal scale–routing decomposition

$$\sum_e dw_e^2 = \frac{m}{4Q}\, dQ^2 + Q\, g_s(ds, ds), \qquad g_s(\xi, \zeta) = \frac{1}{4} \sum_e \frac{\xi_e \zeta_e}{s_e}.$$

For a tangent vector ξ to $\mathcal{S}_0$, differentiating productivity gives

$$d \log \Phi(s)[\xi] = \frac{1}{2} \sum_e \frac{u_e(s) \xi_e}{s_e} = 2 g_s(u - s, \xi).$$

Since $u - s$ is tangent, the routing equation is $ds/d \log Q = \frac{1}{2} \operatorname{grad}_g \log \Phi$. The factor $1/2$ only fixes the clock. This identity explains the gradient structure and the strict Lyapunov formula in Section 4.

Paths as an exponential family. Let $\chi_P \in \{0,1\}^{\mathcal{E}}$ record the edges of a complete path P, and put $\theta_e = \log w_e = \frac{1}{2} \log r_e$. Then the positive scalar path sum and its normalized terms are

$$f(\theta) = \sum_{P \in \mathcal{P}} e^{\langle \theta, \chi_P \rangle}, \qquad A(\theta) = \log f(\theta), \qquad \pi_\theta(P) = e^{\langle \theta, \chi_P \rangle - A(\theta)}.$$

This is a finite exponential family, with edge incidence as its sufficient statistic and A as its log partition function. Differentiating the finite sum gives the standard moment identities (Wainwright and Jordan, 2008)

$$\nabla_\theta A = \mathbb{E}_{\pi_\theta}[\chi_P] = u, \qquad \nabla^2_\theta A = \mathrm{Cov}_{\pi_\theta}(\chi_P).$$

In particular, the response of edge usage to a log-weight perturbation is a path covariance. In the balanced equal-depth case, $\theta_e = \frac{1}{2}\log s_e$ gives $A = \log \Phi$, yielding the covariance operator used in the local stability analysis. The representation can be redundant: $\theta \mapsto \theta + B^T a$ leaves every path score unchanged because $B\chi_P = 0$, and any shift adding the same constant to every path score leaves π unchanged. The covariance is therefore positive semidefinite before restricting to identifiable directions. This statistical Hessian and the capacity metric above arise from different potentials; together they describe path sensitivity and the geometry of learning.

D Conservation, the moment map, and twin aggregation

Proof of Proposition 1. Hidden basis transformations act by $w_e \mapsto g_{v_e^{\mathrm{out}}} w_e g^{-1}_{v_e^{\mathrm{in}}}$, with $g_v = I$ at inputs and outputs. Along each path, adjacent factors cancel, so the predictor and any loss depending only on it are invariant. For symmetric hidden matrices A_v, extended by zero at the boundary, put

$$q_A(w) = \frac{1}{2}\sum_v \mathrm{tr}(A_v c_v(w)), \qquad (\xi_A(w))_e = w_e A_{v_e^{\mathrm{in}}} - A_{v_e^{\mathrm{out}}} w_e.$$

Differentiation gives $\nabla q_A = \xi_A$. This is the infinitesimal basis change with $g_v = \exp(-tA_v)$, so $dL[\xi_A] = 0$. Consequently

$$\frac{dq_A}{dt} = -\langle \nabla q_A, \nabla L\rangle = -dL[\xi_A] = 0.$$

The matrices A_v are arbitrary and each c_v is symmetric; hence every c_v is conserved. This argument uses full hidden basis transformations in linear networks. The corresponding symmetry of a ReLU network is restricted to transformations compatible with its activation, including positive rescalings.

Why the Noether interpretation needs Kähler geometry. The usual Hamiltonian Noether theorem concerns conservation along Hamiltonian flows, not arbitrary gradient flows (Cannas da Silva, 2001, Section 24.1). Here the connection is recovered through the Kähler structure and the complexified symmetry. Complexify the matrix spaces, use $J\delta w = i\delta w$ and $\omega(\xi,\eta) = \langle J\xi, \eta\rangle$, and replace transposes by adjoints in c_v. For Hermitian A_v, the infinitesimal unitary action is $\kappa_A = -J\xi_A$. Thus $dq_A = \iota_{\kappa_A}\omega$ and $\nabla q_A = J\kappa_A$: the q_A are components of the moment map in this convention. A loss depending on the complex network map is invariant under the complexified hidden basis action, including $J\kappa_A$, and therefore

$$\frac{dq_A}{dt} = -dL[J\kappa_A] = 0.$$

Restricting back to real weights gives the conservation law above. The essential extra hypothesis is invariance under the complexified action; unitary invariance alone would not establish conservation under gradient descent.

Twin aggregation. For vertices with identical predecessors, successors, and boundary role, concatenate their feature coordinates, stack incoming maps, and concatenate outgoing maps. This rearranges the same scalar parameters, is an isometry for the Frobenius metric, and preserves the network map under the corresponding identification of boundary coordinates. It therefore intertwines gradient flows. In Figure 1, merging the two outputs produces one two-dimensional output and preserves all three parameters.

E Capacity clock and scale-routing reduction

For positive scalar weights, $f = \sum_P z_P$, $z_P = \prod_{e\in P} w_e$, and $\pi(P) = z_P/f$. With $u_e = \sum_{P\ni e}\pi(P)$ and $r_e = w_e^2$, differentiation gives

$$\frac{\partial f}{\partial w_e} = \frac{f u_e}{w_e}, \qquad \dot r_e = -2\ell'(f) f u_e.$$

Every path entering a hidden vertex also leaves it, so $Bu = 0$ and Br is conserved. For an antichain cut C, which meets every complete path once, $\sum_{e\in C} u_e = 1$. Hence its capacity $Q_C = \sum_{e\in C} r_e$ obeys

$$\dot{Q}_C = -2\ell'(f)f, \qquad \frac{dr}{dQ_C} = u$$

where $\dot{Q}_C \neq 0$. Cut capacities differ by conserved signed sums of vertex charges, obtained by summing the incidence constraint over the difference of their upstream vertex sets. On a growth interval we use the source cut $Q = Q_{C_{\mathrm{in}}}$.

Writing $r = Qs$ gives $ds/d\log Q = u(Q, s) - s$ and $Bs = c/Q$. In balance, s belongs to the fixed unit-flow polytope. If every path has length m, then

$$z_P = Q^{m/2}\prod_{e\in P}\sqrt{s_e}, \qquad f = Q^{m/2}\Phi(s), \qquad \Phi(s) = \sum_P \prod_{e\in P}\sqrt{s_e}.$$

The common scale cancels from $\pi(P)$, so u depends only on s. Fixed points $u(s_\star) = s_\star$ yield capacity rays $r(Q) = Qs_\star$. Appendix G gives the productivity and normalized-norm identities.

F Local routing stability and the path-memory formula

Work on the relative interior of a balanced support of common path depth d, and assume a nonzero routing tangent space. All path expectations below use the positive distribution $\pi_s(P) = \Phi(s)^{-1}\prod_{e\in P}\sqrt{s_e}$.

The parallel-route example. For three disjoint balanced routes of common depth d, write s_i for the constant edge capacity along route i, normalized by $\sum_i s_i = 1$. Then $u_i = s_i^{d/2}/\sum_j s_j^{d/2}$ and $ds_i/d\log Q = u_i - s_i$. At $s_i = 1/3$, for $\sum_i \delta s_i = 0$,

$$\frac{d\,\delta s}{d\log Q} = \left(\frac{d}{2} - 1\right)\delta s.$$

For $d = 2$, $u = s$ everywhere. For $d \neq 2$, an interior fixed point requires $s_i^{d/2-1}$ to be independent of i, hence equal shares. The cases $d < 2$ and $d > 2$ respectively damp and amplify route differences.

Covariance linearization. Let $\chi_P \in \{0, 1\}^E$ be the edge-incidence vector of P, $\Sigma_\pi = \mathrm{Cov}_\pi(\chi_P)$, and $D_s = \mathrm{diag}(s)$. For an edge perturbation δs, differentiation of the normalized path weights gives

$$\delta\log\pi_s(P) = \frac{1}{2}\sum_e(\chi_P(e) - u_e)\frac{\delta s_e}{s_e}, \qquad \delta u = \frac{1}{2}\Sigma_\pi D_s^{-1}\delta s.$$

At $u(s_\star) = s_\star$, the routing Jacobian is therefore

$$J_\star = \frac{1}{2}\Sigma_\pi D_s^{-1} - I.$$

Here the tangent space consists of perturbations satisfying $B\delta s = 0$ and zero total flow through a cut. The covariance maps into this space because every path has the same boundary and crosses each cut once. The affine hull of the path-incidence vectors is the unit-flow space, so the covariance has full rank on its tangent space. Its nonzero spectrum is real and nonnegative: $\Sigma_\pi D_s^{-1}$ is similar to $D_s^{-1/2}\Sigma_\pi D_s^{-1/2}$. A mode of covariance eigenvalue λ evolves as $Q^{\lambda/2-1}$. Equality $\lambda = 2$ is a linear marginality statement; nonlinear stability requires further analysis.

Layers and the path-space operator. An equal-depth source-to-sink DAG has a unique vertex grading by distance from the source. Indeed, two source-to-vertex paths of different lengths, appended to the same vertex-to-sink path, would give complete paths of different lengths. Thus every complete path uses one edge $e_\ell(P)$ between each consecutive pair of layers.

Give edge observables the inner product $\langle h, g\rangle_s = \sum_e s_e h_e g_e$ and centered path observables the $L^2(\pi)$ inner product. Define

$$Th(P) = \sum_{e\in P} h_e - \mathbb{E}_\pi\left[\sum_{e\in P} h_e\right].$$

At a fixed point, $u = s$, and the adjoint acts as

$$(T^*F)_e = \mathbb{E}_\pi[F \mid P \ni e], \qquad T^*T = D_s^{-1}\Sigma_\pi.$$

The nonzero spectra of T^*T and TT^* coincide. Moreover,

$$\langle F, TT^*F\rangle_\pi = \sum_{\ell=1}^{d} \mathrm{Var}_\pi(\mathbb{E}_\pi[F \mid e_\ell(P)])\,.$$

The leading eigenfunction of TT^* lies in the image of T, so maximizing over centered additive observables $F = F_h - \mathbb{E}_\pi F_h$, where $F_h(P) = \sum_{e\in P} h_e$, yields the explained-variance formula in the main text. Each summand is an explained-variance fraction between zero and one. This gives $\lambda_{\max} \le d$, and disjoint parallel paths attain d by retaining route identity at every layer. A chain has no routing perturbation and is excluded from this maximization.

G Convergence, free energy, and stable routing states

Throughout this section the graph is a finite simple DAG with one source, one sink, and common path depth m. Every retained edge lies on a complete path. Routing states are nonnegative balanced unit flows. Stability means Lyapunov stability for the normalized dynamics induced by the smooth weight flow, allowing perturbations outside the active face. It does not mean attraction to an individual point of a continuous fixed family. The parallel-edge depth-one illustration in the main text is separate from this simple-graph classification.

Strict productivity identity. For positive s, $\partial_{s_e} \log \Phi = u_e/(2s_e)$ and $\sum_e u_e = \sum_e s_e = m$, giving

$$\frac{d\log\Phi}{d\log Q} = \frac{1}{2}\sum_e \frac{u_e(u_e - s_e)}{s_e} = \frac{1}{2}\sum_e \frac{(u_e - s_e)^2}{s_e} \ge 0.$$

Equality holds precisely at $u = s$. The metric and gradient identity are derived in Appendix C. Since $\|w\|^2 = mQ$ and $f = Q^{m/2}\Phi$, the function-normalized norm is $m\Phi^{-2/m}$ and decreases along routing.

A smooth description at the boundary. Put $x_e = \sqrt{s_e}$ and $\varphi(x) = \sum_P \prod_{e\in P} x_e$, so $\varphi(x) = \Phi(s)$. On the sphere $\|x\|^2 = m$, Euler's identity gives

$$\frac{dx_e}{d\log Q} = \frac{1}{2}\left(\frac{\partial_e\varphi(x)}{\varphi(x)} - x_e\right), \qquad \frac{dx}{d\log Q} = \frac{1}{2}\,\mathrm{grad}_{\mathbb{S}} \log\varphi(x).$$

This vector field is analytic wherever $\varphi > 0$, including zero coordinates. It preserves nonnegativity. Gauge invariance gives $B(x \odot \nabla\varphi) = 0$, where $\odot$ denotes coordinatewise multiplication, so $d(Bx^2)/d\log Q = -Bx^2$ and the balanced locus is invariant. This smooth extension, rather than an arbitrarily absorbing extension of the capacity equation, specifies boundary dynamics.

Proposition 2 (Convergence in the balanced equal-depth case)**.** *Every positive balanced routing trajectory converges as $Q \to \infty$ to one routing fixed point. The limit need not be stable.*

Proof. The normalized weights remain on a compact sphere and φ is nondecreasing, hence bounded below by its positive initial value. They therefore follow an analytic gradient flow in a neighborhood of a compact positive-output set. The analytic gradient inequality implies finite trajectory length and convergence to one critical point; the standard argument is reviewed in Absil et al. (2005, Section 2). Squaring the coordinates gives a limit $s_\star$, and the smooth stationary equation gives $u(s_\star) = s_\star$ on its active support. Exact symmetric initializations on deep parallel routes illustrate why convergence alone does not imply stability. □

Capacity and usage path laws. Write $m_\ell(v)$ for the vertex mass of a unit flow, with unit source and sink masses. Its capacity-based Markov path law q_s has transition probability $s_{uv}/m_{\ell-1}(u)$ along $u \to v$, so

$$q_s(P) = \frac{\prod_{e\in P} s_e}{\prod_{v\in P\cap\mathcal{V}^\circ} m(v)}.$$

It has edge marginals s. The usage law $\pi_s(P) = \prod_{e\in P}\sqrt{s_e}/\Phi(s)$ is also Markov: if $b(v)$ is the sum of path products from v to the sink, its transition along $u \to v$ is $\sqrt{s_{uv}}b(v)/b(u)$. Thus at a fixed point, where the edge marginals of both laws equal s, the laws coincide. Expectations below are taken over positive active paths.

Proposition 3 (Free-energy dissipation)**.** *For the vertex variables X_ℓ under q_s, define*

$$\mathcal{F}(s) = \frac{1}{2}\sum_{\ell=1}^{m} I(X_{\ell-1}; X_\ell) = \frac{1}{2}\sum_e s_e \log s_e - \sum_{v\in\mathcal{V}^\circ} m(v)\log m(v).$$

Along positive routing dynamics,

$$\frac{d\mathcal{F}}{d\log Q} = -D_{\mathrm{KL}}(\pi_s\|q_s) - D_{\mathrm{KL}}(q_s\|\pi_s) \le 0.$$

The derivative vanishes exactly at a fixed point. The identities apply in the positive interior of an invariant support; continuity of $a\log a$ extends $\mathcal{F}$ to its boundary.

Proof. The mutual-information identity follows from the joint edge law s_{uv} and its adjacent marginals. Direct substitution gives

$$\log\frac{q_s(P)}{\pi_s(P)} = \frac{1}{2}\sum_{e\in P}\log s_e - \sum_{v\in P\cap\mathcal{V}^\circ}\log m(v) + \log\Phi(s), \qquad D_{\mathrm{KL}}(q_s\|\pi_s) = \mathcal{F}(s) + \log\Phi(s).$$

Differentiate $\mathcal{F}$ along $u - s$. The constant derivatives cancel because each layer has unit mass, leaving the difference of the expectations of $\log(q_s/\pi_s)$ under π_s and q_s. This is the stated negative sum of divergences. Equality is equivalent to $q_s = \pi_s$, which implies $u = s$; the converse follows from the Markov observation above. □

Theorem 1 (Stable states are product flows on complete layers)**.** *A routing fixed point is stable if and only if its active support is complete between consecutive active layers and its capacities satisfy*

$$s_{uv} = m_{\ell-1}(u)m_\ell(v).$$

These states have $\mathcal{F} = 0$ and $\Phi = 1$. Every other fixed point has an expanding direction within its active face. Every stable state becomes a chain after twin aggregation.

Proof. If the adjacent vertex variables are independent, their joint probabilities give the displayed product capacities. Positive active marginals require every pair of consecutive active vertices to be joined. Conversely, for such a support and capacities, each path weight is the product of its internal vertex masses. Summing independently over layers gives $\Phi = 1$ and edge usage $u = s$.

Suppose now that a fixed point is not of this form. Since $\pi_s = q_s$ there, some consecutive internal vertices X_j, X_{j+1} are dependent, with $1 \le j \le m-2$. Choose a centered function $F = g(X_j)$ for which $\mathrm{Var}(\mathbb{E}[F \mid X_{j+1}]) > 0$. This is an additive path observable: assign $g(v)$ to each edge entering v in layer j and zero to other edges. Both e_j and e_{j+1} reveal X_j, while e_{j+2} reveals X_{j+1}. Therefore

$$\sum_{\ell=1}^{m}\mathrm{Var}(\mathbb{E}[F \mid e_\ell]) \ge 2\,\mathrm{Var}(F) + \mathrm{Var}(\mathbb{E}[F \mid X_{j+1}]) > 2\,\mathrm{Var}(F).$$

Appendix F consequently gives $\lambda_{\max} > 2$, so the fixed point has a positive Jacobian eigenvalue on its feasible active-face tangent space. It is unstable even before considering dormant edges.

It remains to justify stability of product states, including at the boundary. Place the normalized weights between layers in matrices W_ℓ. On the full sphere $\sum_\ell \|W_\ell\|_F^2 = m$, even allowing signed weights,

$$|\varphi(x)| = |W_m\cdots W_1| \le \prod_\ell \|W_\ell\|_2 \le \prod_\ell \|W_\ell\|_F \le 1.$$

Thus every product state is a global maximum of the analytic sphere function φ. To see pointwise Lyapunov stability rather than merely stability of the maximizing set, apply the local analytic gradient inequality to $E = 1 - \varphi \ge 0$: $\|\operatorname{grad}_{\mathbb{S}}\varphi\| \ge cE^{1-\theta}$ for some $c > 0$, $0 < \theta < 1$. Along the ascent flow, the arc length accumulated while staying in this neighborhood is at most $E(x_0)^\theta/(c\theta)$. Taking x_0 sufficiently close to the chosen maximum prevents exit from any prescribed smaller neighborhood. This proves stability, including balanced nonnegative perturbations across faces. Finally, within each active layer all vertices have identical neighbors; twin aggregation gives a chain. □

Minimum norm and continuous endpoint families. For arbitrary weights on this depth-m graph, the same norm inequality gives $|f(w)| \le (\|w\|^2/m)^{m/2}$. Hence $\|w\|^2/|f|^{2/m} \ge m$, with equality at every product routing state. Stable states therefore globally minimize the normalized norm, not just its restriction to one face. On a support with active internal layer sizes $n_1, \dots, n_{m-1}$, their vertex masses vary over a product of open simplices of dimension $\sum_{\ell=1}^{m-1}(n_\ell - 1)$. These directions are neutral, so stability does not require convergence of neighboring trajectories to the same member of the family.

H MIXED DEPTH, CONSERVED CHARGE, AND ASYMPTOTIC SELECTION

Finite-scale identities. For arbitrary path depths and finite charge, Appendix E gives $ds/d\log Q = u(Q,s) - s$ and $Bs = c/Q$. Summing $Br = c$ over an upstream vertex set shows that capacities of separating cuts differ by a conserved signed sum of charges. Thus their scales differ by constants and become asymptotically equivalent. At nonzero charge, $u = s$ is not a full fixed-point equation on a finite-Q slice, since $Bu = 0$ but $Bs = c/Q$.

Write $d_P = |P|$, $\psi_P(s) = \prod_{e\in P}\sqrt{s_e}$ and $z_P = Q^{d_P/2}\psi_P(s)$. For any path observable a_P independent of Q, differentiation of normalized path weights gives

$$\frac{d\mathbb{E}_\pi[a]}{d\log Q} = \mathrm{Cov}_\pi\left(a, \frac{d\log z_P}{d\log Q}\right), \qquad \frac{d\log z_P}{d\log Q} = \frac{d_P}{2} + \frac{1}{2}\sum_{e\in P}\frac{1}{s_e}\frac{ds_e}{d\log Q}.$$

Taking $a_P = d_P$ proves the main-text depth-race formula. At fixed s, only the nonnegative variance term remains; the covariance accounts for reweighting caused by routing. In a charged network the fixed-s derivative is an algebraic decomposition, not a motion within the moving feasible slice.

Theorem 2 (General positive-scalar routing convergence)**.** *Let the graph be a finite two-terminal DAG, every edge lying on a complete path, and let all initial scalar weights be positive. Consider the positive gradient-ascent continuation of $f(w) = \sum_P \prod_{e\in P} w_e$, or a loss-gradient trajectory following that curve with $Q \to \infty$. Initial charge may be arbitrary and paths may have different depths. Then $s(Q) \to s_\star$, a balanced unit flow. Its positive edge support H has a common path depth d_H, and*

$$\pi_P(Q) \longrightarrow \pi_P^* = \begin{cases} \dfrac{\prod_{e\in P}\sqrt{(s_\star)_e}}{\Phi_H(s_\star)}, & P \subseteq H, \\ 0, & P \not\subseteq H, \end{cases} \qquad u(Q) \longrightarrow s_\star.$$

Here $\Phi_H(s) = \sum_{P\subseteq H}\prod_{e\in P}\sqrt{s_e}$. Moreover,

$$D_{\mathrm{eff}}(Q) \to d_H, \qquad f(Q) = Q^{d_H/2}\big(\Phi_H(s_\star) + o(1)\big), \qquad u_H(s_\star) = s_\star.$$

Thus the endpoint is a fixed point of balanced homogeneous routing on its surviving support. Convergence does not assert stability of that fixed point.

Proof. Escape and a limiting weight direction. Use an auxiliary ascent time t_+, so $dw/dt_+ = \nabla f$. Every derivative $\partial_e f$ is positive in the positive orthant and nondecreasing as the weights increase. A maximal solution cannot stay bounded: at finite maximal time boundedness would permit continuation, while on an infinite interval $\partial_e f(w) \ge \partial_e f(w(0)) > 0$. Monotonicity consequently gives $\|w\| \to \infty$. The identities

$$\frac{dr_e}{dt_+} = 2fu_e, \qquad \frac{dQ}{dt_+} = 2f, \qquad 0 \le \frac{dr_e}{dQ} = u_e \le 1$$

then force $Q \to \infty$.

The secant-limit theorem for gradients of C^2 semialgebraic functions (Grandjean, 2007) applies to the polynomial f: an escaping gradient trajectory has a limiting direction. Hence, with $R = \|w\|$,

$$w/R \to v \ge 0, \qquad \|v\| = 1.$$

Equivalently one may apply the theorem to $-1/f$ on the positive orthant, whose gradient is $f^{-2}\nabla f$ and whose value tends to zero. Set $q_e = v_e^2$. Dividing conservation by R^2 gives $Bq = 0$. A nonzero nonnegative balanced flow on this DAG has positive source flux $\beta = \sum_{e\in C_{\mathrm{in}}} q_e$: otherwise topological induction forces every edge to be zero. Thus $Q/R^2 \to \beta > 0$ and $s \to s_\star = q/\beta$.

Vanishing edges cannot retain output share. Let $H = \{e : (s_\star)_e > 0\}$ and let d be the largest depth of a complete path contained in H. Such a path exists because $s_\star$ is a balanced unit flow. Split $f = Z_{\text{in}} + Z_{\text{out}}$ according to whether a path is contained in H. The already established capacity limit gives

$$Z_{\text{in}}(Q) \sim AQ^{d/2}, \qquad A = \sum_{\substack{P \subseteq H \\ |P|=d}} \prod_{e \in P} \sqrt{(s_\star)_e} > 0.$$

Meanwhile $R_{\text{out}} = \sum_{e \notin H} r_e = o(Q)$ and

$$\frac{dR_{\text{out}}}{dQ} = \sum_{e \notin H} u_e \geq \frac{Z_{\text{out}}}{Z_{\text{in}} + Z_{\text{out}}}.$$

If the latter share failed to tend to zero, it would exceed some fixed $0 < \varepsilon < 1$ at scales $Q_n \to \infty$, implying $Z_{\text{out}}(Q_n) \geq cQ_n^{d/2}$ for $c > 0$. Every path product is nondecreasing, so this lower bound holds throughout $[Q_n, 2Q_n]$. On that interval the asymptotic formula bounds $Z_{\text{in}} \leq CQ_n^{d/2}$. The outside share is therefore at least $c/(c+C) > 0$ throughout the interval. Integrating the last inequality contradicts $R_{\text{out}}(2Q_n) = o(Q_n)$. Hence outside paths have vanishing output share, even if their architectural depths exceed d.

Self-consistency and homogeneous support. Among paths inside H, those shorter than d lose relative weight by a power of Q; the depth-d probabilities converge to their positive coefficients normalized by A. Thus all path probabilities, and hence u, converge. Since

$$s_e(Q) = \frac{r_e(Q_0)}{Q} + \frac{1}{Q} \int_{Q_0}^{Q} u_e(q)\, dq,$$

the limiting usage equals $s_\star$. Every edge of H must consequently belong to a longest path in H.

The union of longest source-to-sink paths in a DAG is graded. At a common vertex, two longest paths cannot have different prefix lengths: the longer prefix combined with the other path's longer suffix would produce a still longer path. Such concatenation is a path because the graph is acyclic. Therefore each active vertex has a well-defined layer, each active edge advances one layer, and every path in H has length $d_H = d$. The stated limits and the homogeneous fixed-point equation follow. □

Stable endpoints and finite-target fitting. If this limiting homogeneous fixed point is stable, Theorem 1 makes it a layerwise-complete product state, hence a chain after twin aggregation. Exceptional trajectories can end at unstable fixed points; the theorem does not remove this distinction. A finite-target loss may stop capacity growth before the asymptotic regime. Charge vanishes after normalization but can affect which face is selected during the preceding trajectory.

A shallower route can be selected. Take two internally disjoint balanced chains of depths three and four, with all weights within each chain equal to a and b. For the ascent time above, $a' = a^2$ and $b' = b^3$. Starting from $a(0) = 1$, $b(0) = 1/2$ gives

$$a(t_+) = (1 - t_+)^{-1}, \qquad b(t_+) = (4 - 2t_+)^{-1/2}.$$

The depth-three chain diverges at $t_+ = 1$, while the depth-four chain remains bounded. Since $Q = a^2 + b^2$, its normalized limit is the depth-three chain. This positive, balanced example shows directly that d_H need not be the largest depth in the original graph.

I KERNEL EVOLUTION AND CONDUCTANCE DEPTH

These formulas hold for positive scalar DAGs, including mixed depths and charge, on a growth interval where the logarithmic derivatives are defined.

Set $g = \nabla_w f$, $\gamma_e = g_e/w_e = fu_e/r_e$, $\sigma_P = \sum_{e \in P} \gamma_e$, and $\mathcal{K} = \|g\|^2$. Then

$$\mathcal{K} = \sum_e \frac{(fu_e)^2}{r_e} = f\mathbb{E}_\pi[\sigma_P], \qquad \dot{f} = -\ell'(f)\mathcal{K}.$$

Each path monomial is multilinear in its distinct edge weights. Consequently, if $H_f = \nabla_w^2 f$,

$$g^T H_f g = f\mathbb{E}_\pi\left[\sigma_P^2 - \sum_{e\in P}\gamma_e^2\right].$$

Using $\nabla\mathcal{K} = 2H_f g$ cancels the loss clock and gives

$$\begin{aligned}\mathcal{R} &= \frac{f}{\mathcal{K}^2}\nabla\mathcal{K}\cdot g \\ &= 2\frac{\mathbb{E}_\pi[\sigma_P^2]}{\mathbb{E}_\pi[\sigma_P]^2}\left(1 - \frac{\mathbb{E}_\pi[\sum_{e\in P}\gamma_e^2]}{\mathbb{E}_\pi[\sigma_P^2]}\right) \\ &= 2(1+\mathrm{CV}_\pi^2(\sigma))(1-m_{\mathrm{eff}}^{-1}).\end{aligned}$$

For each path, $m_P^{\mathrm{series}} = \sigma_P^2/\sum_{e\in P}\gamma_e^2$ lies in $[1,|P|]$. The network depth m_{eff} is their harmonic average with weights proportional to $\pi(P)\sigma_P^2$. This explains both the participation interpretation and the distinction from usage-weighted path length.

On a chain $u_e = 1$ and path dispersion vanishes, yielding the inverse-capacity expression in (8). A depth-one network has $m_{\mathrm{eff}} = 1$ and constant kernel. Low instantaneous $\mathcal{R}$ alone does not bound kernel change over an arbitrarily large predictor excursion. Claims about ReLU representations or transfer must be tied separately to their experimental controls and are not consequences of this scalar identity.

J Boundary expansion and collective nucleation

Work in a balanced graph of common path depth m. Let $s_\star$ be fixed within its active support H, let $Z = \mathcal{E}\setminus H$, and write normalized active and dormant weights as $y = \sqrt{s_H}$ and $x = \sqrt{s_Z}$. Put $y_\star = \sqrt{(s_\star)_H}$ and $\phi_H = \Phi(s_\star) > 0$. The polynomial in normalized weights is $\varphi(y,x) = \sum_P \prod_{e\in P}\sqrt{s_e}$, where the square roots are the coordinates y, x.

Expansion by dormant degree. Let $k = \min_{P\not\subset H}|P\cap Z|$. At the frozen active background, the first term involving dormant variables is

$$\mathcal{H}_k(x) = \sum_{\substack{P\not\subset H\\ |P\cap Z|=k}}\left(\prod_{a\in P\cap H}(y_\star)_a\right)\left(\prod_{e\in P\cap Z}x_e\right).$$

The smooth normalized weight equation from Appendix G gives, in the dormant coordinates,

$$2\frac{dx}{d\log Q} = \frac{\nabla_x\varphi(y,x)}{\varphi(y,x)} - x.$$

For $k = 1$, at least one dormant coordinate has a positive constant forcing at $(y_\star, 0)$. The active-face fixed point is therefore not stationary in the full smooth weight dynamics. For $k > 2$, the forcing is $O(\|x\|^{k-1})$, uniformly for y near $y_\star$, and

$$\frac{d\|x\|^2}{d\log Q} = -\|x\|^2 + O(\|x\|^k).$$

Sufficiently small feasible seeds thus decay in normalized coordinates while the background stays in this neighborhood. This is a local suppression result, not a claim that every sufficiently large seed recruits, nor that physical edge capacities decrease.

The quadratic threshold. When $k = 2$, $\mathcal{H}_2(x) = \frac{1}{2}x^T Ax$, with A defined in (9). Expansion at the fixed state gives

$$\frac{dx}{d\log Q} = \frac{1}{2}(A/\phi_H - I)x + O(\|x\|^2 + \|y-y_\star\|\,\|x\|).$$

The dormant block of the linearization is independent of active perturbations: every path using a dormant edge uses at least two, so the active–dormant mixed derivative vanishes there. An eigenvalue λ of A gives weight-amplitude rate $\frac{1}{2}(\lambda/\phi_H - 1)$. Thus $\lambda_{\max}(A) > \phi_H$ gives growth of an accessible leading mode, whereas equality is marginal. For connected coactivation graph, the Perron eigenvector a is strictly positive and determines the relative dormant weight amplitudes; the corresponding capacity proportions are a_e^2.

Feasibility of a dormant mode. This condition can be checked without treating arbitrary ambient directions as routing perturbations. A nonnegative amplitude vector a is feasible to leading order if an active correction b solves

$$B_H b = -B_Z a^2, \qquad \sum_{e \in C_{\mathrm{in}} \cap H} b_e = - \sum_{e \in C_{\mathrm{in}} \cap Z} a_e^2.$$

Then $s_Z = \varepsilon^2 a^2$, $s_H = (s_\star)_H + \varepsilon^2 b$ is an exactly balanced unit flow for sufficiently small ε. The active support is connected through the source and sink; its hidden incidence constraints and source-cut functional are independent. Hence the only compatibility conditions are $B_Z a^2 = 0$ at vertices absent from H.

In particular, every eigenvector $Aa = \lambda a$ with $\lambda \neq 0$ satisfies these missing-vertex balance conditions. Indeed, $\mathcal{H}_2$ is invariant under hidden scalar gauge changes at an inactive vertex, so

$$\sum_{e \in Z} B_{ve} x_e (Ax)_e = 0.$$

Substituting $x = a$ gives $\lambda \sum_e B_{ve} a_e^2 = 0$. The nonnegative Perron mode with $\lambda_{\max} > \phi_H > 0$ is therefore feasible. If A is reducible, this reasoning applies to a leading component; a seed must reach a supercritical component to exhibit its growth.

Why the threshold is not a transition between arbitrary stable states. The expansion describes initial growth near the specified resident support. If active directions are unstable, they can change the background before the seed becomes macroscopic. The equal-arm scalar experiment in Figure 7 uses an invariant symmetry family; the incumbent also has an unstable arm-imbalance direction. Its dormant-mode comparison remains valid without asserting full-state stability of the incumbent.

For completeness, a layerwise-complete product incumbent cannot have a strictly growing quadratic dormant mode in an equal-depth graph. Here $\phi_H = 1$ and prefix/suffix amplitudes are square roots of the corresponding vertex masses. Each new two-edge route passes through one inactive vertex, and its inlet–outlet coactivation block has entries $C_{uv} = \sqrt{m(u)m(v)}$. Therefore

$$\|C\|_2 = \sqrt{\left(\sum_{u \in I} m(u)\right)\left(\sum_{v \in O} m(v)\right)} \leq 1 = \phi_H.$$

Different inactive vertices give separate blocks. Full access to both adjacent layers gives equality, not strict instability. This also follows from the global productivity bound in Theorem 1.

K FIGURE DEFINITIONS AND EXPERIMENTAL SCOPE

Figure 1. The Y network uses scalar weights and squared loss with target $(1.5, 1)$. The curves are gradient-flow trajectories on three charge levels. Its hidden sign symmetry sends all three weights to their negatives. Under the outgoing-minus-incoming convention, $c = -3/4$ gives the two-sheeted hyperboloid and plane quotient, $c = 0$ the cone, and $c = +3/4$ the one-sheeted hyperboloid and Möbius quotient. The positive sector alone is identified with the capacity polyhedron; general signed states need not be determined by capacities.

Figure 2. Each phase portrait shows the full two-dimensional balanced routing polytope of the graph above it. Corner labels denote surviving routes; thick edges are fixed-point families. In the Wheatstone example, the purple trajectories begin with bridge allocations 2% and 85%.

Figure 5. In (a), frozen-scale vector fields explain the evolution of runs launched at five values of Q_0. In (b), the merge has one unit more incoming than outgoing capacity, so its charge is $c_b = -1$ in our convention. In the rightmost panel, each point specifies charges (c_b, c_m) and its color identifies the surviving pair of routes for the purple initialization. The source figure uses incoming-minus-outgoing charges; reversing both plot axes converts them to our convention without changing any outcome. The purple arrow follows $c_m = 0$ from $c_b = 0$ to $c_b = -1$, matching the illustrated charged trajectory.

Figure 3. In the ReLU experiments, every hidden vertex has width eight, every architecture edge is a trainable matrix, and the target is $\mathrm{ReLU}((x_1 + 0.5x_2)/\sqrt{1.25})$. For interface ℓ with E_ℓ edges, put $p_e = \|W_e\|_F^2 / \sum_{j\in\ell} \|W_j\|_F^2$ and $S_\ell = (E_\ell \sum_{e\in\ell} p_e^2)^{-1}$. Soft support S averages this effective fraction of occupied edges over ten interfaces. The clock is $\log(Q/Q_0)$, with $Q = \sum_e \|W_e\|_F^2/(8{\cdot}10)$ and Q_0 its initial value. Faint curves show six paired seeds; solid support curves and bands show their mean and sample standard deviation. Open circles mark 5% relative training loss. The local response uses separate seed-zero runs with perturbations $\pm\varepsilon$, where $\varepsilon = 10^{-4}$. Writing $\delta s = (s^+ - s^-)/(2\varepsilon)$ and s^{ref} for the symmetric scalar reference, its amplitude is $A = [\sum_e (\delta s_e)^2/(10 s_e^{\mathrm{ref}})]^{1/2}$; the plot reports $\log(A/A_0)$, with A_0 its initial value. Graph memory is mutual information between states separated by a layer lag, measured on the fixed architecture. Dotted gate curves report the mean fraction differing from initialization. Scalar simplex arrows preserve direction and speed ordering but include a baseline length for visibility; their lengths are not proportional to speed. These conventions do not change the integrated trajectories.

Figure 4. A functional path share is $\|f_P(X)\|_F / \sum_R \|f_R(X)\|_F$, evaluated with the current full-network ReLU masks. Initial support allocations instead sum squared edge norms, counting each edge once; overlapping supports do not partition total capacity. The experiment uses width 32, 256 Gaussian training inputs, 1024 independent held-out inputs, and a rank-one ReLU teacher. A marks update 46,494. Capacity $\mathcal{C} = \sum_e \|W_e\|_F^2$ grows a further 80.2-fold while the functional path-share vector changes by at most 0.74 percentage points in total variation. The teacher is $aq\,\mathrm{ReLU}(q^\top x)$, with $q = \mathbf{1}/\sqrt{32}$ and $a = 10^8$. Initial total capacity is 0.36; training minimizes $\|f - Y\|_F^2/(2na)$ by full-batch gradient descent with step 10^{-3}. Seed 7 was selected from eight seeds at two target amplitudes. This is an illustrative run, not an ensemble guarantee; coarse routing stability permits changes in gates and matrices.

The large-target regime is exactly equivalent, by degree-three positive homogeneity, to a unit-amplitude target with weights $V = a^{-1/3}W$ and step size multiplied by $a^{1/3}$. Its initial total squared norm is then $0.36a^{-2/3} \simeq 1.67 \times 10^{-6}$. This makes small-initialization alignment a relevant comparison. Early-alignment analyses concern neuron directions (Boursier et al., 2022; Min et al., 2024; Boursier and Flammarion, 2025); silent alignment concerns the shape of the tangent kernel before its scale grows (Atanasov et al., 2022). Route shares, neuron directions, and kernel shape are different observables. Measuring them together would test whether the mechanisms coincide in this graph; routing stability alone does not establish it.

Figure 12. The graph has edges $s \to a_i$, $a_i \to b_j$ for $i \neq j$, and $b_j \to t$, with $i, j \in \{0, 1, 2\}$. In the path order $(i, j) = (0, 1), (0, 2), (1, 0), (1, 2), (2, 0), (2, 1)$, the initial balanced capacity vector is the path-incidence mixture with weights $(0.08, 0.10, 0.35, 0.08, 0.32, 0.07)$. We integrate $ds/d\log(Q/Q_0) = u(s) - s$, with $Q_0 = 1$. A is the first sampled point where the two paths through b_0 carry 99.9% of path weight, at $\log(Q/Q_0) = 21.41$. We continue to $Q/Q_A = 80.6404$ and choose the scalar target so that final output is 99% of the target. This large target exposes the second stage; it is not a matched-initialization control for the ReLU run. DOP853 and Radau integrations agree within 7.2×10^{-12} in routing coordinates, with balance residual below 4.8×10^{-15}.

Figure 6: scalar chains. The scalar plots combine exact chain identities with numerical gradient flow. Initial output is matched, while initial norms and kernels may differ. Time is normalized separately for each run by its half-fit time; the target output is one. The kernel is divided by its initial value.

Figure 6: ReLU layer changes. Both studies use bias-free, scalar-output ReLU networks trained by full-batch gradient descent. In the left ReLU plot, there are nine weight layers and eight width-44 hidden layers, with 160 unit-sphere inputs in $\mathbb{R}^{10}$ and a standardized six-unit ReLU teacher. Starting from He initialization (seeds 31 and 32), the rescaling $W_l \mapsto e^{-\theta(l-\bar{l})}W_l$, with $\theta = \operatorname{artanh}\varepsilon$ and $\bar{l}$ the mean layer index, preserves the initial function. The six points use $\varepsilon \in \{0, 0.1, 0.2, 0.35, 0.55, 0.75\}$, ordered from upper right to lower left. Let $K_l(x)$ be the contribution of layer l to the NTK diagonal and $\kappa_l(x) = K_l(x)/\sum_j K_j(x)$. The horizontal coordinate, *initial NTK depth*, is $\mathrm{median}_x[\sum_l \kappa_l(x)^2]^{-1}$. For hidden activation matrices $H_l^0, H_l^{\mathrm{end}}$, let $d_l = \max\{1 - \mathrm{CKA}_{\mathrm{lin}}(H_l^0, H_l^{\mathrm{end}}), 10^{-12}\}$, where linear centered kernel alignment is $\mathrm{CKA}_{\mathrm{lin}}(A, B) = \|A^\top B\|_F^2/(\|A^\top A\|_F \|B^\top B\|_F)$ after center-

ing columns. The vertical coordinate, *drift depth*, is $[\sum_l (d_l/\sum_j d_j)^2]^{-1}$. The trend concerns how changes are distributed across layers; these two empirical depths need not agree numerically.

Figure 6: ReLU output rescaling. The right plot uses depths 2, 4, and 6, width 40, 140 unit-sphere inputs in $\mathbb{R}^8$, a standardized five-unit ReLU teacher, and He seeds 21 and 22. Multiplying the output matrix by $\alpha \in \{6, 1, 1/6\}$ and every other matrix by $\alpha^{-1/(d-1)}$ preserves the initial function. "Unscaled" denotes $\alpha = 1$, without imposing exact matrix balance. The ordinate is $\Delta \log \bar{K} = \log(\bar{K}_{\mathrm{end}}/\bar{K}_0)$, where $\bar{K}$ averages the NTK diagonal over inputs.

The left study uses 25,000 updates with learning rate $0.15/\max_x K_0(x)$; the right uses 12,000 updates with $0.2/\max_x K_0(x)$. Every displayed run exhausts this update budget before the code's capacity-growth stopping threshold (1.25 and 1.3, respectively). Thus these are fixed-budget comparisons with initialization-dependent learning rates, not matched-capacity comparisons. Actual capacity ratios range from 1.00000034 to 1.00890 in the left study and from 1.00029 to 1.17527 in the right. Points and bars average two seeds; whiskers show their range, not confidence intervals. Independent reruns reproduce the archived left means within 3.3×10^{-14} and the rounded right means within 4.9×10^{-5}.

Figure 7: matrix learning. Small seeds are added to four dormant matrix edges at $t = 0$, followed by gradient descent. Structured initialization and a scalar rank-one teacher give two recruitment stages. Their representative times, $t = 1.08$ and 13.24, mark maxima of loss descent, not discrete support changes. The dashed control replaces the shared A_2 by an independently trained copy, initialized identically.

Figure 7: equal-depth scalar comparison. Each incumbent arm has paths $s \to a_i \to b_i \to t$ and $s \to a_i \to c_i \to t$, for $i = 1, 2$. Each arm carries flow $1/2$, with fraction $3/4$ through b_i. The private dormant routes are $s \to m_i \to b_i$; merging the junctions and their inlet copies gives $s \to m$ and $m \to b_1, b_2$. Every complete path has depth three, so $ds/d\log(Q/Q_0) = u(s) - s$ is the exact routing reduction during capacity growth. A finite target can stop this growth before the displayed routing trajectory ends.

The private inlet copies and every outlet receive capacity $\varepsilon = 5 \times 10^{-5}$; the shared inlet receives 2ε, preserving total seed $\eta = 10^{-4}$ and matching all initial incumbent capacities. Bypass allocation is $y = s_{L_1} + s_{L_2}$ privately and $y = s_L$ with sharing. Coactivation vertices represent dormant edges, and

$$\lambda_{\mathrm{private}} = \sqrt{3/8} < \phi_H = 1/\sqrt{2} < \lambda_{\mathrm{shared}} = \sqrt{3}/2.$$

Every proper subset of the three shared seed edges is strictly subcritical. The equal-arm trajectory approaches $y = 1$ and a stable product family with a neutral split between output branches. The recovered package checks the exact weight-to-routing reduction at three scales and agreement between DOP853 and Radau integration.