Gradient Descent with Large Step Size Restores Symmetry in Deep Linear Networks with Multi-Pathway

Hee-Sung Kim ▪ Sungyoon Lee
Department of Computer Science, Hanyang University, Seoul, Korea | ICML 2026

ICML International Conference On Machine Learning

Research Question

Does Gradient Flow's "winner-take-all" symmetry breaking survive under *large-step* Gradient Descent?

Keywords: deep linear nets · multi-pathway · edge of stability · flat minima · implicit bias · re-balancing

Multi-Pathway Deep Linear Net

H parallel pathways; pathway h has depth L_h, weights $W_{h\ell} \in \mathbb{R}^{d\times d}$:

$$\Omega_h := W_{hL_h}\cdots W_{h1}, \qquad M := \sum_{h=1}^{H}\Omega_h.$$

Fit a target $M_\star$ with GD at step size η:

$$\mathcal{L} = \tfrac{1}{2}\,\|M - M_\star\|_F^2, \quad W_{h\ell}^{t+1} = W_{h\ell}^{t} - \eta\,\nabla_{W_{h\ell}}\mathcal{L}.$$

Two opposing forces:

- **GF bias** → *one* pathway (sparse, winner-take-all).
- **Large-step GD bias** → re-balancing toward flatter minima.

Mode Decoupling & What We Track

Aligning each pathway with the target SVD makes every pathway map diagonal, so dynamics **decouple across modes** i:

$$\sigma_{hi} := u_{\star i}^\top \Omega_h\, v_{\star i}, \quad \mathcal{L}_i = \tfrac{1}{2}\Big(\textstyle\sum_h \sigma_{hi} - \sigma_{\star i}\Big)^2.$$

σ_{hi} is the singular value mode i carries in pathway h ($\sigma_i = \sum_h \sigma_{hi}$). Depth-balanced init $W_{h\ell}(0) = \alpha_h^{1/L_h} I_d$ keeps GD on this manifold; tiny asymmetric scales α_h seed the symmetry breaking.

We track the per-pathway split σ_{h1} (centre figure; diagonal $\sigma_{11}{+}\sigma_{21}{=}\sigma_{\star 1}$ = minima manifold) and the **sharpness** $\lambda_{\max}$: When the single-path solution becomes GD-unstable $\lambda_{\max} > 2/\eta$, EoS oscillations can trigger re-balancing.

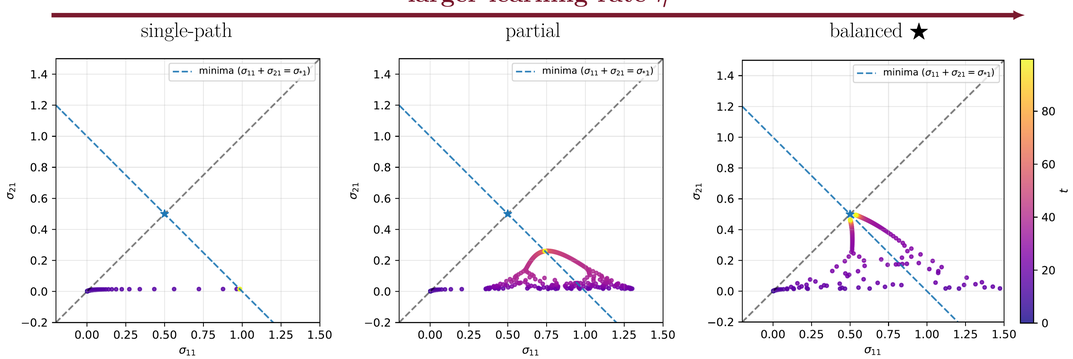
Contributions

- **Sharpness reduction.** Sharing a feature across H pathways cuts sharpness by $H^{2/L-1}$ — *sparse = sharp, balanced = flat.*
- **Re-balancing at the Edge of Stability.** Large-step GD breaks GF's winner-take-all bias in *two phases.*
- **Stability outweighs kinetics.** Even when depth structurally favors a shallow pathway, GD's flat-minima bias *outweighs* this kinetic bias.

Large-step GD disperses the signal across pathways

(1) Early training mimics *Gradient Flow*: the signal collapses into a single pathway;
(2) Oscillations at the Edge of Stability **re-balance** it into a flat, shared minimum.

Re-balancing in action



Trajectories of $(\sigma_{11}, \sigma_{21})$ colored by time t. As η grows, the endpoint drifts from the single-path corner $(\sigma_{\star 1}, 0)$ to the balanced minimum $(\frac{\sigma_{\star 1}}{2}, \frac{\sigma_{\star 1}}{2})$ — exactly the two phases above.

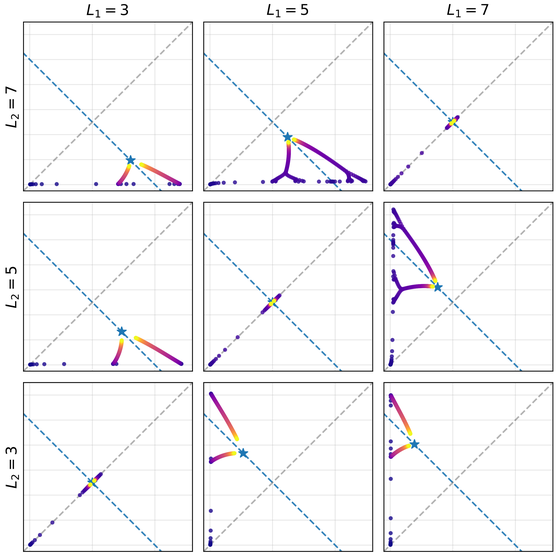
Why balanced = flat: Sharpness Reduction via Parallelism

Mode-i sharpness $\lambda_i = L\,\Sigma_h\,\sigma_{hi}^{2-2/L}$ is minimized by the *equal split* $\sigma_{hi} = \sigma_{\star i}/H$ (for $L > 2$):

$$\lambda_i^{\min} = L\,H^{\frac{2}{L}-1}\sigma_{\star i}^{2-\frac{2}{L}}, \qquad \frac{\lambda_i^{\min}}{\lambda_i^{\min}(H{=}1)} = \boldsymbol{H^{\frac{2}{L}-1}} < 1.$$

*Sparse single-path = **sharpest**; balanced = **flattest**. Equal sharing is the flattest allocation for fixed $L > 2$. The relative sharpness benefit of parallelism grows with depth.*

Kinetic vs. Stability Bias (heterogeneous depths)



$(\sigma_{11}, \sigma_{21})$ at $\eta = 2/\lambda_1^{\min}$, $(L_1, L_2) \in \{3, 5, 7\}^2$.
Despite structural asymmetry, GD approaches the sharpness-optimal distributed allocation (★).

Heterogeneous depths yield $\lambda_i = \Sigma_h\, L_h \sigma_{hi}^{2-2/L_h}$; concentrating all signal in one path remains **sharper than the optimal**. Two biases *conflict*:

- **Kinetic (GF):** the **shallow** path learns faster ⇒ winner-take-all.
- **Stability (GD):** a shallow path is **sharper** ⇒ mass moves to deeper paths.

Large-step GD overrides the depth bias.

Take-Home Message: Beyond Gradient Flow

Gradient Flow explains the initial drift toward pathway specialization, but finite-step GD can select a **different final minimum** by favoring lower-curvature, more distributed representations.
Therefore, pathway-level architectural biases inferred from GF need not persist under practical discrete optimization. **GF-based conclusions should be re-examined when finite learning rates and EoS dynamics matter.**