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

Hee-Sung Kim Sungyoon Lee

ICML 2026

Department of Computer Science
Hanyang University, Seoul, Korea

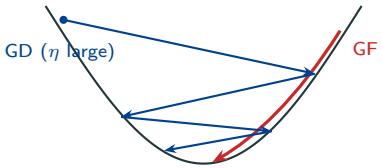
Continuous Theory, Discrete Reality

Deep Linear Networks (DLNs) are the **clean testbed** for the implicit bias of optimization: tractable enough to prove *exactly* which minimum training selects.

Almost all DLN theory uses **Gradient Flow** — the $\eta \to 0$ continuous limit.

But practical training is **discrete GD with large** η, living at the **Edge of Stability**, where sharper minima are actively avoided.

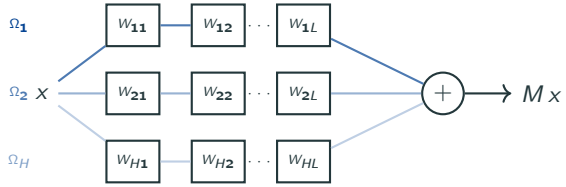
To use **DLNs** as a theory of optimization and implicit bias, we must account for the implicit bias of **discrete GD**.



GF slides into the nearest minimum;

large-step GD oscillates and **seeks flat minima**.

Does GF Predict What GD Does in Multi-pathway DLNs?

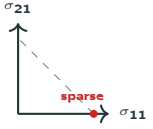


$M = \sum_{h=1}^{H} \Omega_h, \quad \text{minimize } \frac{1}{2}\|M - M_\star\|_F^2$

For the dominant singular mode, σ_{h1} denotes the contribution of pathway h.

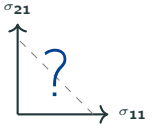
Two predictions, one question:

Gradient Flow (Shi et al. 2022)



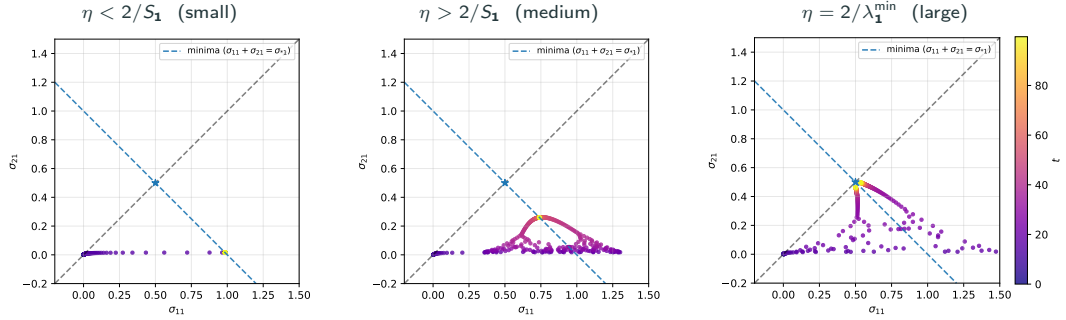
"Winner-takes-all": one path wins, others stay near zero

Discrete GD, large η



What does the Edge of Stability do to symmetry breaking?

The Answer: GD Breaks, Then Restores Symmetry



Trajectory of $(\sigma_{11}, \sigma_{21})$. Blue dashed: global minima $\sigma_{11} + \sigma_{21} = \sigma_{*1}$.

Small η reproduces GF's sparse corner. Past the stability threshold, large-step GD **abandons it** and converges to the balanced, symmetric solution.

Why? Spreading a Feature Across Paths Flattens the Minimum

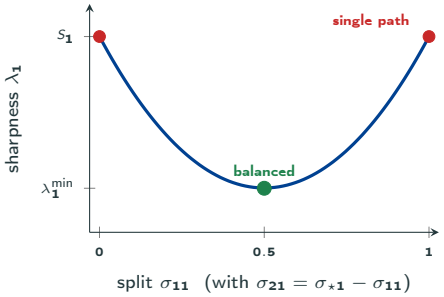
On the depth-balanced manifold, the mode-1 sharpness depends only on **how the target is split** across paths:

$$\lambda_1 \;=\; L\sum_{h=1}^{H}\sigma_{h1}^{2-2/L}, \qquad \textstyle\sum_h \sigma_{h1} = \sigma_{\star 1}.$$

Theorem (Sharpness reduction via parallelism).
λ_1 is minimized by the equal split $\sigma_{h1} = \sigma_{\star 1}/H$, giving

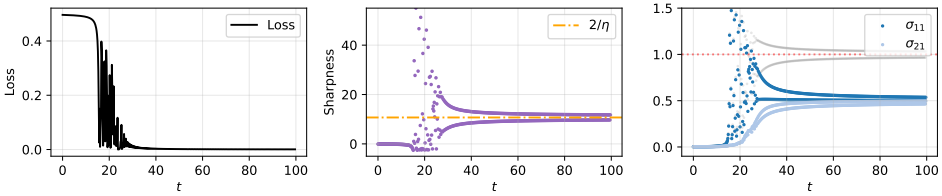
$$\frac{\lambda_1^{\min}}{\lambda_1(\text{single path})} \;=\; H^{\frac{2}{L}-1} \;<\; 1$$

- single path $\Rightarrow$ **sharpest** minimum
- more paths H *or* more depth L $\Rightarrow$ **flatter**



H=2, L=20: flat minimum is $\sim$2$\times$ less sharp.

How? Two Phases and the Edge of Stability



Under large η, training has two phases:

P1. Singular values small, sharpness low $\Rightarrow$ dynamics track GF: the dominant path grows and **symmetry breaks**.

P2. Its sharpness rises past $2/\eta \Rightarrow$ **EoS oscillations** drift mass into the other paths.

Oscillations are **not noise** — they are a *directed drift* toward flat minima.

Kinetic bias vs. stability bias.

Gradient Flow prediction

Architecture creates a built-in bias.

Shallower pathways learn faster
$\Rightarrow$ **winner-takes-all**

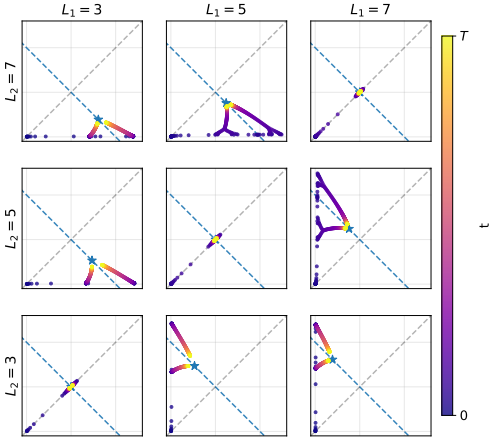
Large-step GD prediction

Stability creates a competing bias.

Sharp shallow-path solution becomes unstable
$\Rightarrow$ **re-balancing**

structural bias *vs.* flat-minima bias

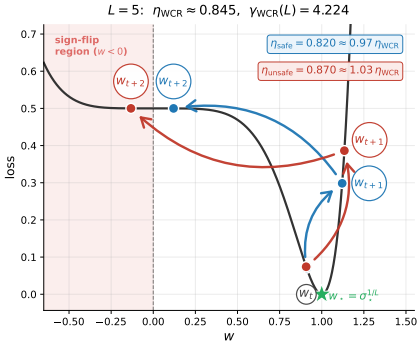
Heterogeneous depth: $L_1 \neq L_2$



Large-step GD counteracts the shallow-path advantage and moves toward the sharpness-balanced allocation.

Other Topics Discussed in the Paper

- **How large can η be?**
  A *Worst-Case Return* threshold, derived from a deep linear chain, caps η from above. The safe overshoot window **widens with depth**: $\gamma_{\mathrm{WCR}}(L) = \Theta(\log L)$.
- **Rank-p balancing: which modes re-balance?**
  A given η re-balances the sharpest modes and leaves the flatter tail GF-like — **larger η balances more modes**.
- **Nonlinear networks.**
  The same *break-then-rebalance* transition appears in Tanh MLPs with multi-pathway.



Deep-chain overshoot–return: crossing η_{WCR} sends the iterate into the sign-flip region ($w < 0$).

Takeaway: Revisit GF-based Architectural Bias under GD

What we showed

- Single-path solutions are sharp; balancing across H paths flattens sharpness by $H^{2/L-1}$ (Thm. 4.2).
- Large-step GD: symmetry breaking (GF-like) $\rightarrow$ re-balancing at the Edge of Stability $\rightarrow$ balanced minimum — overriding even depth-induced asymmetry.

Why it matters

- GF-predicted "winner-takes-all" / sparse-subnetwork biases (e.g., Lottery Ticket) may not survive discrete GD — GD tends to **spread** signal across paths.

Optimization bias can outweigh architectural bias:
large-step GD turns pathway specialization into pathway sharing.

Thank you! Questions?

muwonijr@hanyang.ac.kr