Global Convergence of Four-Layer Matrix Factorization under Random Initialization

Minrui Luo[1,3] Weihang Xu[2] Xiang Gao[4,3] Maryam Fazel[2] Simon Shaolei Du[2]

[1]Georgia Institute of Technology [2]University of Washington [3]Tsinghua University [4]University of Illinois Urbana-Champaign

1 Global convergence beyond NTK

Before this work, no global convergence guarantee had been established for general matrix factorization with $N > 2$ **under random initialization beyond the NTK regime**. NTK describes lazy training; here, **small initialization leads to feature learning**.

Four-layer problem

Let $W_j \in \mathbb{F}^{d\times d}$, $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, and $W = W_4W_3W_2W_1$. The full-rank target has identical singular values $s > 0$; rotating coordinates gives $\Sigma = sI$.

$$\mathcal{L} = \frac{1}{2}\|W - sI\|_F^2 + \frac{a}{4}\sum_{j=1}^{3}\|\Delta_j\|_F^2,$$
$$\Delta_j = W_jW_j^H - W_{j+1}^HW_{j+1}.$$

Main theorem (informal). Initialize all layer entries independently as $W_j(0)_{ab} \sim \epsilon\mathcal{N}(0,1)_{\mathbb{F}}$. For sufficiently small initialization and step size, and sufficiently strong balancing regularization,

$$\epsilon \le s^{1/4}/\operatorname{poly}(d, 1/\delta),$$
$$a \ge s\ \operatorname{poly}\left(d, 1/\delta, \log(s^{1/4}/\epsilon)\right),$$
$$\eta \le \frac{1}{s^{3/2}\operatorname{poly}(a/s, d, 1/\delta, s^{1/4}/\epsilon)},$$

GD reaches $\mathcal{L}(t) \le \xi$ for any $\xi > 0$ after

$$T_\xi = \frac{\operatorname{poly}(d, 1/\delta)}{\eta s\epsilon^2} + O\left(\frac{\log_+(ds^2/\xi)}{\eta s^{3/2}}\right),$$

with success probability at least

$$\boxed{1-\delta \quad (\mathbb{F}=\mathbb{C}), \qquad \frac{1-\delta}{2} \quad (\mathbb{F}=\mathbb{R}).}$$

H denotes conjugate transpose ($\top$ over $\mathbb{R}$); $\log_+(z) = \max\{0, \log z\}$.

2 Start with exact balance and GF

For **unregularized gradient flow**, $\dot{\Delta}_j = 0$. Small initialization already gives $\Delta_j(0) = O(\epsilon^2)$ at fixed dimension. Exact balance, $\Delta_j = 0$, isolates the leading dynamics.

$$W(t) = U(t)\Sigma_w(t)^N V(t)^H.$$

U, V are unitary/orthogonal and Σ_w is diagonal, nonnegative. The product dynamics **close in** (U, V, Σ_w). On this invariant manifold, the balancing penalty and its gradient vanish.

3 Two quantities control balanced GF

Skew-Hermitian error: general depth and spectrum

For every $N \ge 2$ and nonnegative diagonal target Σ,

$$E_\Sigma^2 = \|\Sigma^{1/2}(U - V)\Sigma_w\|_F^2, \qquad \boxed{\frac{d}{dt}E_\Sigma^2 \le 0.}$$

Balanced Gaussian initialization gives a nonzero, simple product spectrum almost surely. Small initialization keeps this error small; growing features align their left and right directions.

Hermitian main term: $N = 4$ and $\Sigma = sI$

Set

$$x_k = \sigma_k\left(\frac{(U+V)\Sigma_w}{2}\right), \quad E_0 = \|(U-V)\Sigma_w\|_F\big|_{t=0}, \quad R = \|\Sigma_w\|_{\mathrm{op}}.$$

Its growth rate satisfies

$$(2s - 2x_k^4 - \tfrac{3}{2}R^2E_0^2)x_k^4 - \tfrac{1}{8}R^2E_0^4x_k^2$$
$$\le \frac{d}{dt}x_k^2 \le s(2R^2 + E_0^2)x_k^2.$$

On favorable balanced Gaussian initialization with sufficiently small ϵ,

$$\boxed{\frac{d}{dt}x_k^2 \ge (\tfrac{5}{4}s - 2x_k^4)x_k^4.}$$

Positive modes grow to target scale and lower-bound layer singular values: $x_k \le \sigma_k(\Sigma_w)$.

4 Why real balanced GF can get trapped

Under real balanced Gaussian initialization,

$$\Pr(\det W(0) < 0) = \frac{1}{2}.$$

On this event, the relative orthogonal orientation has an eigenvalue -1, hence

$$\sigma_{\min}(U + V) = 0 \quad \Longrightarrow \quad x_{\min}(0) = 0.$$

The upper growth bound forces

$$\boxed{x_{\min}(t) \equiv 0, \qquad \mathcal{L}(t) \ge \frac{s^2}{2}.}$$

The last main-term mode never activates: a **saddle obstruction with probability** $1/2$. Complex initialization avoids this zero-mode event almost surely.

This half-probability obstruction concerns *exactly balanced GF*; it is not a failure guarantee for general random-initialized GD.

5 Random initialization and gradient descent

Independent Gaussian layers are only approximately balanced. The penalty $\mathcal{L}_{\mathrm{reg}} = \frac{a}{4}\sum_j \|\Delta_j\|_F^2$ controls imbalance and the errors introduced by discrete updates.

Use the counterparts of the balanced main and skew terms:

$$B_\pm = W_1 \pm W_2^{-1}W_3^HW_4^H.$$

The proof bounds $\sigma_{\min}(B_+)$ from below and $\|B_-\|_F$ from above, while keeping W_2^{-1} well-defined.

STAGE 1 Balance and alignment

Regularization reduces the balanced error, with a small discretization residual. The product changes little, $W(T_1) \approx W(0)$, and extreme layer singular values stay near their initialization scales.

With layer singular values bounded between $\mu_{\min}$ and $\mu_{\max}$,

$$\mathcal{L}_{\mathrm{reg}}(t+1) \le (1-\rho)\mathcal{L}_{\mathrm{reg}}(t) + O(\eta^2a^2),$$
$$\rho = \Omega(\eta a\mu_{\min}^4\mu_{\max}^{-2}).$$

At the end of this stage,

$$\boxed{\mathcal{L}_{\mathrm{reg}} \text{ small}, \qquad \sigma_{\min}(B_+)(T_1) > 0.}$$

The main term can oscillate during alignment; a positive bound is established at its endpoint.

STAGE 2 Feature learning and saddle avoidance

The main term grows while the skew error remains controlled:

$$\sigma_{\min}(B_+)(T_1 + T_2) \ge 2^{3/4}s^{1/4},$$
$$\|B_-\|_F = O(\epsilon\ \operatorname{poly}(d, 1/\delta)).$$

Layer singular values grow from initialization scale to target scale, with minimum singular value $\Omega(s^{1/4})$: **parameters change substantially**. The balanced-GF growth mechanism survives imbalance and discretization errors.

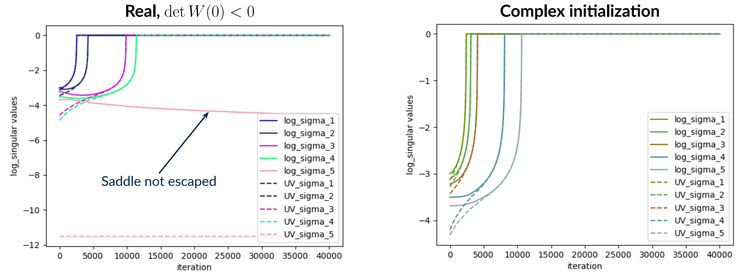
STAGE 3 Local convergence

At $T_* = T_1 + T_2$, every layer has a nonvanishing minimum singular value at target scale. Then

$$\boxed{\mathcal{L}(t) \le \mathcal{L}(T_*)e^{-\eta s^{3/2}(t-T_*)}.}$$

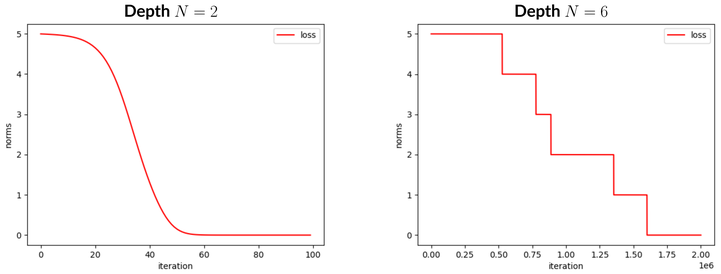
Balance ⟶ learn features ⟶ converge locally

K.1 A real mode fails to escape



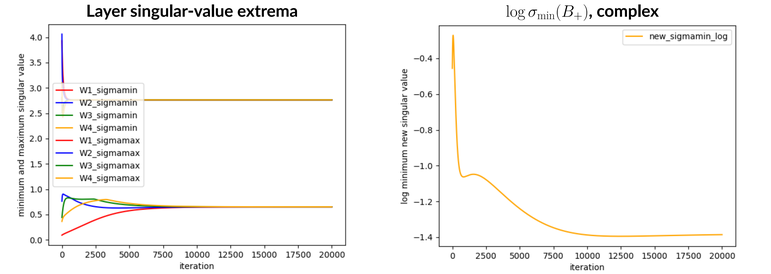
Balanced initialization, $d = 5$, $\Sigma = I$, $\epsilon = 0.05$, $\eta = 0.1$. Solid: $\log\sigma_k(\Sigma_w)$; dashed: $\log x_k$. The real zero mode is clipped at 10^{-5} for display; complex modes all grow.

K.2 Depth prolongs saddle plateaus



Complex balanced initialization, $\Sigma = I$, same d, ϵ, η as K.1. The source loss curves show longer plateaus and sharper transitions at greater depth; note the different iteration scales. Depths other than 4 are numerical illustrations beyond the theorem.

K.3 Regularization aligns the layers



Regularizer alone: reconstruction loss omitted; $a = 1$, $\epsilon = 1$, $\eta = 0.001$. The envelope across all layers contracts. The main term can oscillate before settling at a positive value, supporting the endpoint analysis in Stage 1.