MSIGN: AN OPTIMIZER PREVENTING TRAINING INSTABILITY IN LARGE LANGUAGE MODELS VIA STABLE RANK RESTORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training instability remains a critical challenge in large language model (LLM) pretraining, often manifesting as sudden gradient explosions that waste significant computational resources. We study training failures in a 5M-parameter NanoGPT model scaled via μP, identifying two key phenomena preceding collapse: (1) rapid decline in weight matrix stable rank (ratio of squared Frobenius norm to squared spectral norm), and (2) increasing alignment between adjacent layer Jacobians. We show theoretically how these observations create an amplifying Jacobian channel in a near-rank-one regime, and identify the additional signal-transmission conditions under which it yields exponential gradient growth. To break this instability mechanism, we propose MSign, a new optimizer that periodically applies matrix sign operations to restore stable rank. Experiments on the reported models from 5M to 3B parameters show that MSign prevents the observed collapse events, with measured computational overhead below 7.0% in these configurations.

1 INTRODUCTION

Modern deep learning optimizers and training techniques – including Adam (Kingma & Ba, 2015), layer normalization (Ba et al., 2016), and learning rate schedules – generally enable reliable training. However, as the model scale increases, pretraining large language models (LLMs) becomes increasingly fragile. Training failures manifest as sudden, unrecoverable gradient explosions and corresponding loss growth, which are difficult to predict and can waste substantial computational resources (Chowdhery et al., 2023; Zhang et al., 2022).

Empirical Investigation: Identifying the Failure Mechanism. We systematically study training failures using a 5M-parameter NanoGPT model derived from μP scaling (Yang et al., 2022), which provides a controlled environment for reproducible failure analysis. Through extensive monitoring, we identify two critical phenomena that consistently precede training collapse (Figure 1):

- **Observation 1: Stable Rank Collapse.** The stable rank of weight matrices – defined as $\mathrm{srank}(\mathbf{W}) = \|\mathbf{W}\|_F^2/\|\mathbf{W}\|_2^2$ – declines sharply in the steps preceding failure. This indicates that spectral energy becomes concentrated in the top singular directions, reducing the "effective dimensionality" of the weight matrices.
- **Observation 2: Jacobian Alignment Growth.** The alignment between adjacent layer Jacobians increases, meaning the top singular subspaces of consecutive layers become increasingly correlated. It prevents the typical cancellation effects in matrix products.

We show that stable rank near one controls the residual beyond each Jacobian's dominant rank-one component, while adjacent alignment of the dominant directions creates a constructive path across layers. The resulting lower bound makes the effects of alignment and stable-rank excess explicit.

Our Solution: The MSign Optimizer. To break the stable rank collapse condition, we propose **MSign**, an optimizer that periodically applies $\mathrm{sign}(\mathbf{W}) = \mathbf{U}\mathbf{V}^T$ to nonzero projection weights, using their reduced SVD. This operation maximizes stable rank at fixed positive rank by equalizing all nonzero singular values while preserving the column and row spaces. We restore the original

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.

Frobenius norm and leave an exactly zero weight unchanged. MSign is applied every P steps (typically $P = 100$), with computational overhead below 7.0%.

Experimental Validation. We validate MSign across four reported model configurations spanning 5M to 3B parameters: NanoGPT-5M [1], Sigma-40M (hybrid MHA/MLA attention) Qu et al. (2025); Hu et al. (2025), LLaMA-1B Kumar et al. (2025), and LLaMA-MoE-3B (mixture of experts). In these runs, baseline training fails via gradient explosion while MSign converges stably, preserving higher stable rank and bounded gradient norms. The layer ablation identifies attention projections as the effective intervention target in the tested configuration and provides layer-group intervention evidence consistent with the proposed Jacobian mechanism.

Contributions. We summarize our contributions:

1. **Mechanism identification and theoretical analysis**: We identify stable rank collapse and Jacobian alignment as consistent precursors to LLM training failures, derive a near-rank-one Jacobian-channel bound, and characterize the signal-coupling conditions that turn this channel into large parameter gradients.
2. **Practical solution**: We propose MSign, an optimizer that periodically restores weight stable rank and prevents the observed spectral collapse mode in our reported configurations.
3. **Extensive validation**: We demonstrate MSign's effectiveness across dense and MoE architectures at scales from 5M to 3B parameters, with measured overhead below 7.0%.

2 Literature Review

Training Instability and Low-Rank Dynamics. Training instability in large language models (LLMs) has been widely observed during pretraining, often requiring manual interventions or restarts due to sudden loss spikes and gradient explosions (Chowdhery et al., 2023; Zhang et al., 2022; Zeng et al., 2023). Various remedies have been studied, including gradient clipping (Pascanu et al., 2012), learning rate schedules (Kaplan et al., 2020), and proxy-model diagnosis (Wortsman et al., 2024). Concurrently, low-rank structure has been extensively documented in neural network parameters and gradients (Denil et al., 2013; Arora et al., 2019; Hu et al., 2022; Zhao et al., 2024; Huh et al., 2024). Our work connects low-rank gradient dynamics directly to LLM training instability through stable rank evolution.

Jacobian Analysis and Spectral Regularization. Jacobian spectral properties play a central role in neural network optimization and signal propagation (Glorot & Bengio, 2010; Saxe et al., 2013; Pennington et al., 2017; Yang & Schoenholz, 2017; Fort et al., 2019; Xiao et al., 2018). Stable rank (Rudelson & Vershynin, 2007; Vershynin, 2018) has been used for generalization bounds (Neyshabur et al., 2017), compressibility analysis (Li et al., 2018), and regularization (Sanyal et al., 2020). Our contribution is identifying stable-rank collapse and inter-layer Jacobian alignment as joint empirical precursors and deriving explicit conditions under which the associated near-rank-one channel amplifies parameter gradients.

3 Empirical Observations: Training Failure Phenomena

We identify reproducible training failures in transformer models under moderate learning rates. Our analysis reveals two consistent phenomena preceding training collapse.

3.1 Experimental Setup

We construct a reproducible failure scenario using the modified NanoGPT configuration and training hyperparameters reported in Appendix A.

From μP perspective Yang et al. (2022), this configuration corresponds to a 0.02 std and 6×10^{-4} learning rate at 100M scale, which is within typical ranges.

[1] https://github.com/karpathy/nanoGPT

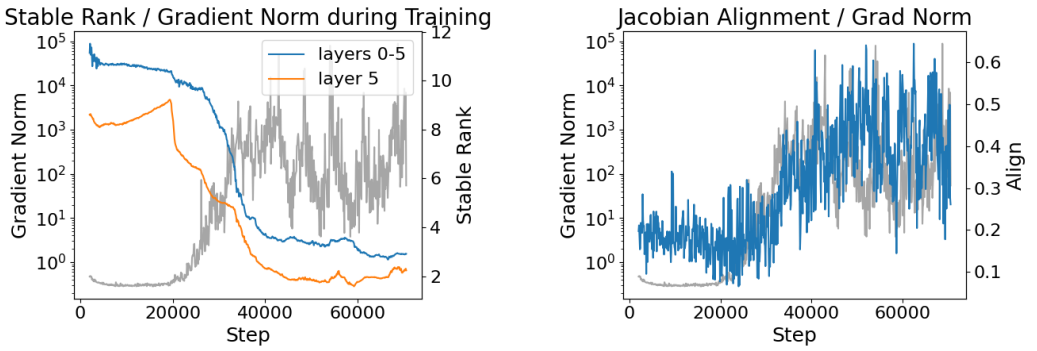

Figure 1: Correlation between training failure indicators and gradient norm explosion. **Left (Observation 1):** Stable rank (geometric mean across early layers) vs. gradient norm over training steps. As stable rank declines sharply around step 20000, gradient norms begin explosive growth. **Right (Observation 2):** Jacobian alignment (average between adjacent layers) vs. gradient norm. Increasing alignment precedes and accompanies gradient explosion.

Model and Jacobian Notation. We consider a standard decoder-only transformer with L stacked blocks. At each block ℓ, hidden states $\mathbf{H}^{(\ell-1)} \in \mathbb{R}^{T\times d}$ (sequence length T, hidden dimension d) are first processed by multi-head self-attention with query/key/value/output projections $\mathbf{W}_Q^{(\ell)}, \mathbf{W}_K^{(\ell)}, \mathbf{W}_V^{(\ell)}, \mathbf{W}_O^{(\ell)}$, and then by a position-wise MLP with weights $\mathbf{W}_1^{(\ell)} \in \mathbb{R}^{d\times d_{\text{ff}}}$ and $\mathbf{W}_2^{(\ell)} \in \mathbb{R}^{d_{\text{ff}}\times d}$, with residual connections and LayerNorm around each sublayer:

$$\mathbf{H}^{(\ell)} = F^{(\ell)}(\mathbf{H}^{(\ell-1)}), \quad \ell = 1, \dots, L. \tag{1}$$

We define the *layer Jacobian* as $\mathbf{J}^{(\ell)} = \frac{\partial \operatorname{vec}(\mathbf{H}^{(\ell)})}{\partial \operatorname{vec}(\mathbf{H}^{(\ell-1)})}$. Later references to the stable ranks of $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O, \mathbf{W}_1, \mathbf{W}_2$ and to "Jacobian alignment between adjacent layers" all refer to this architecture and these $\mathbf{J}^{(\ell)}$.

3.2 Observations during training failure

Observation 1: Sharp Stable Rank Decline The first observable phenomenon is the rapid decline in weight stable rank. As shown in Figure 1 (left panel), the geometric averaged stable rank of the first several layers drops sharply around step 20000, preceding the gradient explosion. This phenomenon indicates energy concentration in top singular values.

Definition 3.1 (Stable Rank)**.** *The stable rank of a matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$ is:*

$$\operatorname{srank}(\mathbf{W}) = \frac{\|\mathbf{W}\|_F^2}{\|\mathbf{W}\|_2^2} = \frac{\sum_i s_i^2}{s_1^2}, \tag{2}$$

where $s_1 \geq s_2 \geq \cdots \geq 0$ are the singular values.

Observation 2: Increasing Jacobian Alignment Between Adjacent Layers The second critical phenomenon is the increasing alignment between Jacobians of adjacent layers, as illustrated in Figure 1 (right panel).

Definition 3.2 (Matrix Alignment)**.** *For any two matrices $\mathbf{A}, \mathbf{B}$ for which the product $\mathbf{AB}$ is well-defined, their alignment is the cosine similarity between the top right singular vector of $\mathbf{A}$ and the top left singular vector of $\mathbf{B}$:*

$$\mathit{Align}(\mathbf{A}, \mathbf{B}) = |\mathbf{v}_{A,1}^T \mathbf{u}_{B,1}|, \tag{3}$$

where $\mathbf{v}_{A,1}$ is the first right singular vector of $\mathbf{A}$ and $\mathbf{u}_{B,1}$ is the first left singular vector of $\mathbf{B}$. For simplicity, we denote the alignment of Jacobians for adjacent layers, $\mathit{Align}(\mathbf{J}^{(\ell+1)}, \mathbf{J}^{(\ell)})$, as $\mathit{Align}(\ell+1, \ell)$.

High alignment correlates strongly with both weight scale growth and stable rank decline, as well as gradient growth during the failure phase.

3.3 Conjecture: Low Stable Rank + Jacobian Alignment Drives Training Failure

Based on these two consistent phenomena, we conjecture that the combination of low weight stable rank and high Jacobian alignment creates a destabilizing feedback mechanism that leads to training failure. The remainder of this paper develops this hypothesis theoretically and proposes a solution.

4 Theoretical Analysis: Understanding the Failure Mechanism

We now provide theoretical analysis of a sufficient failure mechanism suggested by the observations. Our analysis is divided into two parts: (1) deriving a near-rank-one Jacobian channel and the additional signal-coupling conditions that make it amplify parameter gradients, and (2) analyzing stable-rank feedback under explicit mixed-moment conditions.

4.1 Part I: From Observations to Training Failure

In this section, we establish a conditional chain: near-rank-one layer Jacobians with sufficiently large leading singular values and coherent alignment form an amplified channel; terminal excitation and local parameter sensitivity then convert that channel into large weight gradients.

Remark 4.1 (Simplifying Assumption)**.** *We adopt the standard assumption that large gradient norms indicate training instability (Pascanu et al., 2012; Philipp et al., 2017). Our analysis derives lower bounds for gradient norms, with the understanding that such bounds indicate increased risk of divergence.*

4.1.1 High Jacobian Alignment + High Layer Jacobian Norm ⇒ High Total Jacobian Norm

The total Jacobian $\mathbf{J}_{total} = \prod_{\ell=1}^{L} \mathbf{J}^{(\ell)}$ determines how perturbations at the input propagate to the output. Stable rank near one means that each Jacobian is close to its dominant rank-one component, while the observed adjacent alignment measures how coherently these components connect across depth. The following theorem quantifies the resulting dominant path and its robustness to the residual spectral tail.

To formalize this, we use the definition of matrix alignment (Definition 3.2). This quantity measures how well the "output direction" of $\mathbf{J}^{(\ell)}$ matches the "input direction" of $\mathbf{J}^{(\ell+1)}$, determining whether the composition $\mathbf{J}^{(\ell+1)}\mathbf{J}^{(\ell)}$ preserves or diminishes the spectral norm.

Theorem 4.2 (Near-Rank-One Jacobian Product Lower Bound)**.** *For $L \geq 1$, let each nonzero compatible layer Jacobian have dominant singular triplet $(\sigma_\ell, \mathbf{u}_\ell, \mathbf{v}_\ell)$ and write*

$$\mathbf{J}^{(\ell)} = \mathbf{R}_\ell + \mathbf{E}_\ell, \qquad \mathbf{R}_\ell = \sigma_\ell \mathbf{u}_\ell \mathbf{v}_\ell^T. \tag{4}$$

Assume $\sigma_\ell \geq M > 0$, $\varepsilon_\ell \geq 0$, adjacent dominant directions satisfy $|\mathbf{v}_{\ell+1}^T \mathbf{u}_\ell| \geq a > 0$, and

$$\operatorname{srank}(\mathbf{J}^{(\ell)}) \leq 1 + \varepsilon_\ell^2. \tag{5}$$

Then $\|\mathbf{E}_\ell\|_2 \leq \varepsilon_\ell \sigma_\ell$ and

$$\|\mathbf{J}_{total}\|_2 \geq \left(\prod_{\ell=1}^{L} \sigma_\ell \right) \left(a^{L-1} - \sum_{\ell=1}^{L} \varepsilon_\ell \right). \tag{6}$$

In particular, if the bracket is nonnegative, this is at least $M^L(a^{L-1} - \sum_\ell \varepsilon_\ell)$. For exact rank-one Jacobians, $\varepsilon_\ell = 0$ and hence $\|\mathbf{J}_{total}\|_2 \geq M^L a^{L-1}$.

Corollary 4.3 (Alignment Margin Absorbs Stable-Rank Excess)**.** *Under the setup of Theorem 4.2, fix $\delta \geq 0$ with $a + \delta \leq 1$. If*

$$|\mathbf{v}_{\ell+1}^T \mathbf{u}_\ell| \geq a + \delta \quad \text{and} \quad \operatorname{srank}(\mathbf{J}^{(\ell)}) \leq 1 + \left(\frac{(a+\delta)^{L-1} - a^{L-1}}{L} \right)^2 \tag{7}$$

for every layer (with the alignment condition imposed for $\ell = 1, \ldots, L-1$), then

$$\|\mathbf{J}_{total}\|_2 \geq M^L a^{L-1}. \tag{8}$$

Proof. Choose $\varepsilon_\ell = ((a+\delta)^{L-1} - a^{L-1})/L$ for every layer. The stable-rank condition controls each spectral tail by $\|\mathbf{E}_\ell\|_2 \le \varepsilon_\ell \sigma_\ell$, while $\sum_{\ell=1}^{L} \varepsilon_\ell = (a+\delta)^{L-1} - a^{L-1}$. Applying Theorem 4.2 with adjacent alignment $a+\delta$ leaves the factor a^{L-1}. □

The proof is provided in Appendix B.1.

4.1.2 Stable Rank and Layer-Jacobian Upper Envelopes

We analyze the relationship between stable rank and layer Jacobian norm across transformer layer types.

Linear Layers.

Theorem 4.4 (Stable Rank Controls Jacobian Norm: Linear Layer)**.** *For a linear layer with nonzero weight matrix $\mathbf{W} \in \mathbb{R}^{m\times n}$ and fixed Frobenius norm $\|\mathbf{W}\|_F = F > 0$, the operator norm satisfies:*

$$\|\mathbf{W}\|_2 = \frac{F}{\sqrt{\mathrm{srank}(\mathbf{W})}}. \tag{9}$$

The proof is provided in Appendix B.3. For a linear layer this identity exactly shows that, at fixed Frobenius norm, decreasing weight stable rank increases the layer Jacobian norm.

Attention and MLP Layers. For self-attention and MLP blocks, stable rank enters useful upper envelopes. For single-head attention with projections $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O$, the output Jacobian operator norm is bounded by $\|D\mathbf{Y}(\mathbf{H})\|_{F\to F} \le \|\mathbf{A}\|_2\|\mathbf{W}_V\|_2\|\mathbf{W}_O\|_2 + 4\chi\|\mathbf{H}\|_2^2\|\mathbf{W}_Q\|_2\|\mathbf{W}_K\|_2\|\mathbf{W}_V\|_2\|\mathbf{W}_O\|_2/\sqrt{d_k}$. For a two-layer MLP, the corresponding bound is $\|\mathbf{J}_{\mathrm{MLP}}\|_2 \le \frac{L_\phi \|\mathbf{W}_1\|_F \|\mathbf{W}_2\|_F}{\sqrt{\mathrm{srank}(\mathbf{W}_1)\cdot\mathrm{srank}(\mathbf{W}_2)}}$. Lower stable rank enlarges these admissible upper envelopes at fixed Frobenius norms. The realized nonlinear-block Jacobian additionally depends on activation and directional alignment, represented explicitly by the leading singular-value lower bounds and spectral-tail controls in the amplification result below. Formal upper-bound statements and derivations are provided in Appendix B.4.

4.1.3 High Total Jacobian Norm ⇒ Large Weight Gradient Norm

We now show that high total Jacobian norms translate into large weight gradients. By the chain rule, the gradient with respect to weight vector $\hat{\mathbf{v}}_{out}^{(i)}$ at layer i decomposes as:

$$\frac{\partial L}{\partial \hat{\mathbf{v}}_{out}^{(i)}} = \left(\frac{\partial \mathbf{h}^{(i)}}{\partial \hat{\mathbf{v}}_{out}^{(i)}}\right)^T \left(\frac{\partial \mathbf{h}^{(L)}}{\partial \mathbf{h}^{(i)}}\right)^T \frac{\partial L}{\partial \mathbf{h}^{(L)}}. \tag{10}$$

Theorem 4.5 (Weight Gradient Norm Lower Bound)**.** *Fix $i < L$ and let the suffix Jacobians $\mathbf{J}^{(i+1)}, \ldots, \mathbf{J}^{(L)}$ satisfy the decomposition and assumptions of Theorem 4.2, with top singular values σ_k, and residual parameters ε_k. Let $\mathbf{g}_L = \partial L/\partial \mathbf{h}^{(L)}$ and suppose*

$$|\mathbf{u}_L^T \mathbf{g}_L| \ge (a+\delta)\|\mathbf{g}_L\|_2, \tag{11}$$

while all adjacent dominant-direction alignments are at least $a+\delta$, where $\delta \ge 0$ and $a+\delta \le 1$. For $\mathbf{A}_i = \partial \mathbf{h}^{(i)}/\partial \hat{\mathbf{v}}_{out}^{(i)}$, define the dominant-direction sensitivity $\gamma_i \triangleq \|\mathbf{A}_i^T \mathbf{v}_{i+1}\|_2 \ge \gamma > 0$ and assume $\|\mathbf{A}_i\|_2 \le \kappa\gamma_i$. If the suffix stable-rank residual obeys

$$\kappa \sum_{k=i+1}^{L} \varepsilon_k \le (a+\delta)^{L-i} - a^{L-i}, \tag{12}$$

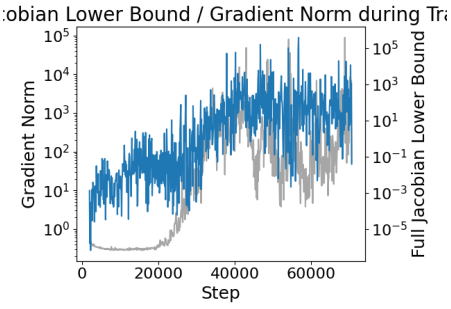

Figure 2: Jacobian-channel diagnostic versus the observed gradient norm. Their co-evolution is consistent with the near-rank-one transmission mechanism analyzed in Theorem 4.5.

then

$$\left\| \frac{\partial L}{\partial \hat{\mathbf{v}}_{out}^{(i)}} \right\|_2 \geq \gamma (aM)^{L-i} \|\mathbf{g}_L\|_2. \tag{13}$$

The proof is provided in Appendix B.6. Figure 2 compares the Jacobian-channel diagnostic with the observed gradient norm.

Theorem 4.6 (Total Gradient Norm Lower Bound)**.** *Suppose Theorem 4.5 holds columnwise for n_w columns in each layer $i < L$, and the terminal layer satisfies $\|\nabla_{\mathbf{W}^{(L)}} L\|_F^2 \geq n_w \gamma^2 \|\mathbf{g}_L\|_2^2$. Then*

$$\sum_{i=1}^{L} \left\| \frac{\partial L}{\partial \mathbf{W}^{(i)}} \right\|_F^2 \geq C \sum_{k=0}^{L-1} (aM)^{2k}, \tag{14}$$

where $C = \gamma^2 n_w \|\mathbf{g}_L\|_2^2$. For $(aM)^2 \neq 1$, the sum is $((aM)^{2L} - 1)/((aM)^2 - 1)$; for $aM = 1$, it equals L. When $aM > 1$, the certificate grows exponentially with depth.

The proof is provided in Appendix B.8.

4.1.4 Summary: The Failure Pathway

Combining these results yields a sufficient failure pathway: (1) Weight stable rank determines the operator norm of linear layers at fixed Frobenius norm and provides upper envelopes for attention and MLP Jacobians; (2) Near-rank-one Jacobians with $\sigma_\ell \geq M$, controlled spectral tails, and adjacent dominant-direction alignment form an amplified Jacobian channel (Theorem 4.2); and (3) Terminal excitation and controlled local sensitivity convert this channel into large weight gradients when task loss excites it (Theorems 4.5, 4.6). Together, these provide a formal mechanism by which spectral concentration and Jacobian alignment trigger gradient explosion and training instability.

4.2 Part II: The Positive Feedback Mechanism

Having established this conditional amplification pathway, we now analyze how spectral concentration can self-reinforce during training.

Low-Rank Signal Propagation. Gradients in backpropagation inherit outer-product structures ($\nabla_{\mathbf{W}} L = \tilde{\mathbf{H}}^T \mathbf{H}$). If hidden states $\mathbf{H}^{(\ell-1)}$ or cohidden states $\tilde{\mathbf{H}}^{(\ell)}$ become low-rank, the resulting weight gradients $\nabla_{\mathbf{W}} L$ are strictly rank-bounded by $\min(\text{rank}(\mathbf{H}), \text{rank}(\tilde{\mathbf{H}}))$. This causes parameter updates to concentrate in low-dimensional subspaces.

Accelerated Stable Rank Collapse. When the dominant component of the input–cohidden mixed moment shares the singular bases of a weight matrix $\mathbf{W} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, gradient descent can enlarge the leading singular value faster in relative terms than the remaining singular values. Theorem C.2 gives the exact common-basis result, while Proposition C.3 removes the common-basis condition in the small-step regime and Corollary C.4 derives its decrease condition from input–cohidden alignment, energy concentration, and the weight spectrum. Together these results provide a concrete self-reinforcing route toward spectral energy concentration. Formal derivations are provided in Appendix C.

5 The MSign Optimizer: Breaking the Feedback Loop

Based on our theoretical analysis and empirical observations, we propose the MSign optimizer to target the observed spectral-collapse mode by periodically restoring the stable rank of weight matrices via the matrix sign operation.

5.1 The Matrix Sign Operation

Definition 5.1 (Matrix Sign Operator)**.** *For any nonzero matrix $\mathbf{W} \in \mathbb{R}^{m \times n}$ with* reduced *(thin) SVD $\mathbf{W} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, where $r = \text{rank}(\mathbf{W}) > 0$, $\mathbf{U} \in \mathbb{R}^{m \times r}$, $\mathbf{S} \in \mathbb{R}^{r \times r}$, and $\mathbf{V} \in \mathbb{R}^{n \times r}$, we define:*

$$\text{sign}(\mathbf{W}) = \mathbf{U}\mathbf{V}^T \in \mathbb{R}^{m \times n}. \tag{15}$$

Algorithm 1 MSign Optimizer

Input: parameters θ, gradients g, learning rate η, period P, step t, target layers
Procedure StepWithMSign:
$\theta \leftarrow \text{BaseOptimizerStep}(\theta, g, \eta)$ ▷ Standard optimizer update
if $t \bmod P == 0$ **then**
  for each parameter $\mathbf{W}$ in target layers **do**
    if $\mathbf{W}$.ndim ≥ 2 **then**
      $F \leftarrow \|\mathbf{W}\|_F$ ▷ Record Frobenius norm
      if $F > 0$ **then**
        $\mathbf{U}, \mathbf{S}, \mathbf{V}^T \leftarrow \text{ReducedSVD}(\mathbf{W})$
        $\mathbf{W} \leftarrow \frac{F}{\|\mathbf{U}\mathbf{V}^T\|_F}\mathbf{U}\mathbf{V}^T$ ▷ Matrix sign with norm restoration
      end if
    end if
  end for
end if

This operation sets all nonzero singular values to 1, creating a rank-r partial isometry with $\|\text{sign}(\mathbf{W})\|_2 = 1$, $\|\text{sign}(\mathbf{W})\|_F = \sqrt{r}$, and maximal stable rank r among matrices of rank r. The reduced SVD ensures that $\mathbf{U}\mathbf{V}^T$ has shape $m \times n$. For $\mathbf{W} = 0$, we define the MSign intervention as a no-op.

5.2 Practical Implementation

Scaling. For $\mathbf{W} \neq 0$, we preserve the original Frobenius norm by

$$\mathbf{W}_{\text{new}} = \frac{\|\mathbf{W}\|_F}{\|\text{sign}(\mathbf{W})\|_F}\text{sign}(\mathbf{W}). \tag{16}$$

Here $\|\text{sign}(\mathbf{W})\|_F = \sqrt{r} > 0$. For $\mathbf{W} = 0$, the guarded algorithm leaves $\mathbf{W}_{\text{new}} = 0$. In our current implementation, we adopt this Frobenius-norm preserving rescaling; when stable rank is sufficiently low, it may excessively amplify minor singular values, and more principled rescaling schemes are left for future work.

Cost-Reduction Strategies. Computing full SVD at every step is prohibitively expensive. We reduce invocation frequency and scope as follows:

Periodic Application. MSign is applied every P steps (e.g., $P = 100$), preserving its stabilizing effect in the tested configurations while substantially reducing computational overhead.

Selective Layer Targeting. Based on our empirical findings, MSign can be selectively applied to the most critical layers:

- **Attention-only**: Apply only to self-attention weights ($\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O$).
- **2D-parameters-only**: Apply to all 2D parameter tensors (excluding biases and layer norm parameters).

5.3 Computational Cost Analysis

The arithmetic overhead of MSign is dominated by the SVD computation ($O(d^3)$ per targeted weight matrix). Because MSign is applied periodically every P steps (reported default $P = 100$), the amortized theoretical FLOPs ratio R is small:

$$R = \frac{\text{FLOPs}_{\text{MSign}}}{\text{FLOPs}_{\text{original}} \times P} \approx \frac{52d^3}{(72BTd^2 + 12BT^2d) \times P} < 0.1\%, \tag{17}$$

for standard transformer dimensions, batch size B, and sequence length T. Detailed FLOPs derivations and concrete numerical breakdowns are provided in Appendix A.3.

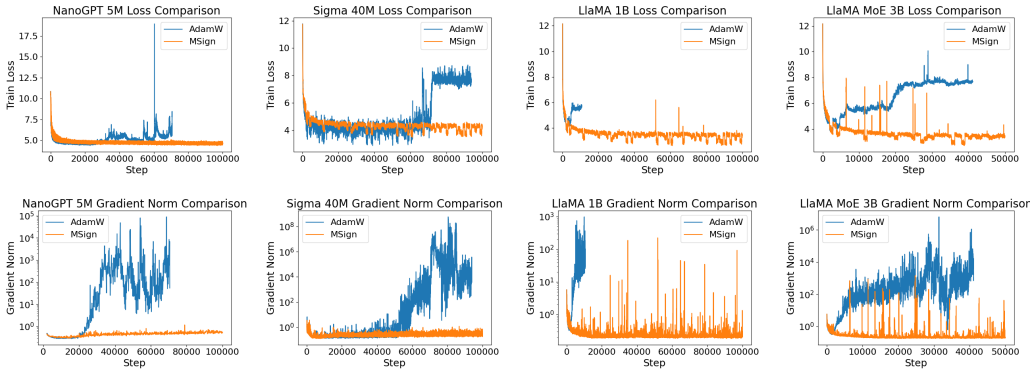

Figure 3: MSign prevents the observed failures in the reported runs. **Top row:** Training loss comparison between baseline (blue) and MSign (orange). Baseline training collapses with sudden loss spikes, while MSign maintains stable convergence. **Bottom row:** Corresponding gradient norm dynamics. Baseline runs exhibit rapid gradient growth preceding collapse, while MSign keeps gradient norms bounded throughout training. Training is terminated after failure to conserve computational resources. The intervention behavior is consistent with disrupting the proposed stable-rank feedback mechanism. From left to right: NanoGPT-5M, Sigma-40M, LLaMA-1B, LLaMA-MoE-3B.

6 Experiment

We validate our approach across multiple experimental settings of increasing scale. Our experiments demonstrate: (1) the effectiveness of MSign in preventing training failures, and (2) the minimal impact on training throughput.

6.1 Experimental Setting

We conduct experiments on four model configurations spanning different scales and architectures: NanoGPT-5M (No RoPE) [2]; Sigma-40M (Hybrid Attention) Qu et al. (2025); Hu et al. (2025); LLaMA-1B (Full RoPE) Kumar et al. (2025); LLaMA-MOE-3B (Mixture of Experts) (Modified based on LLaMA-1B). Detailed configurations are provided in Table 4 in Appendix A.

Table 1: Training throughput (tokens/second) across model scales. Measured overhead significantly exceeds theoretical predictions due to implementation factors discussed in Section 6.3.

	NanoGPT-5M	**Sigma-40M**	**LLaMA-1B**	**LLaMA-MoE-3B**
AdamW	102,708	6,504	1,742	544
MSign	105,199	6,097	1,640	520
Measured Overhead	-2.4%	6.7%	6.2%	4.6%
Theoretical Overhead	$< 0.01\%$	$< 0.01\%$	0.03%	0.09%

6.2 Main Result

MSign Prevents Observed Collapse Events. Figure 3 demonstrates that MSign prevents collapse across all four model scales (NanoGPT-5M, Sigma-40M, LLaMA-1B, and LLaMA-MoE-3B). In baseline training (blue curves), loss spikes and divergence occur predictably (e.g., step 30k for NanoGPT-5M, 50k for Sigma-40M, 2k for LLaMA-1B, 3k for MoE-3B), preceded by rapid gradient norm growth up to 10^1–10^7. In contrast, MSign-enabled runs (orange) maintain stable convergence with gradient norms bounded near 10^0. MSign is equally effective on MoE architectures. Furthermore, without MSign, weight matrix stable rank declines sharply to $\mathrm{srank} \approx 1.2$ before collapse, whereas MSign ($P = 100$) maintains stable rank above 5 throughout training, confirming that MSign disrupts the spectral-collapse feedback loop.

[2] https://github.com/karpathy/nanoGPT

Table 2: Layer selection ablation: test perplexity (↓) on NanoGPT-5M and Sigma-40M.

Target Layers	NanoGPT	Sigma
AdamW	Failed	Failed
MSign (All 2D)	102.6	74.00
MSign (Attn only)	118.6	75.68
MSign (MLP only)	Failed	Failed

Table 3: Application period ablation: test perplexity (↓) and throughput (tokens/s) across models.

Period P	Test PPL		Throughput
	NanoGPT	Sigma	Sigma
$P = 10$	103.9	92.7	18,236
$P = 100$	102.6	74.0	24,559
$P = 1000$	99.4	75.7	25,082
$P = 10000$	104.2	**69.5**	25,270
$P = 100000$	Failed	Failed	—

6.3 Throughput Analysis

Table 1 compares training throughput across model scales. For NanoGPT-5M, the measured overhead (-2.4%) falls within system noise, consistent with the theoretical prediction ($< 0.1\%$). For larger models, however, measured overhead (4.6–6.7%) significantly exceeds theoretical values. This discrepancy arises from implementation factors not captured in FLOPs analysis: (1) `all_gather` communication for distributed SVD, (2) disruption of FlashAttention kernel fusion and continuous stream execution, and (3) pipeline bubbles in distributed training. Despite these overheads, the 4–7% throughput reduction remains modest compared to the computational waste from training failures.

6.4 Ablation Study

We conduct detailed ablation studies on the NanoGPT-5M and Sigma-40M configurations to identify the key factors affecting MSign effectiveness.

Ablation 1: Layer Selection: Attention vs. MLP. Table 2 shows that applying MSign to MLP layers alone fails to prevent collapse, whereas attention-only application stabilizes training. Thus attention projections are the effective intervention target in this tested configuration, consistent with the proposed inter-layer Jacobian mechanism. Applying MSign to all 2D parameters achieves the best reported perplexity (e.g., 102.6 vs. 118.6 on NanoGPT-5M), indicating an additional empirical benefit from including MLP layers.

Ablation 2: Application Period P. The application period P controls the trade-off between computational overhead and stable rank maintenance. Table 3 demonstrates the robustness of MSign across application periods P from 10 to 10,000. All tested values from $P = 10$ to $10,000$ prevent collapse. As shown in Figure 4, $P = 10,000$ exhibits greater loss and gradient-norm variability between steps 20k–40k than the shorter-period configurations, indicating weaker control between interventions. We use $P = 100$ as the default because it combines stable training dynamics with modest computational overhead.

7 Conclusion

We identify stable rank collapse and Jacobian alignment as consistent precursors to LLM training instability and derive a sufficient amplification mechanism under explicit realized-gain, tail, alignment, and task-coupling conditions. The proposed MSign optimizer periodically restores weight matrix stable rank. Across models from 5M to 3B parameters, MSign reliably prevents the observed training collapse with minimal computational overhead ($< 7\%$).

Limitations and Future Work. While our theoretical analysis models the self-reinforcing collapse, extending the feedback alignment assumptions to general non-aligned scenarios remains an open problem. Future work includes adaptive scheduling, fused GPU kernels, and broader spectral regularization.

IMPACT STATEMENT

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

AI USE STATEMENT

Generative AI tools were used for language editing and LaTeX formatting assistance. All AI-assisted output was reviewed by the authors for correctness, originality, and appropriate attribution. The authors take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REFERENCES

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. volume 32, 2019.

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016. URL http://arxiv.org/abs/1607.06450.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. ISSN 1532-4435.

Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando De Freitas. Predicting parameters in deep learning. volume 26, 2013.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. 2019.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Qingguo Hu, Zhenghao Lin, Ziyue Yang, Yucheng Ding, Xiao Liu, Yuting Jiang, Ruizhe Wang, Tianyu Chen, Zhongxin Guo, Yifan Xiong, Rui Gao, Lei Qu, Jinsong Su, Peng Cheng, and Yeyun Gong. Sigma-MoE-Tiny technical report, 2025. URL https://arxiv.org/abs/2512.16248.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=BH8TYy0r6u.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. 2020. URL https://arxiv.org/abs/2001.08361.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Abhay Kumar, Louis Owen, Nilabhra Roy Chowdhury, and Fabian G"ura. Zclip: Adaptive spike mitigation for LLM pre-training, 2025. URL https://doi.org/10.48550/arXiv.2504.02507.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *ICLR (Poster)*, 2018. URL https://openreview.net/forum?id=ryup8-WCW.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *CoRR*, abs/1707.09564, 2017. URL http://arxiv.org/abs/1707.09564.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *30th International Conference on Machine Learning, ICML 2013*, pp. 1310–1318, 11 2012.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *CoRR*, abs/1711.04735, 2017. URL http://arxiv.org/abs/1711.04735.

George Philipp, Dawn Song, and Jaime G Carbonell. The exploding gradient problem demystified - definition, prevalence, impact, origin, tradeoffs, and solutions. *arXiv preprint arXiv:1712.05577*, 2017.

Lei Qu, Lianhai Ren, Peng Cheng, Rui Gao, Ruizhe Wang, Tianyu Chen, Xiao Liu, Xingjian Zhang, Yeyun Gong, Yifan Xiong, Yucheng Ding, Yuting Jiang, Zhenghao Lin, Zhongxin Guo, and Ziyue Yang. SIGMA: An AI-empowered training stack on early-life hardware, 2025. URL https://arxiv.org/abs/2512.13488.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4):21–es, July 2007. ISSN 0004-5411. doi: 10.1145/1255443.1255449. URL https://doi.org/10.1145/1255443.1255449.

Amartya Sanyal, Philip H. Torr, and Puneet K. Dokania. Stable rank normalization for improved generalization in neural networks and gans. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1enKkrFDB.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. volume abs/1312.6120, 2013. URL https://api.semanticscholar.org/CorpusID:17272965.

Lloyd N. Trefethen and David Bau III. *Numerical Linear Algebra*. SIAM, Philadelphia, 1997.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *International Conference on Learning Representations*, 2024.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *International conference on machine learning*, pp. 5393–5402. PMLR, 2018.

Greg Yang and Samuel S. Schoenholz. Mean field residual networks: On the edge of chaos. *CoRR*, abs/1712.08969, 2017. URL http://arxiv.org/abs/1712.08969.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. *CoRR*, abs/2203.03466, 2022. doi: 10.48550/ARXIV.2203.03466. URL https://doi.org/10.48550/arXiv.2203.03466.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: An open bilingual pre-trained model. 2023. URL https://arxiv.org/abs/2210.02414.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. 2024.

A Experimental Details

A.1 Model Configurations

Table 4 summarizes the experimental configurations across four model scales used in our experiments.

Table 4: Experimental configurations across four model scales

Setting	NanoGPT-5M	Sigma-40M	LLaMA-1B	LLaMA-MoE-3B
Parameters	5M	40M	1B	1B (active) / 3B (total)
Layers	24	28	16	16
Hidden dim	48	128	2048	2048
Attention heads	6	4	16	16
Attention type	MHA	MHA/MLA alternating	MHA	MHA
Architecture	Dense	Dense	Dense	MoE (16 experts, top-4)
MLP intermediate	192	448	5440	1360 per expert
Position encoding	None	RoPE (MHA only)	RoPE	RoPE
Layernorm	LayerNorm	RMSNorm	RMSNorm	RMSNorm
Activation	GELU	SiLU	SiLU	SiLU
Initial Std[1,2]	0.02	0.006	0.02	0.02
Dataset	OpenWebText	Nemotron-cc	Nemotron-cc	Nemotron-cc
Batch size (tokens)	123k	66k	66k	66k
Sequence length	256	2048	2048	2048
Learning rate[3]	6×10^{-4}	3.5×10^{-4}	1×10^{-3}	1×10^{-3}
Optimizer	AdamW	AdamW	AdamW	AdamW

[1]For NanoGPT-5M and Sigma-40M, the μP scaling rules are applied: the initial 2D weights in transformer layers are scaled by $4\times$, and query weights ($\mathbf{W}_Q$) are initialized to zero.
[2]For NanoGPT-5M and Sigma-40M, the output projection weights ($\mathbf{W}_O, \mathbf{W}_{down}$) are further divided by $\sqrt{2 \cdot \#\text{layers}}$ as part of the original architecture design.
[3]For NanoGPT-5M and Sigma-40M, the μP scaling rules are applied: the learning rate for 2D weights in transformer layers are scaled by $16\times$, and for embedding weights are scaled by $4\times$.

A.2 Common Training Settings

The following hyperparameters and architectural choices are shared across all experiments unless otherwise specified:

Optimizer settings. We use the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$. Weight decay is set to 0.1 for all experiments. Gradient clipping with max norm 1.0 is applied in all experiments. For the learning rate scheduler, we warm up for 2000 steps and linearly decay to $1/10$ learning rate.

Architecture details. Bias terms are disabled in all linear layers (attention projections and MLP layers). Pre-LayerNorm is used in all transformer blocks.

A.3 Computational Cost Details

Table 5 provides a detailed arithmetic-cost breakdown for a typical transformer layer. The SVD cost is $O(d^3)$ per targeted matrix; amortization over P steps makes the modeled FLOPs ratio small. This estimate excludes communication, synchronization, kernel-fusion disruption, and pipeline effects, which explain why the measured wall-clock penalty in Table 1 is 4.6–6.7% on the larger reported models.

A.4 Throughput Model Analysis

To quantify the relationship between application period P and training throughput, we fit a simple analytical model. Let f denote the baseline computation per token (fixed), and F denote the additional computation per MSign application (fixed). When the period is P, the amortized overhead per token

Table 5: Computational cost comparison per layer (hidden dim d, intermediate dim $4d$, batch size B, sequence length T, applied every P steps)

Component	FLOPs per step	MSign FLOPs†	Amortized MSign
Attention Block			
$\mathbf{W}_Q/\mathbf{W}_K/\mathbf{W}_V/\mathbf{W}_O$	$6BTd^2$	$4 \times 13d^3$	$52d^3/P$
Attention computation	$12BT^2d + O(BT^2)$	0	0
Total	$24BTd^2 + 12BT^2d + O(BT^2)$	$52d^3$	$52d^3/P$
MLP Block			
$\mathbf{W}_{up}/\mathbf{W}_{down}$ ($d \to 4d/4d \to d$)	$24BTd^2$	$2 \times 19d^3$	$38d^3/P$
Activation	$O(BTd)$	0	0
Total	$48BTd^2 + O(BTd)$	$38d^3$	$38d^3/P$
Entire Layer			
Original	$72BTd^2 + 12BT^2d + O(BT^2 + BTd)$	0	0
+ MSign (Attn only)	$72BTd^2 + 12BT^2d + O(BT^2 + BTd)$	$52d^3$	$52d^3/P$
+ MSign (All 2D)	$72BTd^2 + 12BT^2d + O(BT^2 + BTd)$	$90d^3$	$90d^3/P$

†SVD FLOPs computed as $2mn\min(m,n) + 11\min(m,n)^3$ per matrix (Trefethen & Bau III, 1997). For $d \times d$ matrix: $13d^3$; for $d \times 4d$ matrix: $19d^3$.

is F/P. The throughput $T(P)$ (tokens/s) can be modeled as:

$$T(P) = \frac{T_\infty}{1 + r/P}, \quad \text{where } r = F/f, \tag{18}$$

and T_∞ is the asymptotic throughput as $P \to \infty$ (i.e., baseline without MSign overhead).

Using the Sigma-40M throughput measurements from Table 3, we perform least-squares fitting by linearizing: $1/T(P) = (1/T_\infty)(1 + r/P)$. This yields:

$$T_\infty \approx 25{,}350 \text{ tokens/s}, \quad r \approx 3.9. \tag{19}$$

The fitted model predicts:

P	10	100	1000	10000
Measured	18,236	24,559	25,082	25,270
Predicted	18,273	24,399	25,251	25,340

The close agreement validates the model. However, the fitted $r \approx 3.9$ significantly exceeds the theoretical prediction. From Section 5.3, the FLOPs-based overhead ratio is $R = 52d^3/(72BTd^2 \cdot P) < 0.1\%$ for typical configurations, implying $r_{\text{theory}} \ll 1$.

This gap arises from the implementation factors discussed in Section 6.3: (1) `all_gather` synchronization latency for distributed SVD computation, (2) disruption of FlashAttention kernel fusion and continuous CUDA stream execution, and (3) pipeline bubbles in distributed training. These factors introduce latency-dominated overhead that scales poorly with batch size, explaining why the effective r far exceeds the FLOPs-based prediction. Future work on asynchronous MSign execution and fused SVD kernels could potentially close this gap.

A.5 Application Period Analysis

Figure 4 provides detailed training dynamics for different application periods P on NanoGPT-5M. The left figure shows training loss trajectories, while the right figure shows gradient norm evolution.

Several observations emerge from this analysis:

- $P = 10$ **and** $P = 100$: Both show smooth, stable training dynamics with minimal variance in loss and gradient norm. The frequent interventions maintain higher stable rank throughout the reported trajectories.

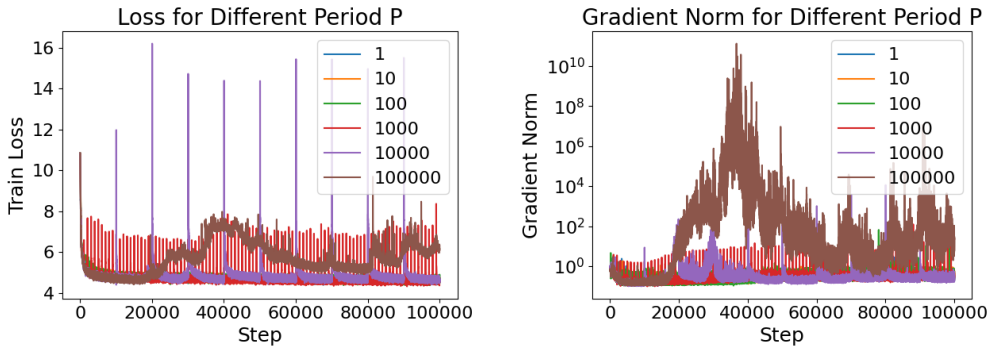

Figure 4: Training dynamics under different MSign application periods on NanoGPT-5M. **Left:** Training loss. **Right:** Gradient norm. All periods from $P = 10$ to $P = 10000$ reach stable final perplexity; $P = 10000$ shows higher gradient norms between steps 20,000 and 40,000, motivating the shorter default period.

- $P = 1000$: Training remains stable but shows slightly increased variance compared to smaller P values. The longer intervals between MSign applications allow some transient stable rank decline, but recovery occurs before instability develops.
- $P = 10000$: While training eventually converges (PPL 104.2), the dynamics exhibit clear signs of intermittent instability, the gradient norm shows periodic spikes and the loss curve has higher variance. This suggests that 10,000 steps is near the boundary where stable rank can decline sufficiently to trigger partial feedback loop activation before the next MSign intervention.

For the reported configurations, these findings motivate $P = 100$ as a conservative operating point. It combines smooth trajectories with a small modeled FLOPs ratio; the measured end-to-end overhead remains deployment-dependent and is reported separately.

B Proofs of Main Results

B.1 Proof of Theorem 4.2 (Near-Rank-One Jacobian Product Lower Bound)

By the definition of stable rank and the truncated SVD,

$$\|\mathbf{E}_\ell\|_F^2 = \sum_{j\geq 2} \sigma_{\ell,j}^2 = \sigma_\ell^2\big(\mathrm{srank}(\mathbf{J}^{(\ell)}) - 1\big), \tag{20}$$

so $\|\mathbf{E}_\ell\|_2 \leq \|\mathbf{E}_\ell\|_F \leq \varepsilon_\ell \sigma_\ell$. The dominant path has the exact form

$$\mathbf{R}_L \cdots \mathbf{R}_1 = \left(\prod_{\ell=1}^{L} \sigma_\ell\right)\left(\prod_{\ell=1}^{L-1} \mathbf{v}_{\ell+1}^T \mathbf{u}_\ell\right)\mathbf{u}_L \mathbf{v}_1^T, \tag{21}$$

and therefore its norm is at least $(\prod_\ell \sigma_\ell) a^{L-1}$. The product telescope gives

$$\mathbf{J}_L \cdots \mathbf{J}_1 - \mathbf{R}_L \cdots \mathbf{R}_1 = \sum_{j=1}^{L} \mathbf{J}_L \cdots \mathbf{J}_{j+1} \mathbf{E}_j \mathbf{R}_{j-1} \cdots \mathbf{R}_1. \tag{22}$$

Since $\|\mathbf{J}_\ell\|_2 = \|\mathbf{R}_\ell\|_2 = \sigma_\ell$, submultiplicativity yields

$$\|\mathbf{J}_L \cdots \mathbf{J}_1 - \mathbf{R}_L \cdots \mathbf{R}_1\|_2 \leq \left(\prod_{\ell=1}^{L} \sigma_\ell\right) \sum_{\ell=1}^{L} \varepsilon_\ell. \tag{23}$$

The reverse triangle inequality now gives the stated result. □

Proof of Corollary 4.3: Choose $\varepsilon_\ell = ((a+\delta)^{L-1} - a^{L-1})/L$ for every layer. The stable-rank condition controls each spectral tail by $\|\mathbf{E}_\ell\|_2 \le \varepsilon_\ell \sigma_\ell$, while $\sum_{\ell=1}^{L} \varepsilon_\ell = (a+\delta)^{L-1} - a^{L-1}$. Applying Theorem 4.2 with adjacent alignment $a+\delta$ leaves the factor a^{L-1}. □

B.2 Near-Rank-r Extension of the Jacobian Product Bound

The rank-one result extends to a dominant subspace of dimension r, although stable rank alone no longer controls all required quantities. For each compatible Jacobian, let

$$\mathbf{J}^{(\ell)} = \mathbf{R}_{\ell,r} + \mathbf{E}_{\ell,r}, \qquad \mathbf{R}_{\ell,r} = \mathbf{U}_{\ell,r}\mathbf{S}_{\ell,r}\mathbf{V}_{\ell,r}^T \tag{24}$$

be its top-r truncated SVD, and define the adjacent subspace-overlap matrix

$$\mathbf{C}_{\ell,r} = \mathbf{V}_{\ell+1,r}^T \mathbf{U}_{\ell,r}. \tag{25}$$

Write $m_\ell = \sigma_r(\mathbf{J}^{(\ell)})$, $s_\ell = \sigma_1(\mathbf{J}^{(\ell)})$, and $\varepsilon_{\ell,r} = \|\mathbf{E}_{\ell,r}\|_2 / s_\ell = \sigma_{r+1}(\mathbf{J}^{(\ell)})/s_\ell$.

Proposition B.1 (Near-Rank-r Jacobian Product Bound)**.** *Suppose $m_\ell > 0$ and $\sigma_{\min}(\mathbf{C}_{\ell,r}) \ge a_r > 0$ for all adjacent layers. Then*

$$\begin{aligned}\sigma_r(\mathbf{J}^{(L)} \cdots \mathbf{J}^{(1)}) \ge & \left(\prod_{\ell=1}^{L} m_\ell\right) a_r^{L-1} \\ & - \left(\prod_{\ell=1}^{L} s_\ell\right) \sum_{\ell=1}^{L} \varepsilon_{\ell,r}.\end{aligned} \tag{26}$$

Proof. The dominant product factors through the $r \times r$ core

$$\mathbf{R}_{L,r} \cdots \mathbf{R}_{1,r} = \mathbf{U}_{L,r}\mathbf{S}_{L,r}\mathbf{C}_{L-1,r} \cdots \mathbf{C}_{1,r}\mathbf{S}_{1,r}\mathbf{V}_{1,r}^T. \tag{27}$$

Thus repeated minimum-singular-value multiplication gives $\sigma_r(\mathbf{R}_{L,r} \cdots \mathbf{R}_{1,r}) \ge (\prod_\ell m_\ell) a_r^{L-1}$. The product telescope and $\|\mathbf{J}^{(\ell)}\|_2 = \|\mathbf{R}_{\ell,r}\|_2 = s_\ell$ give

$$\|\mathbf{J}^{(L)} \cdots \mathbf{J}^{(1)} - \mathbf{R}_{L,r} \cdots \mathbf{R}_{1,r}\|_2 \le \left(\prod_{\ell=1}^{L} s_\ell\right) \sum_{\ell=1}^{L} \varepsilon_{\ell,r}. \tag{28}$$

Weyl's singular-value perturbation inequality completes the proof. □

If the measured overlap has a margin $a_r + \delta_r$, the residual can again be absorbed. In particular, if

$$\left(\prod_{\ell=1}^{L} s_\ell\right) \sum_{\ell=1}^{L} \varepsilon_{\ell,r} \le \left(\prod_{\ell=1}^{L} m_\ell\right) \left((a_r+\delta_r)^{L-1} - a_r^{L-1}\right), \tag{29}$$

then applying Proposition B.1 with overlap $a_r + \delta_r$ yields the clean bound

$$\sigma_r(\mathbf{J}^{(L)} \cdots \mathbf{J}^{(1)}) \ge \left(\prod_{\ell=1}^{L} m_\ell\right) a_r^{L-1}. \tag{30}$$

The rank-r extension therefore controls the top-r channel conditioning m_ℓ/s_ℓ, the spectral tail $\varepsilon_{\ell,r}$, and the subspace overlap explicitly, in addition to the stable-rank profile.

B.3 Proof of Theorem 4.4 (Stable Rank Controls Jacobian Norm: Linear Layer)

Consider a linear layer that computes $\mathbf{h}_{out} = \mathbf{W}\mathbf{h}_{in}$, where $\mathbf{W} \in \mathbb{R}^{m \times n}$ is the weight matrix. The Jacobian of this transformation is:

$$\mathbf{J} = \frac{\partial \mathbf{h}_{out}}{\partial \mathbf{h}_{in}} = \mathbf{W}, \tag{31}$$

since the linear map $\mathbf{h} \mapsto \mathbf{W}\mathbf{h}$ has constant derivative equal to $\mathbf{W}$ itself.

The operator norm (also called spectral norm or 2-norm) of a matrix equals its largest singular value:

$$\|\mathbf{W}\|_2 = \sigma_1(\mathbf{W}) = \max_{\|\mathbf{x}\|_2=1} \|\mathbf{W}\mathbf{x}\|_2. \tag{32}$$

This represents the maximum factor by which the matrix can stretch any unit vector, i.e., the worst-case signal amplification.

From the definition of stable rank (Definition 1):

$$\operatorname{srank}(\mathbf{W}) = \frac{\|\mathbf{W}\|_F^2}{\|\mathbf{W}\|_2^2} = \frac{\sum_i \sigma_i^2}{\sigma_1^2}, \tag{33}$$

where $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$ are the singular values of $\mathbf{W}$.

Let $F = \|\mathbf{W}\|_F$ denote the Frobenius norm, which is fixed by assumption. Substituting into the stable rank definition:

$$\operatorname{srank}(\mathbf{W}) = \frac{F^2}{\|\mathbf{W}\|_2^2}. \tag{34}$$

Solving for the operator norm:

$$\|\mathbf{W}\|_2^2 = \frac{F^2}{\operatorname{srank}(\mathbf{W})} \quad \Rightarrow \quad \|\mathbf{W}\|_2 = \frac{F}{\sqrt{\operatorname{srank}(\mathbf{W})}}. \tag{35}$$

This formula reveals a fundamental trade-off: for a matrix with fixed total "energy" (Frobenius norm F), the operator norm, which determines signal amplification, is inversely proportional to the square root of stable rank.

Intuitively, stable rank measures how "spread out" the singular values are. When $\operatorname{srank}(\mathbf{W}) \approx \operatorname{rank}(\mathbf{W})$ (high stable rank), the singular values are roughly uniform, and $\|\mathbf{W}\|_2 \approx F/\sqrt{\operatorname{rank}(\mathbf{W})}$ is relatively small. When $\operatorname{srank}(\mathbf{W}) \approx 1$ (low stable rank), almost all energy is concentrated in the top singular value, and $\|\mathbf{W}\|_2 \approx F$, the maximum possible for given Frobenius norm. □

B.4 Layer-Specific Jacobian Norm Bounds

Here we state and prove the formal Jacobian norm bounds for Attention and MLP layers mentioned in Section 4.1.

Theorem B.2 (Jacobian Norm Bound: Attention Layer)**.** *Consider a single-head attention layer with $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d_k}$ and $\mathbf{W}_O \in \mathbb{R}^{d_k\times d}$. Let $\mathbf{H} \in \mathbb{R}^{n\times d}$, $\mathbf{S} = \mathbf{H}\mathbf{W}_Q\mathbf{W}_K^T\mathbf{H}^T$, $\mathbf{Z} = \mathbf{S}/\sqrt{d_k}$, $\mathbf{A} = \operatorname{softmax}(\mathbf{Z})$ row-wise, and $\mathbf{Y} = \mathbf{A}\mathbf{H}\mathbf{W}_V\mathbf{W}_O$. Matrix spaces are equipped with the Frobenius norm, and $\|D\mathcal{F}(\mathbf{H})\|_{F\to F}$ denotes the induced norm of a Fréchet derivative; $\|\cdot\|_2$ denotes matrix spectral norm. Define $\gamma_{\min}$ as the minimum top-two margin among the rows of the* scaled *logits $\mathbf{Z}$, and set $\chi = \min\{(n-1)e^{-\gamma_{\min}}, 1\}$. Then*

$$\|D\mathbf{A}(\mathbf{H})\|_{F\to F} \leq \frac{4\chi}{\sqrt{d_k}}\|\mathbf{H}\|_2\|\mathbf{W}_Q\|_2\|\mathbf{W}_K\|_2, \tag{36}$$

and

$$\begin{aligned}\|D\mathbf{Y}(\mathbf{H})\|_{F\to F} \leq{}& \|\mathbf{A}\|_2\|\mathbf{W}_V\|_2\|\mathbf{W}_O\|_2 \\ &+ \frac{4\chi\|\mathbf{H}\|_2^2}{\sqrt{d_k}}\|\mathbf{W}_Q\|_2\|\mathbf{W}_K\|_2\|\mathbf{W}_V\|_2\|\mathbf{W}_O\|_2.\end{aligned} \tag{37}$$

Theorem B.3 (Jacobian Norm Bound: MLP Layer)**.** *Consider a two-layer MLP with nonzero weights $\mathbf{W}_1 \in \mathbb{R}^{d\times d_{ff}}$ and $\mathbf{W}_2 \in \mathbb{R}^{d_{ff}\times d}$. Suppose ϕ is differentiable at the realized preactivations z_j and $|\phi'(z_j)| \leq L_\phi$ for every j. Then its Jacobian satisfies*

$$\|\mathbf{J}_{MLP}\|_2 \leq \frac{L_\phi\|\mathbf{W}_1\|_F\|\mathbf{W}_2\|_F}{\sqrt{\operatorname{srank}(\mathbf{W}_1)\cdot\operatorname{srank}(\mathbf{W}_2)}}, \tag{38}$$

where one may take $L_\phi = \sup_z |\phi'(z)|$ for a globally differentiable activation.

Proof of Theorem B.2: Matrix arguments and outputs of Fréchet derivatives carry the Frobenius norm, derivative maps carry the induced $F \to F$ norm, and multiplicative matrix factors carry spectral norm. For a perturbation $\Delta\mathbf{H}$,

$$D\mathbf{S}(\mathbf{H})[\Delta\mathbf{H}] = \Delta\mathbf{H}\mathbf{W}_Q\mathbf{W}_K^T\mathbf{H}^T + \mathbf{H}\mathbf{W}_Q\mathbf{W}_K^T(\Delta\mathbf{H})^T. \tag{39}$$

The inequality $\|AXB\|_F \le \|A\|_2\|X\|_F\|B\|_2$ therefore gives

$$\|D\mathbf{S}(\mathbf{H})\|_{F\to F} \le 2\|\mathbf{H}\|_2\|\mathbf{W}_Q\|_2\|\mathbf{W}_K\|_2. \tag{40}$$

For row i of the scaled logits $\mathbf{Z} = \mathbf{S}/\sqrt{d_k}$, let $\mathbf{a}_i = \mathrm{softmax}(\mathbf{z}_i)$ and let γ_i be the top-two margin of $\mathbf{z}_i$. Its softmax Jacobian is $\mathrm{diag}(\mathbf{a}_i) - \mathbf{a}_i\mathbf{a}_i^T$, which is positive semidefinite and has spectral norm at most

$$1 - \|\mathbf{a}_i\|_2^2 \le 2(1 - \max_j a_{i,j}) \le 2\min\{(n-1)e^{-\gamma_i}, 1\}. \tag{41}$$

Under vectorization, the row-wise derivative is block diagonal in the Frobenius norm. Thus, with $\gamma_{\min} = \min_i \gamma_i$ and $\chi = \min\{(n-1)e^{-\gamma_{\min}}, 1\}$,

$$\|D\mathbf{A}(\mathbf{H})\|_{F\to F} \le \frac{2\chi}{\sqrt{d_k}}\|D\mathbf{S}(\mathbf{H})\|_{F\to F} \le \frac{4\chi}{\sqrt{d_k}}\|\mathbf{H}\|_2\|\mathbf{W}_Q\|_2\|\mathbf{W}_K\|_2. \tag{42}$$

Finally, the product rule gives

$$D\mathbf{Y}(\mathbf{H})[\Delta\mathbf{H}] = D\mathbf{A}(\mathbf{H})[\Delta\mathbf{H}]\mathbf{H}\mathbf{W}_V\mathbf{W}_O + \mathbf{A}\Delta\mathbf{H}\mathbf{W}_V\mathbf{W}_O. \tag{43}$$

Applying the same mixed Frobenius–spectral norm inequality and the preceding bound yields exactly the two-term estimate stated in the theorem. □

Discussion: Connection to Stable Rank. Substituting $\|\mathbf{W}_i\|_2 = \|\mathbf{W}_i\|_F/\sqrt{\mathrm{srank}(\mathbf{W}_i)}$ into the bound reveals that low stable rank in a projection matrix enlarges the corresponding Jacobian upper envelope. In particular, the V–O pathway contribution to this envelope scales as $\|\mathbf{W}_V\|_2\|\mathbf{W}_O\|_2 \propto (\mathrm{srank}(\mathbf{W}_V)\mathrm{srank}(\mathbf{W}_O))^{-1/2}$ at fixed Frobenius norms; tightness requires additional activation and directional conditions.

B.5 Proof of Theorem B.3 (Jacobian Norm Bound: MLP Layer)

Consider the forward pass through a two-layer MLP. Given input $\mathbf{h} \in \mathbb{R}^d$, we first compute the hidden representation:

$$\mathbf{z} = \mathbf{h}\mathbf{W}_1 \in \mathbb{R}^{d_{\text{ff}}}, \tag{44}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d\times d_{\text{ff}}}$ projects from hidden dimension d to feedforward dimension d_{ff} (typically $d_{\text{ff}} = 4d$).

Next, we apply the element-wise activation function ϕ (such as GELU or SiLU):

$$\mathbf{a} = \phi(\mathbf{z}) = [\phi(z_1), \phi(z_2), \ldots, \phi(z_{d_{\text{ff}}})] \in \mathbb{R}^{d_{\text{ff}}}. \tag{45}$$

Finally, we project back to the hidden dimension:

$$\mathbf{y} = \mathbf{a}\mathbf{W}_2 = \phi(\mathbf{z})\mathbf{W}_2 \in \mathbb{R}^d, \tag{46}$$

where $\mathbf{W}_2 \in \mathbb{R}^{d_{\text{ff}}\times d}$.

To compute the Jacobian $\frac{\partial\mathbf{y}}{\partial\mathbf{h}}$, we apply the chain rule through this three-stage computation:

$$\frac{\partial\mathbf{y}}{\partial\mathbf{h}} = \frac{\partial\mathbf{y}}{\partial\mathbf{a}} \cdot \frac{\partial\mathbf{a}}{\partial\mathbf{z}} \cdot \frac{\partial\mathbf{z}}{\partial\mathbf{h}}. \tag{47}$$

Each factor has a simple form:

- $\frac{\partial\mathbf{z}}{\partial\mathbf{h}} = \mathbf{W}_1$: the Jacobian of a linear transformation is the weight matrix itself.
- $\frac{\partial\mathbf{a}}{\partial\mathbf{z}} = \mathrm{diag}(\phi'(\mathbf{z}))$: since ϕ acts element-wise, its Jacobian is diagonal with entries $\phi'(z_i)$.

- $\frac{\partial \mathbf{y}}{\partial \mathbf{a}} = \mathbf{W}_2$: again a linear transformation.

Combining these:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{h}} = \mathbf{W}_1 \cdot \text{diag}(\phi'(\mathbf{z})) \cdot \mathbf{W}_2. \tag{48}$$

To bound the operator norm, we use the submultiplicativity property $\|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2$:

$$\left\| \frac{\partial \mathbf{y}}{\partial \mathbf{h}} \right\|_2 = \|\mathbf{W}_1 \cdot \text{diag}(\phi'(\mathbf{z})) \cdot \mathbf{W}_2\|_2 \leq \|\mathbf{W}_1\|_2 \cdot \|\text{diag}(\phi'(\mathbf{z}))\|_2 \cdot \|\mathbf{W}_2\|_2. \tag{49}$$

The operator norm of a diagonal matrix is the maximum absolute value of its diagonal entries:

$$\|\text{diag}(\phi'(\mathbf{z}))\|_2 = \max_i |\phi'(z_i)| \leq \sup_{z \in \mathbb{R}} |\phi'(z)| = L_\phi, \tag{50}$$

where L_ϕ is the Lipschitz constant of ϕ. For commonly used activations: GELU has $L_{\text{GELU}} \approx 1.13$ (achieved near $z \approx -0.75$), and SiLU has $L_{\text{SiLU}} \approx 1.1$ (achieved near $z \approx 1.28$).

Thus:

$$\left\| \frac{\partial \mathbf{y}}{\partial \mathbf{h}} \right\|_2 \leq \|\mathbf{W}_1\|_2 \cdot L_\phi \cdot \|\mathbf{W}_2\|_2. \tag{51}$$

To express this in terms of stable rank, recall from Definition 1 that $\text{srank}(\mathbf{W}) = \|\mathbf{W}\|_F^2 / \|\mathbf{W}\|_2^2$, which can be rearranged to:

$$\|\mathbf{W}\|_2 = \frac{\|\mathbf{W}\|_F}{\sqrt{\text{srank}(\mathbf{W})}}. \tag{52}$$

Substituting this for both weight matrices:

$$\|\mathbf{J}_{\text{MLP}}\|_2 \leq L_\phi \cdot \frac{\|\mathbf{W}_1\|_F}{\sqrt{\text{srank}(\mathbf{W}_1)}} \cdot \frac{\|\mathbf{W}_2\|_F}{\sqrt{\text{srank}(\mathbf{W}_2)}} = \frac{L_\phi \|\mathbf{W}_1\|_F \|\mathbf{W}_2\|_F}{\sqrt{\text{srank}(\mathbf{W}_1) \cdot \text{srank}(\mathbf{W}_2)}}. \tag{53}$$

Thus low stable ranks and fixed or growing Frobenius norms enlarge this MLP Jacobian upper envelope. They do not alone force the realized Jacobian norm to increase. □

Discussion: Unified View Across Layer Types. For linear layers the stable-rank relation is exact. For attention and MLP blocks, the formulas give stable-rank-dependent upper envelopes. The near-rank-one decomposition, explicit $\sigma_\ell \geq M$, residual control, and alignment assumptions in Theorem 4.2 provide the separate lower-bound ingredients needed for an amplification certificate.

B.6 Proof of Theorem 4.5 (Weight Gradient Norm Lower Bound)

Let $\mathbf{R}_{i+1:L} = \mathbf{R}_L \cdots \mathbf{R}_{i+1}$ and $\mathbf{J}_{i+1:L} = \mathbf{J}_L \cdots \mathbf{J}_{i+1}$. The dominant backpropagated signal is

$$\mathbf{R}_{i+1:L}^T \mathbf{g}_L = \left(\prod_{k=i+1}^{L} \sigma_k \right) \left(\prod_{k=i+1}^{L-1} \mathbf{v}_{k+1}^T \mathbf{u}_k \right) (\mathbf{u}_L^T \mathbf{g}_L) \mathbf{v}_{i+1}. \tag{54}$$

Hence its image under $\mathbf{A}_i^T$ has norm at least

$$\gamma_i (a + \delta) \left(\prod_{k=i+1}^{L} \sigma_k \right) a^{L-i-1} \|\mathbf{g}_L\|_2. \tag{55}$$

The suffix version of the product telescope in Theorem 4.2 gives

$$\|\mathbf{J}_{i+1:L} - \mathbf{R}_{i+1:L}\|_2 \leq \left(\prod_{k=i+1}^{L} \sigma_k \right) \sum_{k=i+1}^{L} \varepsilon_k. \tag{56}$$

Applying $\|\mathbf{A}_i\|_2 \leq \kappa \gamma_i$ to this residual and using the reverse triangle inequality gives the coefficient

$$\gamma_i \left(\prod_{k=i+1}^{L} \sigma_k \right) \left((a + \delta)^{L-i} - \kappa \sum_{k=i+1}^{L} \varepsilon_k \right). \tag{57}$$

Condition (12), $\gamma_i \geq \gamma$, and $\sigma_k \geq M$ reduce this to $\gamma (aM)^{L-i}$. □

Discussion: Gradient Decomposition Interpretation. The three-part decomposition in (10) has a clear interpretation:

- $\frac{\partial L}{\partial \mathbf{h}^{(L)}}$: The **loss gradient** at the final layer output. This is the "signal" that backpropagation aims to transmit.
- $\frac{\partial \mathbf{h}^{(L)}}{\partial \mathbf{h}^{(i)}} = \mathbf{J}^{(i+1:L)}$: The **cumulative Jacobian** is the transmission channel. Theorem 4.2 quantifies its dominant rank-one path, while Proposition B.1 quantifies its transmitted top-r subspace.
- $\frac{\partial \mathbf{h}^{(i)}}{\partial \hat{\mathbf{v}}_{out}^{(i)}}$: The **local parameter map** converts the transmitted hidden-state signal into a parameter gradient.

The chain-rule decomposition itself is rank-independent. In the near-rank-one case, transmission is tracked through one dominant singular direction; in the near-rank-r extension below, it is tracked through an r-dimensional subspace using terminal projection, subspace overlap, and restricted local sensitivity.

Discussion: Interpretation of the conditions. For rank one, the alignment margin measures how strongly the task gradient excites the dominant path, and γ_i measures the local response to that direction. For rank r, τ measures terminal energy in the dominant subspace, a_r controls transmission between adjacent dominant subspaces, and γ is the minimum local gain on the transmitted subspace. In both cases, κ controls the relative amplification of discarded-tail residuals, so the Jacobian approximation error propagates explicitly into the weight-gradient bound.

B.7 Near-Rank-r Extension of the Weight-Gradient Lower Bound

Proposition B.4 (Near-Rank-r Weight-Gradient Lower Bound)**.** *Fix $i < L$ and let $\mathbf{g}_L = \partial L / \partial \mathbf{h}^{(L)}$. In addition to the suffix version of Proposition B.1, assume*

$$\|\mathbf{U}_{L,r}^T \mathbf{g}_L\|_2 \geq \tau \|\mathbf{g}_L\|_2, \qquad \sigma_{\min}(\mathbf{A}_i^T \mathbf{V}_{i+1,r}) \geq \gamma > 0, \qquad \|\mathbf{A}_i\|_2 \leq \kappa\gamma, \tag{58}$$

where $\mathbf{A}_i = \partial \mathbf{h}^{(i)} / \partial \hat{\mathbf{v}}_{out}^{(i)}$. Then

$$\begin{aligned} \left\| \frac{\partial L}{\partial \hat{\mathbf{v}}_{out}^{(i)}} \right\|_2 \geq \gamma \Bigg[& \tau \left(\prod_{k=i+1}^{L} m_k \right) a_r^{L-i-1} \\ & - \kappa \left(\prod_{k=i+1}^{L} s_k \right) \sum_{k=i+1}^{L} \varepsilon_{k,r} \Bigg] \|\mathbf{g}_L\|_2. \end{aligned} \tag{59}$$

Proof. The top-r suffix maps the terminal component $\mathbf{U}_{L,r}^T \mathbf{g}_L$ through the r-dimensional core with minimum gain $(\prod_{k=i+1}^{L} m_k) a_r^{L-i-1}$. Its output lies in $\mathrm{range}(\mathbf{V}_{i+1,r})$, where $\mathbf{A}_i^T$ has minimum gain γ. The suffix telescope contributes operator error at most $(\prod_{k=i+1}^{L} s_k) \sum_{k=i+1}^{L} \varepsilon_{k,r}$; applying $\|\mathbf{A}_i\|_2 \leq \kappa\gamma$ and the reverse triangle inequality gives Equation (59). □

In particular, suppose both the terminal projection and every adjacent top-r overlap have margin $a_r + \delta_r$, and

$$\kappa \left(\prod_{k=i+1}^{L} s_k \right) \sum_{k=i+1}^{L} \varepsilon_{k,r} \leq \left(\prod_{k=i+1}^{L} m_k \right) \left((a_r + \delta_r)^{L-i} - a_r^{L-i} \right). \tag{60}$$

Then the margin absorbs the residual and

$$\left\| \frac{\partial L}{\partial \hat{\mathbf{v}}_{out}^{(i)}} \right\|_2 \geq \gamma \left(\prod_{k=i+1}^{L} m_k \right) a_r^{L-i} \|\mathbf{g}_L\|_2. \tag{61}$$

If $m_k \geq M$, this is at least $\gamma (a_r M)^{L-i} \|\mathbf{g}_L\|_2$.

B.8 Proof of Theorem 4.6 (Total Gradient Norm Lower Bound)

For each layer, summing the squared columnwise bound of Theorem 4.5 gives

$$\left\| \frac{\partial L}{\partial \mathbf{W}^{(i)}} \right\|_F^2 \geq n_w \gamma^2 (aM)^{2(L-i)} \|\mathbf{g}_L\|_2^2 = C(aM)^{2(L-i)}. \tag{62}$$

Summing over i and setting $k = L - i$ yields

$$\sum_{i=1}^{L} \left\| \frac{\partial L}{\partial \mathbf{W}^{(i)}} \right\|_F^2 \geq C \sum_{k=0}^{L-1} (aM)^{2k}. \tag{63}$$

For $(aM)^2 \neq 1$, the finite geometric sum equals $((aM)^{2L} - 1)/((aM)^2 - 1)$; for $aM = 1$, it equals L. Thus $aM > 1$ gives the stated exponential certificate. □

C Positive Feedback Mechanism

Here we state and prove the formal theorems for the self-reinforcing feedback mechanism introduced in Section 4.2.

Theorem C.1 (Low-Rank Propagation in Attention Layers)**.** *Consider an attention layer with query, key, value, and output projection matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O$. If the hidden states $\mathbf{H}^{(\ell-1)}$ and cohidden states $\tilde{\mathbf{H}}^{(\ell)}$ have rank at most r, then the gradients $\nabla_{\mathbf{W}_Q} L, \nabla_{\mathbf{W}_K} L, \nabla_{\mathbf{W}_V} L, \nabla_{\mathbf{W}_O} L$ all have rank at most r.*

Theorem C.2 (Exact Common-Basis Stable Rank Feedback)**.** *Let $\mathbf{W} = \mathbf{U}\,\mathrm{diag}(s_1, \ldots, s_n)\mathbf{V}^T$, where $s_1 > 0$ and $s_1 \geq s_i \geq 0$. Suppose the input–cohidden mixed moment is diagonal in the same singular bases,*

$$\nabla_{\mathbf{W}} L = \mathbb{E}[\tilde{\mathbf{h}}_{out} \mathbf{h}_{in}^T] = \mathbf{U}\,\mathrm{diag}(c_1, \ldots, c_n)\mathbf{V}^T, \tag{64}$$

with $c_i \leq 0$. Define $d_i = -c_i$ and assume

$$s_1 d_i \leq s_i d_1, \qquad i = 1, \ldots, n. \tag{65}$$

Then, for every step size $\eta \geq 0$, the gradient-descent update $\mathbf{W}^+ = \mathbf{W} - \eta \nabla_{\mathbf{W}} L$ satisfies $\mathrm{srank}(\mathbf{W}^+) \leq \mathrm{srank}(\mathbf{W})$. If $\eta > 0$ and the inequality in (65) *is strict for at least one non-leading direction, then the stable rank decreases strictly.*

C.1 Proof of Theorem C.1 (Low-Rank Propagation in Attention Layers)

The key insight is that gradients in neural networks are computed as outer products. For any weight matrix $\mathbf{W}$ in a linear layer $\mathbf{y} = \mathbf{W}\mathbf{x}$, the gradient with respect to $\mathbf{W}$ is:

$$\nabla_{\mathbf{W}} L = \frac{\partial L}{\partial \mathbf{y}} \mathbf{x}^T = \tilde{\mathbf{y}} \mathbf{x}^T, \tag{66}$$

where $\tilde{\mathbf{y}} = \frac{\partial L}{\partial \mathbf{y}}$ is the cohidden state (gradient of loss w.r.t. the output).

This outer product structure implies a fundamental constraint: $\mathrm{rank}(\tilde{\mathbf{y}}\mathbf{x}^T) \leq \min(\mathrm{rank}(\tilde{\mathbf{y}}), \mathrm{rank}(\mathbf{x}^T)) = 1$ for single vectors. When we batch over B samples, the gradient becomes:

$$\nabla_{\mathbf{W}} L = \sum_{b=1}^{B} \tilde{\mathbf{y}}^{(b)} (\mathbf{x}^{(b)})^T = \tilde{\mathbf{Y}}^T \mathbf{X}, \tag{67}$$

where $\tilde{\mathbf{Y}} \in \mathbb{R}^{B \times d_{out}}$ and $\mathbf{X} \in \mathbb{R}^{B \times d_{in}}$. By the rank inequality for matrix products:

$$\mathrm{rank}(\tilde{\mathbf{Y}}^T \mathbf{X}) \leq \min(\mathrm{rank}(\tilde{\mathbf{Y}}^T), \mathrm{rank}(\mathbf{X})) \leq \min(\mathrm{rank}(\tilde{\mathbf{Y}}), \mathrm{rank}(\mathbf{X})). \tag{68}$$

For attention layers specifically, let's trace through each gradient:

Query gradient: The query projection is $\mathbf{Q} = \mathbf{H}\mathbf{W}_Q$. By the chain rule:

$$\nabla_{\mathbf{W}_Q} L = \mathbf{H}^T \frac{\partial L}{\partial \mathbf{Q}}. \tag{69}$$

This is a product of $\mathbf{H}^T \in \mathbb{R}^{d\times n}$ (where n is sequence length) and $\frac{\partial L}{\partial \mathbf{Q}} \in \mathbb{R}^{n\times d_k}$. If $\mathbf{H}$ has rank at most r, then $\mathbf{H}^T$ also has rank at most r, and hence $\nabla_{\mathbf{W}_Q} L$ has rank at most r.

The same argument applies to $\nabla_{\mathbf{W}_K} L$ and $\nabla_{\mathbf{W}_V} L$, since they all have $\mathbf{H}^T$ as the left factor:

$$\nabla_{\mathbf{W}_Q} L = \mathbf{H}^{(\ell-1)T} \frac{\partial L}{\partial \mathbf{Q}^{(\ell-1)}} \quad \Rightarrow \quad \text{rank} \le \text{rank}(\mathbf{H}^{(\ell-1)}) \le r, \tag{70}$$

$$\nabla_{\mathbf{W}_K} L = \mathbf{H}^{(\ell-1)T} \frac{\partial L}{\partial \mathbf{K}^{(\ell-1)}} \quad \Rightarrow \quad \text{rank} \le r, \tag{71}$$

$$\nabla_{\mathbf{W}_V} L = \mathbf{H}^{(\ell-1)T} \frac{\partial L}{\partial \mathbf{V}^{(\ell-1)}} \quad \Rightarrow \quad \text{rank} \le r. \tag{72}$$

Output projection gradient: For $\mathbf{W}_O$, the computation is $\mathbf{y} = (\mathbf{AV})\mathbf{W}_O$. The gradient is:

$$\nabla_{\mathbf{W}_O} L = (\mathbf{AV})^T \frac{\partial L}{\partial \mathbf{y}} = (\text{Attn Output})^T \tilde{\mathbf{H}}^{(\ell)}. \tag{73}$$

If the cohidden states $\tilde{\mathbf{H}}^{(\ell)}$ have rank at most r, then $\nabla_{\mathbf{W}_O} L$ has rank at most r.

This completes the proof: all four attention gradients have rank bounded by $\max(\text{rank}(\mathbf{H}^{(\ell-1)}), \text{rank}(\tilde{\mathbf{H}}^{(\ell)})) \le r$. □

C.2 Proof of Theorem C.2 (Exact Common-Basis Stable Rank Feedback)

The shared-basis assumption gives the exact finite-step update

$$\mathbf{W}^+ = \mathbf{U}\,\text{diag}(s_1 + \eta d_1, \ldots, s_n + \eta d_n)\mathbf{V}^T. \tag{74}$$

All updated diagonal entries are nonnegative. Moreover, condition (65) implies

$$\frac{s_i + \eta d_i}{s_1 + \eta d_1} \le \frac{s_i}{s_1}, \tag{75}$$

because both denominators are positive and cross multiplication reduces the claim to $s_1 d_i \le s_i d_1$. Since $s_i/s_1 \le 1$, Equation (75) also shows that the first updated diagonal entry remains maximal. Consequently,

$$\text{srank}(\mathbf{W}^+) = \sum_{i=1}^{n} \left(\frac{s_i + \eta d_i}{s_1 + \eta d_1}\right)^2 \le \sum_{i=1}^{n} \left(\frac{s_i}{s_1}\right)^2 = \text{srank}(\mathbf{W}). \tag{76}$$

If $\eta > 0$ and the relative-growth inequality is strict for some $i > 1$, the corresponding squared normalized singular value decreases strictly, so the summed inequality is strict. □

Proposition C.3 (First-Order Stable Rank Feedback without Common Bases)**.** *Let $\mathbf{W} \in \mathbb{R}^{m\times n}$ have singular values $s_1 > s_2 \ge \cdots \ge s_q \ge 0$, where $q = \min\{m, n\}$, and corresponding singular vectors $(\mathbf{u}_i, \mathbf{v}_i)$. Let $\mathbf{G} = \mathbb{E}[\tilde{\mathbf{h}}_{out}\mathbf{h}_{in}^T]$ be arbitrary and define only its matched directional coefficients*

$$c_i = \mathbf{u}_i^T \mathbf{G} \mathbf{v}_i. \tag{77}$$

For $\mathbf{W}(\eta) = \mathbf{W} - \eta\mathbf{G}$,

$$\frac{d}{d\eta}\,\text{srank}(\mathbf{W}(\eta))\Big|_{\eta=0} = \frac{2}{s_1^3} \sum_{i=2}^{q} s_i(s_i c_1 - s_1 c_i). \tag{78}$$

Therefore, if

$$\sum_{i=2}^{q} s_i(s_i c_1 - s_1 c_i) < 0, \tag{79}$$

then there exists $\eta_0 > 0$ such that $\text{srank}(\mathbf{W} - \eta\mathbf{G}) < \text{srank}(\mathbf{W})$ for every $0 < \eta < \eta_0$.

Proof. The strict gap $s_1 > s_2$ makes the leading singular value differentiable, with $s_1'(0) = -\mathbf{u}_1^T \mathbf{G} \mathbf{v}_1 = -c_1$. Also,

$$\frac{d}{d\eta} \|\mathbf{W} - \eta \mathbf{G}\|_F^2 \Big|_{\eta=0} = -2\langle \mathbf{W}, \mathbf{G} \rangle_F = -2 \sum_{i=1}^{q} s_i c_i. \tag{80}$$

Differentiating $\|\mathbf{W}(\eta)\|_F^2 / s_1(\eta)^2$ gives Equation (78); its $i = 1$ term cancels identically. A negative derivative implies the stated strict decrease on a sufficiently small right neighborhood of zero. □

This criterion places no restriction on the off-diagonal coefficients $\mathbf{u}_i^T \mathbf{G} \mathbf{v}_j$ for $i \neq j$. A convenient sufficient condition is $s_i c_1 \leq s_1 c_i$ for every $i > 1$, with strict inequality for at least one index with $s_i > 0$; the aggregate condition (79) is weaker because different directions may compensate.

Corollary C.4 (Dominant-Channel Alignment Criterion)**.** *Let* $\alpha_i = \mathbf{u}_i^T \tilde{\mathbf{h}}_{out}$ *and* $\beta_i = \mathbf{v}_i^T \mathbf{h}_{in}$*, and define*

$$A_1 = \mathbb{E}[\alpha_1^2], \quad B_1 = \mathbb{E}[\beta_1^2], \qquad A_\perp = \sum_{i=2}^{q} \mathbb{E}[\alpha_i^2], \quad B_\perp = \sum_{i=2}^{q} \mathbb{E}[\beta_i^2]. \tag{81}$$

Suppose $\operatorname{srank}(\mathbf{W}) > 1$ *and the leading input–cohidden channel has negative alignment*

$$-\mathbb{E}[\alpha_1 \beta_1] \geq \chi \sqrt{A_1 B_1}, \qquad \chi > 0. \tag{82}$$

If

$$\chi \sqrt{A_1 B_1} \sqrt{\operatorname{srank}(\mathbf{W}) - 1} > \sqrt{A_\perp B_\perp}, \tag{83}$$

then condition (79) *holds, and stable rank decreases for all sufficiently small positive steps.*

Proof. Write $\mathbf{s}_\perp = (s_2, \ldots, s_q)$ and $\mathbf{c}_\perp = (c_2, \ldots, c_q)$. Cauchy–Schwarz, first across random variables and then across coordinates, gives

$$\|\mathbf{c}_\perp\|_2^2 \leq \sum_{i=2}^{q} \mathbb{E}[\alpha_i^2] \mathbb{E}[\beta_i^2] \leq A_\perp B_\perp. \tag{84}$$

Using $c_1 \leq -\chi \sqrt{A_1 B_1}$,

$$\sum_{i=2}^{q} s_i (s_i c_1 - s_1 c_i) \leq -\chi \sqrt{A_1 B_1} \|\mathbf{s}_\perp\|_2^2 + s_1 \|\mathbf{s}_\perp\|_2 \|\mathbf{c}_\perp\|_2 \tag{85}$$

$$\leq \|\mathbf{s}_\perp\|_2 \left[-\chi \sqrt{A_1 B_1} \|\mathbf{s}_\perp\|_2 + s_1 \sqrt{A_\perp B_\perp} \right]. \tag{86}$$

Finally, $\|\mathbf{s}_\perp\|_2 / s_1 = \sqrt{\operatorname{srank}(\mathbf{W}) - 1}$, so condition (83) makes the bracket strictly negative. □

Equivalently, if $p_{out} = A_1/(A_1 + A_\perp)$ and $p_{in} = B_1/(B_1 + B_\perp)$, the sufficient condition is

$$\chi \sqrt{\frac{p_{out} p_{in}}{(1 - p_{out})(1 - p_{in})}} > \frac{1}{\sqrt{\operatorname{srank}(\mathbf{W}) - 1}}. \tag{87}$$

Thus stronger leading input–cohidden alignment, greater leading-channel energy concentration, or a broader current singular spectrum can each supply the first-order decrease margin.