SPECTRAL ALIGNMENT FOR EARLY DETECTION OF LOSS EXPLOSION IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Loss explosions can invalidate expensive language-model training runs, while scalar monitors such as weight and gradient norms often provide ambiguous or late warnings. We introduce Spectral Alignment (SA), a scale-robust measure of the signed cosine alignment between layer inputs and a weight matrix's principal left singular direction. Mixed-sign SA permits cancellation in the first-order spectral update. We show that a sufficiently one-sided distribution, together with compatible downstream gradients, produces positive spectral drift, and derive conditions under which a persistent SA warning precedes a spectral-norm threshold. In two language-model failure settings, this directional alignment collapse emerges well before loss explosion and provides a clearer warning than the monitored scalar baselines. SA also identifies checkpoints yielding lower retraining loss than later pre-failure checkpoints. Sampled-layer monitoring and power iteration make its cost configurable, supporting early intervention and recovery.

1 INTRODUCTION

Training a large language model can consume weeks and millions of dollars (Brown et al., 2020; Touvron et al., 2023), making late loss explosions costly. Low-precision computation improves efficiency but can accumulate rounding error and interact adversely with optimization (Micikevicius et al., 2017); high learning rates can also drive divergence (Wortsman et al., 2023). A run may appear healthy while these failures develop. Detecting instability before the loss explodes would let practitioners stop high-risk runs, preserve compute, and identify checkpoints for recovery.

Conventional monitors track gradient and weight norms (Takase et al., 2025; Fishman et al., 2024), stable rank (Sanyal et al., 2020), and activation magnitudes (Rybakov et al., 2024; Wortsman et al., 2023; Kimi Team et al., 2025). Raw magnitudes vary across models and layers, and a rising norm may reflect either healthy feature learning or developing failure. Interpretation often requires model-specific baselines, with clear warnings emerging only after instability becomes critical.

Feature learning couples layer inputs to principal singular directions of their weights (Yang et al., 2022; 2023; Imani et al., 2021). The spectral norm measures maximum permitted amplification; input alignment determines whether that direction is occupied. Mixed signed alignments permit cancellation in the first-order spectral update. A sufficiently one-sided batch removes this cancellation and produces spectral growth under compatible gradients. This distributional change can be detected before a spectral-norm threshold is crossed.

We define *Spectral Alignment* (SA) as this cosine alignment and call a persistent loss of signed diversity *directional alignment collapse*. SA is invariant to positive scalar rescaling of inputs and weights. Our detector tracks persistent one-sided shifts with growing magnitude (Figure 1). Our contributions are:

- We introduce SA, a data-conditioned, scale-robust monitor of input alignment with a layer's most amplifying direction.
- We connect directional alignment collapse to spectral growth under compatible gradients and derive conditions for warnings before a spectral-norm threshold.
- In two language-model failure settings, SA gives earlier, clearer warnings than scalar baselines and selects checkpoints yielding lower retraining loss.

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.

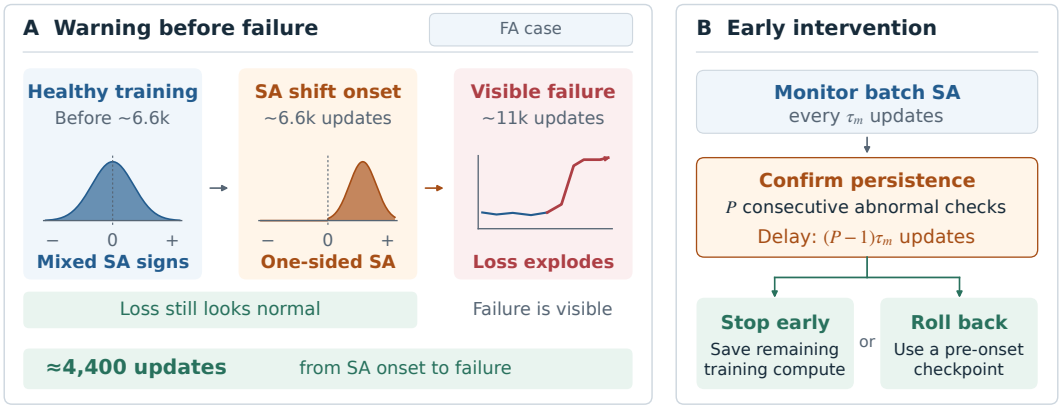


Figure 1: SA as an actionable early-warning monitor. In the FA run, a persistent one-sided SA shift appears near step 6,600 while the loss remains normal, creating an approximately 4,400-step intervention window before failure near step 11,000. Practitioners can stop the run or resume from a healthy checkpoint, reducing wasted computation.

2 Related Work

2.1 Monitoring and Mitigating Training Instability

Activation, attention, and gradient magnitudes can reveal numerical stress or loss spikes (Rybakov et al., 2024; Takase et al., 2025; Wortsman et al., 2023; Fishman et al., 2024). These inexpensive signals depend on the model and training configuration and often reflect accumulated instability, requiring stable references to identify abnormal trajectories.

Other work targets particular causes or modifies optimization. Weight-scaling reparameterization addresses non-uniform parameter norms, while analyses of low-precision FlashAttention trace failures to representation structure and biased rounding errors (Nishida et al., 2024; Qiu & Yao, 2026). Optimizer-level methods instead mitigate abnormal updates through adaptive clipping, momentum reset, or anomaly-aware suppression (Huang et al., 2025; Wang et al., 2025; Kumar et al., 2025). A complementary line improves stability by modifying the training recipe or the Transformer architecture itself, including initialization and warmup choices (Xiong et al., 2020; Huang et al., 2020; Liu et al., 2020), normalization and residual scaling (Wang et al., 2022; Liu et al., 2023; Shleifer et al., 2022), and attention-specific normalization such as Query-Key normalization (Henry et al., 2020; Dehghani et al., 2023). SA complements these mitigation methods: it monitors ordinary linear layers without changing optimization and uses a data-conditioned distributional signal to identify intervention windows and rollback checkpoints.

2.2 Spectral Geometry and Representation Alignment

Training stability is connected to weight and Jacobian spectra. Dynamical isometry controls Jacobian singular values, spectral normalization constrains layer spectral norms, and orthogonal or unitary parametrizations preserve signal magnitude (Pennington et al., 2017; Miyato et al., 2018; Arjovsky et al., 2016). Stable rank summarizes effective matrix rank (Sanyal et al., 2020). These methods characterize parameter or Jacobian geometry without directly measuring current input alignment.

Dominant representation directions can align with prediction targets, and spectral scaling conditions support feature learning at large width (Imani et al., 2021; Yang et al., 2022; 2023). These μP-style constraints prescribe parameter and update scales under width transfer. SA instead monitors signed input alignment with a weight matrix's principal direction. Its signal is a persistent loss of batchwise sign diversity, which, to our knowledge, has not been used for online instability warning or checkpoint selection.

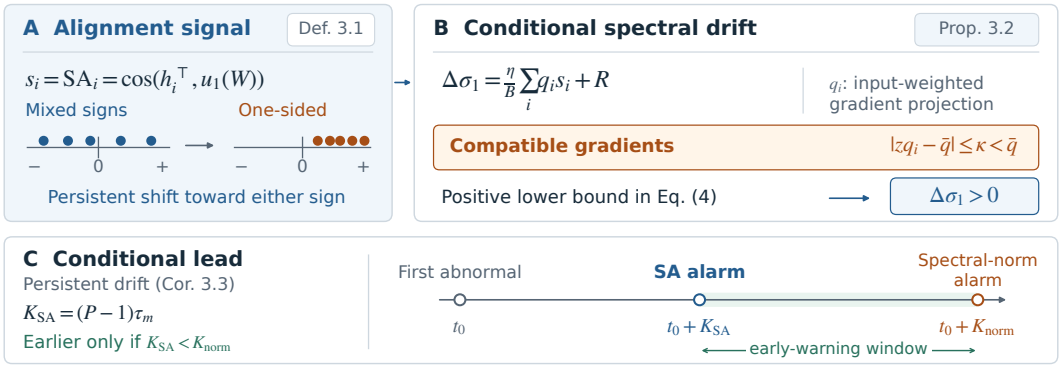


Figure 2: Theoretical roadmap from SA to early warning. Definition 3.1 (Section 3.1) makes amplification data-conditioned; Proposition 3.2 and Corollary 3.3 (Section 3.2) connect one-sided SA to conditional spectral drift and temporal lead.

3 Spectral Alignment for Early Detection

This section defines SA, establishes a conditional mechanism for early warning, and specifies its monitoring rule. Fixing layer l, let $\mathbf{W} = \mathbf{W}_l$, $\mathbf{h}_i = \mathbf{h}^{(l-1,i)}$, and $\mathbf{f}_i = \mathbf{h}_i\mathbf{W}$.

3.1 Definition of Spectral Alignment

Prior spectral analyses connect feature learning to principal singular directions (Imani et al., 2021; Yang et al., 2023). Write the compact SVD as

$$\mathbf{W} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^\top, \quad \mathbf{f}_i = \mathbf{h}_i\mathbf{W} = \sum_{j=1}^{r} \sigma_j \left\langle \mathbf{h}_i^\top, \mathbf{u}_j \right\rangle \mathbf{v}_j^\top, \tag{1}$$

where $r = \mathrm{rank}(\mathbf{W})$ and $\sigma_1 \ge \sigma_2 \ge \cdots$. The spectral norm σ_1 measures maximum permitted amplification, while $\langle \mathbf{h}_i^\top, \mathbf{u}_1 \rangle$ measures whether the current input occupies that direction. We normalize this projection to isolate directional coupling.

Definition 3.1 (Spectral Alignment)**.** Spectral Alignment (SA) of sample i at layer l is the cosine similarity between $\mathbf{h}_i$ and the principal left singular vector of $\mathbf{W}$:

$$\mathrm{SA}_l^{(i)} = \frac{\left\langle \mathbf{h}_i^\top, \mathbf{u}_1(\mathbf{W}) \right\rangle}{\|\mathbf{h}_i\|_2 \, \|\mathbf{u}_1(\mathbf{W})\|_2}. \tag{2}$$

The singular-vector pair $(\mathbf{u}_1, \mathbf{v}_1)$ is defined only up to a joint sign flip. Such a flip reflects the entire SA distribution about zero but does not change whether both signs coexist. The relevant signal is therefore not an individual value or its absolute sign, but the batch distribution $\{\mathrm{SA}_l^{(i)}\}_{i \in S}$. We call a persistent loss of signed diversity *directional alignment collapse*; the detector consequently requires a persistent one-sided shift with growing magnitude rather than treating transient skew as pathological.

3.2 Why Distributional SA Can Provide Early Warning

We next explain why the *distribution* of SA can become informative before scalar norms are visibly abnormal. For a mini-batch of nonzero inputs, let $\boldsymbol{\delta}_i = \partial \ell_i / \partial \mathbf{f}_i$ and $\mathbf{G} = \nabla_{\mathbf{W}} \mathcal{L}_B$. Let $\sigma_1 > \sigma_2$ be the two largest singular values of $\mathbf{W}$, with principal vectors $\mathbf{u}_1, \mathbf{v}_1$, and set $s_i = \mathrm{SA}_i$. Define the effective coefficient $q_i = -\|\mathbf{h}_i\|_2 \langle \boldsymbol{\delta}_i^\top, \mathbf{v}_1 \rangle$, which combines the input magnitude with the downstream-gradient projection; the minus sign comes from gradient descent. For a sufficiently small update $\mathbf{W}^+ = \mathbf{W} - \eta\mathbf{G}$ with learning rate η, singular-value perturbation gives

$$\mathbf{G} = \frac{1}{B}\sum_{i=1}^{B} \mathbf{h}_i^\top \boldsymbol{\delta}_i, \qquad \Delta\sigma_1 = -\eta \mathbf{u}_1^\top \mathbf{G} \mathbf{v}_1 + R = \frac{\eta}{B}\sum_{i=1}^{B} q_i s_i + R. \tag{3}$$

Here, $\Delta\sigma_1 = \sigma_1(\mathbf{W}^+) - \sigma_1(\mathbf{W})$, and R is the second-order remainder of the singular-value perturbation expansion, satisfying $|R| = O(\eta^2\|\mathbf{G}\|_2^2/(\sigma_1 - \sigma_2))$. Thus, $q_i s_i$ determines sample i's first-order contribution to spectral growth, with SA providing the forward directional factor.

Define the signed and dominant/opposite SA masses by $M_\pm = B^{-1}\sum_i [\pm s_i]_+$, with $[u]_+ = \max(u, 0)$, $M_d = \max(M_+, M_-)$, and $M_o = \min(M_+, M_-)$. Each mass is an average *magnitude* over the batch rather than a count of samples, so many weakly aligned samples and few strongly aligned ones can contribute equally. Because the singular-vector pair is defined only up to a joint sign flip, the raw sign of s_i carries no meaning, and we let $z \in \{-1, +1\}$ denote the orientation that maps the majority side to positive values; z is a bookkeeping convention rather than a quantity we monitor, and the statements below are invariant to this flip.

Proposition 3.2 (Dominant-side compatibility and spectral drift)**.** *Suppose the sign-adjusted coefficients zq_i lie within κ of a common positive reference value $\bar{q}$: $|zq_i - \bar{q}| \le \kappa$ for every i, where $0 \le \kappa < \bar{q}$. Here, κ bounds the deviation from $\bar{q}$, and $\kappa < \bar{q}$ ensures that all zq_i are positive and reinforce the dominant SA side. Then*

$$\Delta\sigma_1 \ge \eta\big[(\bar{q} - \kappa)M_d - (\bar{q} + \kappa)M_o\big] - |R|. \tag{4}$$

Hence, compatible downstream gradients and a sufficiently one-sided SA distribution produce positive spectral drift whenever the right-hand side is positive.

Directional alignment collapse removes forward-side cancellation, but growth also requires compatible downstream gradients: reversing their projections reverses the first-order drift without changing SA. A joint singular-vector sign flip changes s_i, q_i, and z while preserving the bound. Proofs and the necessity construction appear in Appendix B and Proposition B.2.

Corollary 3.3 (Temporal lead over a spectral-norm threshold)**.** *SA is evaluated every τ_m updates, so consecutive checkpoints are τ_m updates apart, and t_0 is the first of $P \ge 1$ consecutive abnormal checkpoints; P controls how much persistence is required and is set so that isolated checkpoints do not trigger a warning. The alarm is confirmed only after the persistence delay $K_{\mathrm{SA}} = (P-1)\tau_m$. Write K_{norm} for the number of updates after t_0 at which the spectral norm of the same layer reaches a user-chosen threshold T. Under the compatible-drift conditions in Appendix B, SA fires first whenever $K_{\mathrm{SA}} < K_{\mathrm{norm}}$, and the appendix gives a computable sufficient condition together with bounds on K_{norm}.*

The persistence window must therefore be shorter than the time compatible updates need to produce a detectable norm change.

Scale robustness. For $\alpha, \beta > 0$, $\mathrm{SA}(\beta\mathbf{h}, \alpha\mathbf{W}) = \mathrm{SA}(\mathbf{h}, \mathbf{W})$ under a consistent singular-vector orientation: positive scalar rescaling of inputs or weights leaves SA unchanged. A common sign flip preserves sign diversity and the warning summaries. Proposition B.3 proves this and a stronger homogeneous-ReLU result.

Scope. The analysis assumes a separated leading singular value and small gradient-descent or mini-batch SGD updates. A small spectral gap can make the principal direction unstable. Extension to AdamW requires its preconditioner to approximately preserve the relevant principal-direction projection. Appendix B details these conditions.

3.3 SA Monitoring Rule

Every τ_m updates, we estimate the principal left singular vector of each monitored layer using k power iterations and compute SA on a sampled mini-batch, at cost $O(kn_{l-1}n_l)$ per layer and checkpoint. Because SA tracks a batchwise sign pattern, approximate singular vectors can suffice; too few iterations can still obscure the warning when the dominant direction is poorly estimated. We use $k = 10$; Figure 11 shows that power iteration preserves the SVD-based warning pattern on the monitored GPT-2 layer.

For each SA batch, let $\mu_+ = \mathbb{E}[\mathrm{SA} \mid \mathrm{SA} > 0]$ and $\mu_- = \mathbb{E}[|\mathrm{SA}| \mid \mathrm{SA} < 0]$, and summarize its magnitude and sign diversity by

$$a = \big|{}^1\!/_B \sum_{i=1}^{B} \mathrm{SA}_i\big|, \qquad b = \left|\mathrm{median}_i(\mathrm{SA}_i)\right|, \qquad r = \min(\mu_+, \mu_-)/\max(\mu_+, \mu_-). \tag{5}$$

Detector	Easy to deploy	Monitoring cost	Scale robust	Early warning	Actionable output
Training loss	High	Negligible	Limited	No; confirms visible failure	Late stopping or heuristic rollback
Weight spectral norm	Moderate	Low–medium	No	Inconsistent in our failures	Trend or alarm only
Gradient norm	High	Low	No	Inconsistent in our failures	Alarm only
Maximum activation	Moderate	Low	No	Often late or ambiguous	Alarm only
Stable rank	Moderate	Low–medium	Yes	Inconsistent in our failures	Trend or alarm only
Spectral Alignment	Moderate	Configurable	**Yes**	**Early in evaluated failures**	**Intervention window and checkpoint selection**

Table 1: Comparison of passive training-instability detectors. Ratings assume periodic monitoring of selected layers and summarize the evaluated failures rather than universal guarantees.

We set $r = 0$ if either side is empty. The pair (a, b) measures the magnitude of the one-sided shift, while r distinguishes loss of sign diversity from a broad distribution that merely has a nonzero mean. In our experiments, a checkpoint is abnormal when $a > 0.2$, $b > 0.2$, and $r < 0.1$; the detector fires after P consecutive abnormal checkpoints. Persistence filters transient skew. These empirical thresholds detect a forward-side risk signal; the theory does not make them sufficient conditions for failure. Numerical values and trajectories appear in Tables 3 and 4 and Figure 9.

Monitoring cost is configurable through the selected layers, batch size, frequency, and number of power iterations. Inputs and weights can instead be saved for offline computation at the expense of storage and I/O. In our experiments, a preliminary spectral-norm screen localizes a sensitive module, after which the same module is monitored throughout the run. Other deployment settings may monitor a small fixed layer set or rotate the monitored layers to trade temporal resolution for broader coverage. A run-level signal can be obtained by aggregating across monitored matrices, for example using the largest sign-imbalance score, the top-k average score, or the fraction of monitored matrices that satisfy the warning criterion.

3.4 Practical Comparison with Conventional Detectors

Table 1 summarizes deployment effort, cost, scale robustness, warning behavior, and actionable output. Training loss is nearly free but confirms failure only after it becomes visible. Weight, gradient, and activation summaries are also inexpensive, yet their raw scales vary across models and layers; stable rank is scale robust but does not measure data–weight coupling. All of these methods are passive monitors that leave the objective and optimizer unchanged.

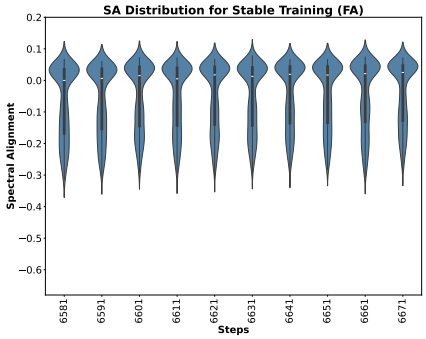
SA adds sampled inputs, principal-vector estimation, and persistence calibration, with cost controlled by the choices in Section 3.3. In return, it is invariant to positive scalar rescaling of layer inputs and weights and provides both an intervention window and a checkpoint-selection signal in our evaluated failures.

4 Experiments

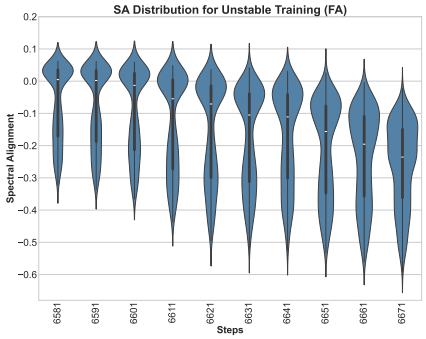
We evaluate SA's warning lead, consistency across two failure settings, and checkpoint-selection utility. These experiments test detection rather than the per-sample gradient compatibility in Proposition 3.2.

4.1 Experimental Settings and Monitoring Protocol

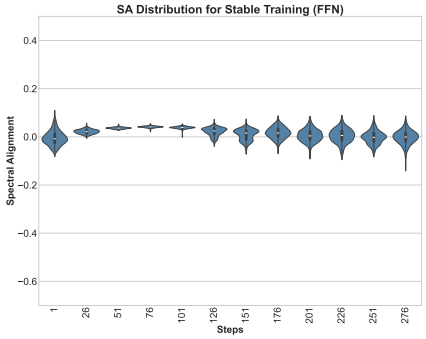
Flash-Attention instability. We reproduce low-precision Flash Attention (FA) instability in a 12-layer, 12-head GPT-2-style model trained on OpenWebText with bfloat16 mixed precision (nanoGPT



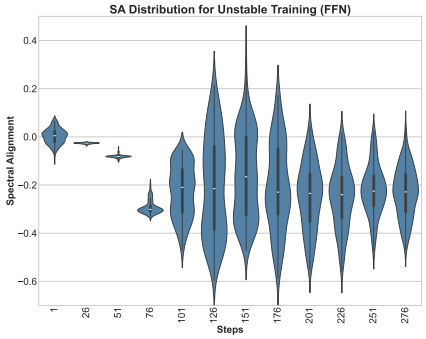
(a) SA distribution for stable training (FA)



(b) SA distribution for unstable training (FA)



(c) SA distribution for stable training (FFN)



(d) SA distribution for unstable training (FFN)

Figure 3: SA distributions across monitored checkpoints in stable and unstable training. Stable runs retain mass on both sides of zero, whereas unstable runs develop a persistent one-sided distribution. The negative direction shown here is determined by the chosen singular-vector orientation; the relevant signal is the loss of signed diversity.

Issue 303, 2023; nanoGPT Issue 524, 2024). The run explodes near step 11,000; vanilla attention or high-precision FA stabilizes training. Following the preliminary spectral-norm screen, we monitor the second attention layer's key projection. Full hyperparameters appear in Appendix A.1.

FFN instability. Following Wortsman et al. (2023), we train Qwen2.5-0.5B on English C4 with learning rates 1×10^{-3} (stable) and 5×10^{-3} (unstable). The latter run explodes near step 647. The preliminary screen selects layer 10's `gate_proj` matrix in the SwiGLU feed-forward network (FFN). Appendix A.1 gives full hyperparameters.

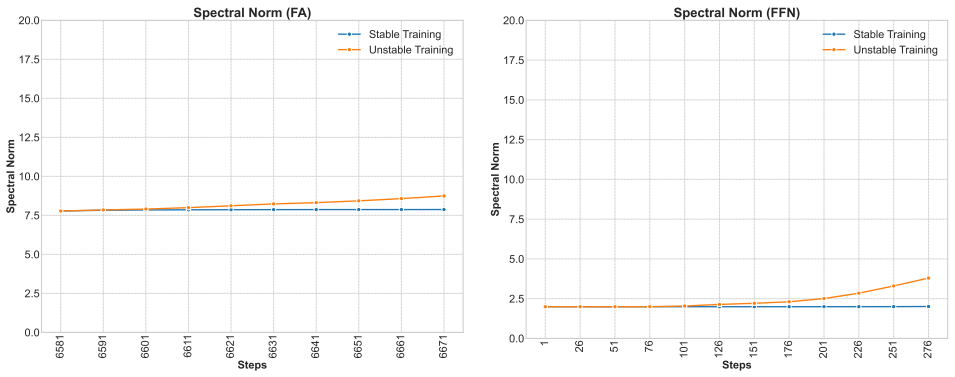
Both settings use the rule in Section 3.3. We report t_0, the first checkpoint in a sustained abnormal window. A P-checkpoint online alarm is confirmed at $t_0 + (P-1)\tau_m$; its lead over failure is shorter than the reported onset-to-failure interval by $(P-1)\tau_m$.

4.2 Directional Alignment Collapse Precedes Visible Instability

Flash Attention. Figures 3(a) and 3(b) show the SA distributions from steps 6,581 to 6,671. The stable run exhibits mild skew at individual checkpoints but retains substantial mass on both sides of zero and does not enter a sustained abnormal window. In the unstable run the distribution instead becomes persistently one-sided and its magnitude continues to grow; under the plotted orientation, the mean falls below -0.2 and the negative tail extends below -0.6. We therefore identify $t_0 \approx 6{,}600$ as the onset of the sustained abnormal window, well before the spectral norm responds and about 4,400 updates before the loss explodes (Table 2).

Setting	Monitored matrix	Abnormal-window onset t_0	Loss explosion	Onset-to-failure interval	Rollback comparison
Flash Attention	Layer-2 key projection	$\sim$ 6,600	$\sim$ 11,000	$\sim$ 4,400 updates	6,580 vs. 7,200
FFN	Layer-10 `gate_proj`	76	647	571 updates	75 vs. 200

Table 2: Observed SA timing and rollback comparisons. The reported t_0 is the first checkpoint in the sustained abnormal window; a P-checkpoint online alarm occurs after the additional persistence delay $(P-1)\tau_m$.



(a) Spectral norm of key weight (FA)

(b) Spectral norm of gate weight (FFN)

Figure 4: Weight spectral norm for stable and unstable training in the FA and FFN settings. Absolute scales and pre-failure trends differ across the two cases.

FFN. Figures 3(c) and 3(d) show the corresponding distributions from steps 1 to 276. The stable run temporarily leans to the positive side, but its magnitude remains bounded and both signs remain represented, so the transient skew is not treated as abnormal. In the unstable run, sign diversity collapses persistently by step 76; under the plotted orientation, the mean and lower tail reach approximately -0.2 and -0.4. Thus $t_0 = 76$ again precedes the loss explosion, which occurs at step 647 (Table 2). Results for other layers, additional scalar summaries, and longer windows are reported in Appendices A.2 and A.5.

Both cases exhibit loss of SA sign diversity before spectral growth and loss explosion, consistent with Proposition 3.2 and Corollary 3.3.

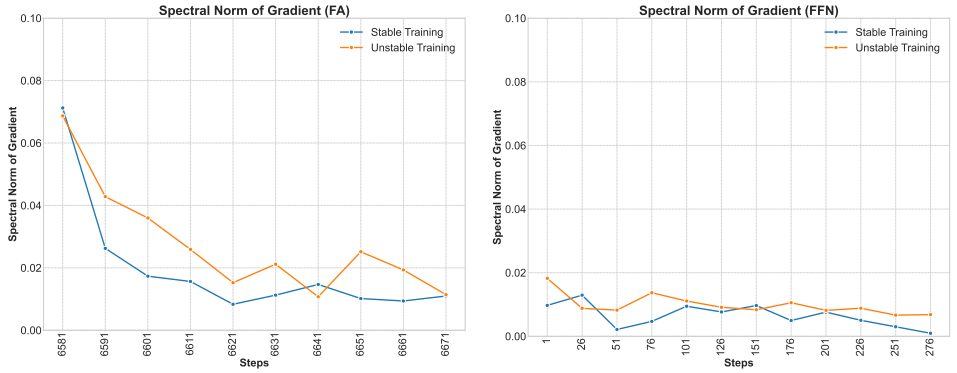
Relation to downstream symptoms. SA measures input alignment with a layer's most amplifying direction. It complements downstream diagnostics, such as logit growth or attention-entropy collapse, by probing the directional coupling that can precede amplification.

4.3 Comparison with Conventional Indicators

Weight spectral norm. The stable and unstable trajectories can be distinguished retrospectively, but neither their scales nor their trends give a common warning pattern (Figure 4): a value near 7.5 is routine in the FA setting, whereas the FFN run fails at a much smaller scale, and the FA trajectory becomes clearly abnormal only after the SA distribution has entered its sustained one-sided regime.

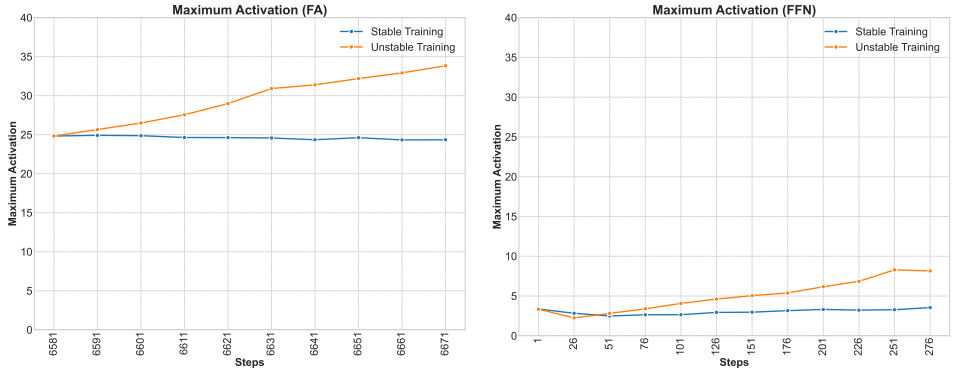
Gradient spectral norm. In FA, gradient spectral norm decreases from roughly 0.08 to 0.01 as instability develops; in FFN, it remains small in both runs (Figure 5). Neither growth nor a shared absolute threshold identifies both failures.

Maximum activation. Maximum activation separates the FA trajectories relatively early, but does not separate the FFN ones, where the curves remain close until later (Figure 6). Its absolute scale is



(a) Gradient spectral norm (FA) (b) Gradient spectral norm (FFN)

Figure 5: Gradient spectral norm for stable and unstable training. It decreases or remains small during the pre-failure periods considered here.



(a) Maximum activation for key value (FA) (b) Maximum activation for gate output (FFN)

Figure 6: Maximum activation for stable and unstable training. Early separation occurs in the FA case but not consistently in the FFN case.

also setting dependent, with values near 20 in stable FA training but much smaller values accompanying FFN failure, and large activations are not inherently pathological (Sun et al., 2024).

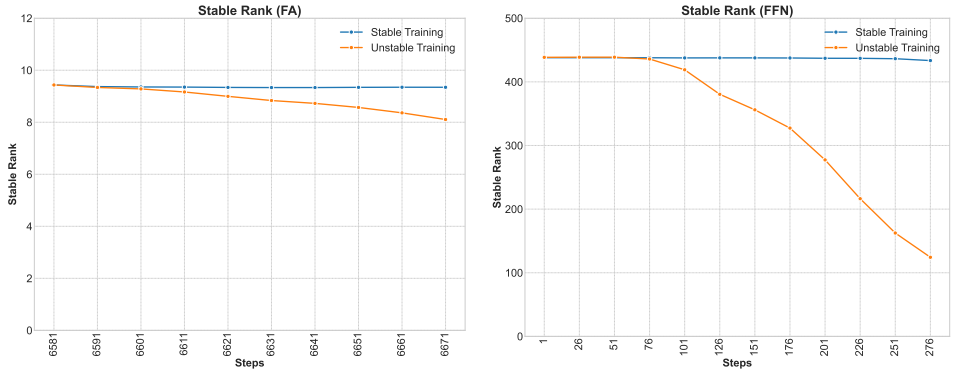
Stable rank. Stable rank is scale robust, but it measures parameter geometry rather than data–weight coupling: both FA runs remain near a low stable rank, whereas the unstable FFN run falls sharply from a much larger initial value while its stable control changes only gradually (Figure 7).

Across both evaluated failures, SA consistently shows a persistent loss of sign diversity with increasing magnitude, while stable controls do not. Conventional indicators lack a common early-warning pattern even after normalization by their initial values (Appendix A.4).

4.4 SA-Guided Checkpoint Selection

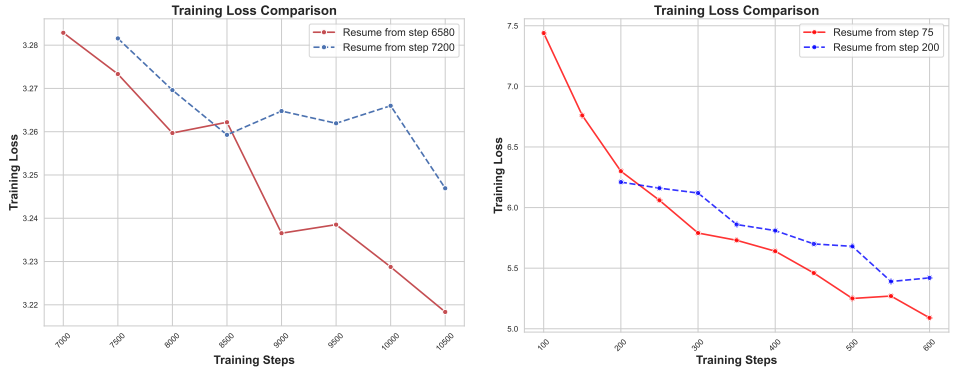
Once an SA warning is confirmed, we select the latest saved checkpoint before the sustained abnormal window. We compare this choice with a later checkpoint that follows alignment collapse but still precedes visible loss explosion. A warning can also prompt denser checkpointing and targeted checks of downstream symptoms, helping practitioners decide whether to stop or roll back.

For FA, we resume with vanilla attention from step 6,580, preceding the abnormal window near step 6,600, and compare against step 7,200. The SA-guided checkpoint yields lower validation loss over the displayed retraining horizon (Figure 8(a)).



(a) Stable rank of key weight (FA) (b) Stable rank of gate weight (FFN)

Figure 7: Weight stable rank for stable and unstable training. Its scale and pre-failure trajectory differ substantially between the two settings.



(a) Retraining on the FA case (b) Retraining on the FFN case

Figure 8: SA-guided checkpoint recovery in the two failure settings. In each case, resuming from the checkpoint immediately before the sustained abnormal SA window yields lower subsequent loss than resuming from a later pre-failure checkpoint.

For FFN, we resume with learning rate 1×10^{-3} from step 75, preceding the abnormal window at step 76, and compare against step 200. The SA-guided checkpoint again yields lower loss (Figure 8(b)), supporting its utility as a rollback boundary in both settings.

5 Conclusion

Spectral Alignment monitors signed input alignment with a layer's principal singular direction. Our analysis connects one-sided SA to spectral growth under compatible gradients. In two language-model failure settings, SA warns before loss explosion and identifies checkpoints yielding lower retraining loss. Its scale invariance and configurable monitoring cost support practical early intervention and recovery.

Limitations The theory analyzes linear-layer updates; extensions covering attention, LayerNorm, and full optimizer dynamics remain open. Our experiments cover two failure settings. Broader evaluation across architectures, datasets, and training recipes is needed to characterize false positives, false negatives, and threshold transfer.

Ethics Statement

This paper presents work whose goal is to improve the reliability and efficiency of language model training. We do not foresee direct negative societal impacts beyond the general risks associated with large-scale language model development.

AI Use Statement

We used generative AI tools solely for language polishing to improve the clarity and readability of the manuscript. No generative AI tools were used for research ideation, theoretical development or proofs, experimental design, method implementation, data processing, or interpretation of results. The authors take full responsibility for the content of this paper.

Reproducibility Statement

We describe the monitoring rule in Section 3.3 and report the full experimental settings, including models, datasets, optimizers, learning rates, and batch sizes, in Appendix A.1. Appendix A.2 lists the per-checkpoint SA summaries on which the warning criterion operates, and Section 4.3 reports the conventional indicators used for comparison.

References

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *ICML*, pp. 1120–1128. PMLR, 2016.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, volume 33, pp. 1877–1901, 2020.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *ICML*, 2023.

Maxim Fishman, Brian Chmiel, Ron Banner, and Daniel Soudry. Scaling FP8 training to trillion-token LLMs. *arXiv preprint*, 2024.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of EMNLP*, 2020.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2501.06842.

Xiao Shi Huang, Felipe Perez, Jimmy Ba, and Maksims Volkovs. Improving transformer optimization through better initialization. In *ICML*, 2020.

Ehsan Imani, Wei Hu, and Martha White. Representation alignment in neural networks. *arXiv preprint*, 2021.

Kimi Team et al. Kimi K2: Open agentic intelligence. *arXiv preprint*, 2025.

Abhay Kumar, Louis Owen, Nilabhra Roy Chowdhury, and Fabian Güra. ZClip: Adaptive spike mitigation for LLM pre-training. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2504.02507.

Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. Understanding the difficulty of training transformers. In *EMNLP*, 2020.

Yijin Liu, Xianfeng Zeng, Fandong Meng, and Jie Zhou. BranchNorm: Robustly scaling extremely deep transformers. *arXiv preprint*, 2023.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint*, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint*, 2018.

nanoGPT Issue 303. nanoGPT issue #303: Gradient explosion when training with bfloat16, 2023. URL https://github.com/karpathy/nanoGPT/issues/303. Accessed 2025-09-07.

nanoGPT Issue 524. nanoGPT issue #524: Training loss becomes NaN when using bfloat16, 2024. URL https://github.com/karpathy/nanoGPT/issues/524. Accessed 2025-09-07.

Kosuke Nishida, Kyosuke Nishida, and Kuniko Saito. Initialization of large language models via reparameterization to mitigate loss spikes. In *EMNLP*, 2024. doi: 10.18653/v1/2024.emnlp-main.1264. URL https://aclanthology.org/2024.emnlp-main.1264/.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *NeurIPS*, 2017.

Haiquan Qiu and Quanming Yao. Why low-precision transformer training fails: An analysis on Flash Attention. In *ICLR*, 2026. URL https://arxiv.org/abs/2510.04212.

Oleg Rybakov, Mike Chrzanowski, Peter Dykas, Jinze Xue, and Ben Lanir. Methods of improving LLM training stability. *arXiv preprint*, 2024.

Amartya Sanyal, Philip H. Torr, and Puneet K. Dokania. Stable rank normalization for improved generalization in neural networks and GANs. In *ICLR*, 2020.

Sam Shleifer, Jason Weston, and Myle Ott. NormFormer: Improved transformer pretraining with extra normalization. In *ICLR*, 2022.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *COLM*, 2024.

Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike No More: Stabilizing the pre-training of large language models, 2025. URL https://openreview.net/forum?id=G84F1h2IiD.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint*, 2023.

Guoxia Wang, Shuai Li, Congliang Chen, Jinle Zeng, Jiabin Yang, Dianhai Yu, Yanjun Ma, and Li Shen. AdaGC: Enhancing LLM pretraining stability via adaptive gradient clipping. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2502.11034.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, and Furu Wei. DeepNet: Scaling transformers to 1,000 layers. *arXiv preprint*, 2022.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint*, 2023.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *ICML*, 2020.

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. *arXiv preprint*, 2022.

Greg Yang, James B. Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint*, 2023.

A Detailed Experimental Results

A.1 Experimental Details

For the FA experiment, we use a 12-layer, 12-head GPT-2-style decoder-only Transformer with embedding dimension 768 and context length 1024. Dropout is disabled, bias terms are removed from LayerNorm and linear layers, and training uses bfloat16 mixed precision. We train on OpenWebText with AdamW, maximum learning rate 1×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, no weight decay, a cosine schedule with 2000 warmup steps, and gradient clipping at 1.0. Training is run on 4 NVIDIA A100 GPUs (80GB), with per-GPU micro-batch size 32 and 4 gradient accumulation steps.

For the FFN experiment, we train Qwen2.5-0.5B with context length 128 and bfloat16 mixed precision on the English portion of C4. We use AdamW with weight decay 0.1. Stable and unstable runs use learning rates 1×10^{-3} and 5×10^{-3}, respectively. The effective batch size is 128, using per-device batch size 16 and 8 gradient accumulation steps, implemented with Huggingface Accelerator.

A.2 Additional Quantitative Results for Spectral Alignment

This section provides the numerical SA summaries used by the warning criterion in Section 3.3. The results support the distinction made in the main text: stable runs can show mild temporary skew, while unstable runs show persistent one-sided collapse together with increasing SA magnitude. The reported mean and median are the signed batch averages whose absolute values define a and b in Section 3.3, and the positive and negative conditional means define $\mu_\pm$ and hence the ratio r.

Table 3: SA distributional metrics for the GPT-2-style FA experiment, measured on layer 1 attention q projection.

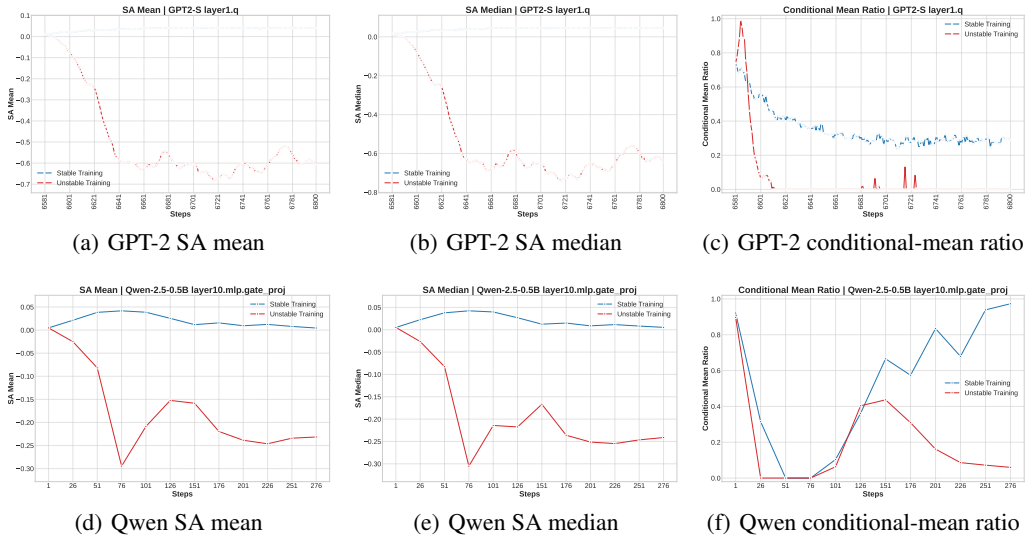
Metric	6581	6591	6601	6611	6621	6631	6641	6651	6661	6671
Stable Training										
sa_mean	0.0121	0.0169	0.0218	0.0274	0.0296	0.0321	0.0338	0.0358	0.0395	0.0392
sa_median	0.0158	0.0205	0.0261	0.0315	0.0337	0.0370	0.0387	0.0411	0.0450	0.0445
sa_positive_mean	0.0322	0.0328	0.0351	0.0365	0.0376	0.0403	0.0408	0.0425	0.0447	0.0447
sa_negative_mean	-0.0238	-0.0211	-0.0194	-0.0169	-0.0150	-0.0155	-0.0150	-0.0140	-0.0138	-0.0135
Unstable Training										
sa_mean	0.0121	-0.0119	-0.0754	-0.1772	-0.2420	-0.4478	-0.5851	-0.5991	-0.6239	-0.6097
sa_median	0.0157	-0.0062	-0.0818	-0.2046	-0.2581	-0.4862	-0.6278	-0.6460	-0.6797	-0.6556
sa_positive_mean	0.0322	0.0231	0.0050	0.0037	0	0	0	0	0	0
sa_negative_mean	-0.0239	-0.0411	-0.0785	-0.1773	-0.2420	-0.4478	-0.5851	-0.5991	-0.6239	-0.6097

Table 4: SA distributional metrics for the Qwen2.5-0.5B FFN experiment, measured on layer 10 `gate_proj`.

Metric	1	26	51	76	101	126	151	176	201	226	251	276
Stable Training												
sa_mean	0.0048	0.0213	0.0384	0.0417	0.0387	0.0254	0.0118	0.0156	0.0094	0.0123	0.0079	0.0045
sa_median	0.0051	0.0224	0.0379	0.0423	0.0398	0.0268	0.0125	0.0149	0.0087	0.0115	0.0082	0.0053
sa_positive_mean	0.0342	0.0308	0.0384	0.0417	0.0405	0.0396	0.0317	0.0345	0.0296	0.0321	0.0294	0.0287
sa_negative_mean	-0.0315	-0.0097	0	0	-0.0042	-0.0143	-0.0211	-0.0198	-0.0247	-0.0218	-0.0276	-0.0295
Unstable Training												
sa_mean	0.0047	-0.0251	-0.0818	-0.2946	-0.2087	-0.1523	-0.1584	-0.2195	-0.2386	-0.2462	-0.2341	-0.2315
sa_median	0.0052	-0.0264	-0.0827	-0.3052	-0.2142	-0.2175	-0.1668	-0.2361	-0.2514	-0.2548	-0.2465	-0.2412
sa_positive_mean	0.0348	0	0	0	0.0163	0.1246	0.1385	0.0942	0.0467	0.0245	0.0198	0.0161
sa_negative_mean	-0.0312	-0.0251	-0.0818	-0.2946	-0.2584	-0.3089	-0.3172	-0.3048	-0.2891	-0.2843	-0.2746	-0.2694

A.3 Numerical Results of Other Indicators

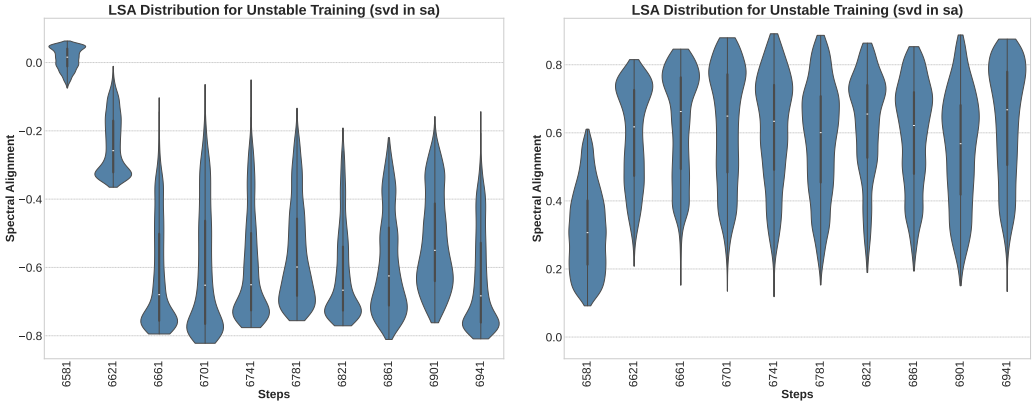
We provide the detailed numerical values for other indicators in the two experiments in Table 5 and Table 6.



(a) GPT-2 SA mean (b) GPT-2 SA median (c) GPT-2 conditional-mean ratio

(d) Qwen SA mean (e) Qwen SA median (f) Qwen conditional-mean ratio

Figure 9: Scalar summaries derived from SA distributions. In the unstable runs, mean/median SA move into the warning regime and the conditional-mean ratio collapses persistently, while the stable runs remain bounded or only show transient skew.

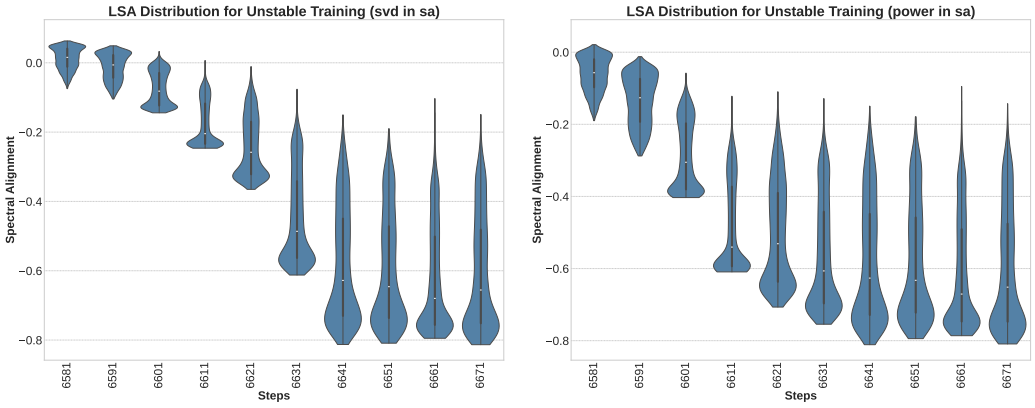


(a) Attention q projection (b) Attention k projection

Figure 10: Longer-window SA distributions for the GPT-2-style FA experiment. The longer window is consistent with the short pre-failure window used in the main text.

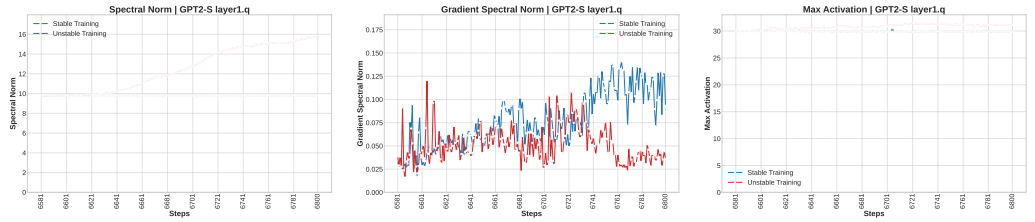
Table 5: Detailed numerical values for metrics in the FA Experiment.

Step	Stable Rank		Spectral Norm		Grad. Spectral Norm		Max Activation	
	Stable	Unstable	Stable	Unstable	Stable	Unstable	Stable	Unstable
6581	9.43	9.43	7.78	7.78	0.071	0.068	24.83	24.83
6601	9.36	9.28	7.85	7.90	0.017	0.035	24.87	26.49
6621	9.34	8.99	7.86	8.11	0.008	0.015	24.62	28.98
6641	9.33	8.73	7.87	8.31	0.015	0.011	24.35	31.39
6661	9.35	8.36	7.87	8.57	0.009	0.019	24.33	32.92

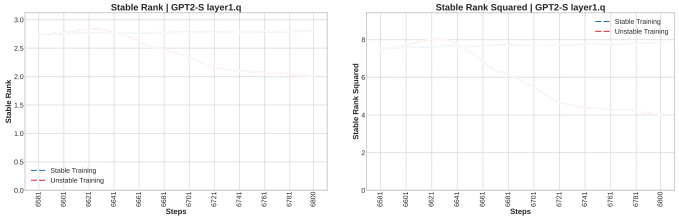


(a) SVD-based SA (b) Power-iteration-based SA

Figure 11: Comparison between SVD-based and power-iteration-based SA on the same GPT-2-style FA layer. The warning pattern is preserved under the cheaper power-iteration approximation.



(a) Spectral norm (b) Gradient spectral norm (c) Maximum activation



(d) Stable rank (e) Squared stable rank

Figure 12: Additional scalar baselines for the monitored GPT-2-style FA layer over the same window as Figure 9. These plots provide the corresponding conventional-metric context for the SA summaries.

Table 6: Detailed numerical values for metrics in the FFN Experiment.

Step	Stable Rank		Spectral Norm		Grad. Spectral Norm		Max Activation	
	Stable	Unstable	Stable	Unstable	Stable	Unstable	Stable	Unstable
1	437.93	438.54	1.99	1.99	0.013	0.013	3.37	3.35
51	437.91	438.75	1.99	1.99	0.005	0.0032	2.49	2.81
101	437.61	418.93	1.99	2.04	0.012	0.0061	2.65	4.07
151	437.63	356.03	1.99	2.21	0.013	0.0033	2.97	5.05
201	437.01	277.35	2.00	2.51	0.011	0.0032	3.32	6.17
251	436.36	162.45	2.00	3.30	0.006	0.0017	3.28	8.29

A.4 Normalized Metrics of Other Indicators

In this section, we evaluate the normalized metrics (i.e., metrics divided by their initial values) for the targeted layers across both stable and unstable training runs. We track the Stable Rank, Spectral

Norm, Gradient Spectral Norm, and Maximum Activation. The quantitative results for the FA and FFN experiments are presented in Table 7 and Table 8, respectively. Our analysis indicates that none of these metrics exhibit a consistent threshold capable of distinguishing between stable and unstable trajectories. This lack of separability is expected; standard parameter initialization schemes constrain these metrics to similar initial magnitudes, rendering their relative evolution insufficient for reliable early detection of training instability.

Table 7: Normalized metrics evolution for the FA experiment. The metrics for stable and unstable runs remain indistinguishable during the transition phase.

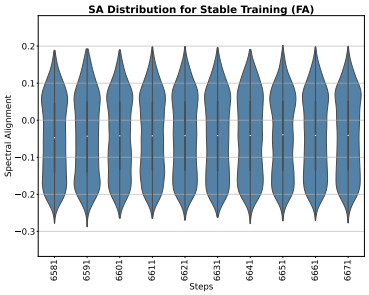
Step	**Stable Rank**		**Spectral Norm**		**Grad. Spectral Norm**		**Max Activation**	
	Stable	Unstable	Stable	Unstable	Stable	Unstable	Stable	Unstable
6581	0.05	0.05	7.04	7.04	11.15	10.75	10.18	10.18
6601	0.05	0.05	7.10	7.15	2.71	5.63	10.20	10.86
6621	0.05	0.05	7.11	7.34	1.30	2.39	10.10	11.88
6641	0.05	0.05	7.12	7.52	2.30	1.68	9.98	12.87
6661	0.05	0.04	7.12	7.75	1.47	3.03	9.97	13.50

Table 8: Normalized metrics evolution for the FFN experiment. Despite divergence in later steps, early metrics do not provide a clear distinguishable boundary.

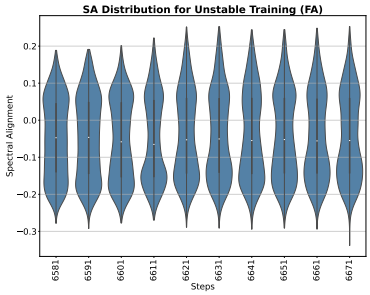
Step	**Stable Rank**		**Spectral Norm**		**Grad. Spectral Norm**		**Max Activation**	
	Stable	Unstable	Stable	Unstable	Stable	Unstable	Stable	Unstable
1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
51	1.00	1.00	1.00	1.00	0.22	0.45	0.74	0.84
101	1.00	0.96	1.00	1.02	0.97	0.61	0.79	1.22
151	1.00	0.81	1.00	1.11	1.00	0.46	0.88	1.51
201	1.00	0.63	1.00	1.26	0.78	0.45	0.98	1.84
251	1.00	0.357	1.00	1.65	0.351	0.357	0.97	2.48

A.5 SA Results of Other Layers

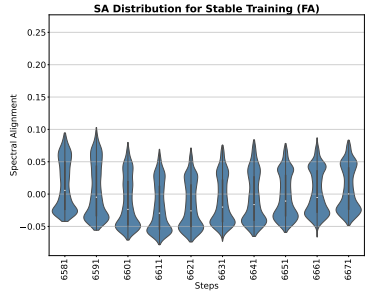
We also sample some normal layers from the FA case and some normal and abnormal layers from the FFN case to show their SA distributions. In Figure 13, we ramdomly sample other normal layers from the FA experiment. These layers have centered SA distributions for both stable and unstable training, indicating healthy sign diversity. In Figure 14, we sample two normal layers and one abnormal layers from the FFN experiment. The normal layers have centered SA distributions for both stable and unstable training. In contrast, the abnormal layers have collapsed SA distributions for unstable training. These results are in expectation as the normal layers have diverse sign patterns, while the abnormal layers have collapsed sign patterns.



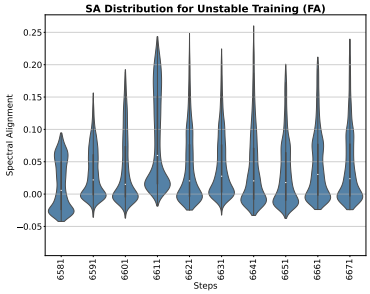
(a) Normal layer in the stable case (attention k_proj in layer 0)



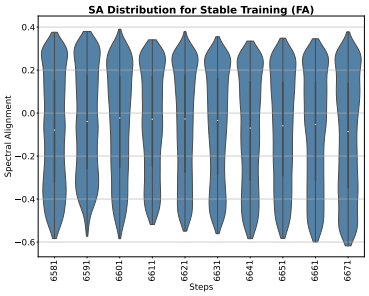
(b) Normal layer in the unstable case (attention k_proj in layer 0)



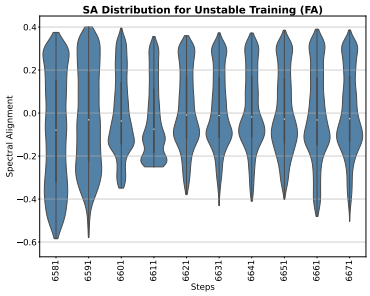
(c) Normal layer in the stable case (attention q_proj in layer 2)



(d) Normal layer in the unstable case (attention q_proj in layer 2)

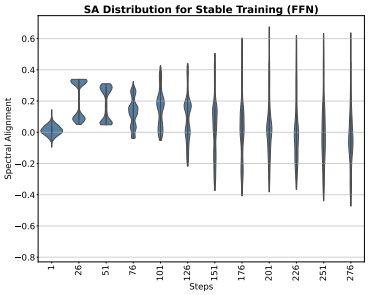


(e) Normal layer in the stable case (attention k_proj in layer 11)

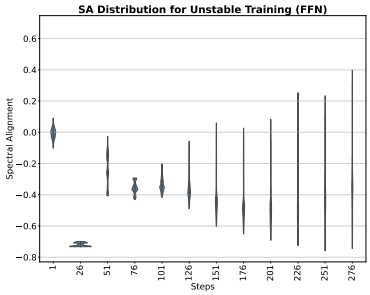


(f) Normal layer in the unstable case (attention k_proj in layer 11)

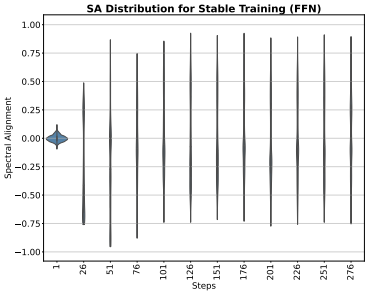
Figure 13: SA distributions of other layers in FA experiment.



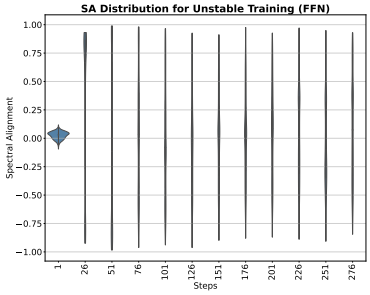
(a) Normal layer in the stable case (mlp gate_proj in layer5)



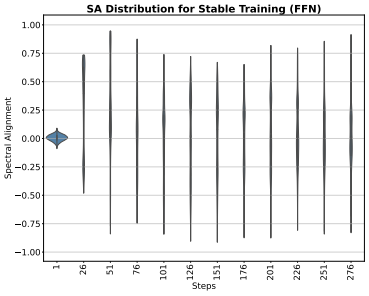
(b) Abnormal layer in the unstable case (mlp gate_proj in layer5)



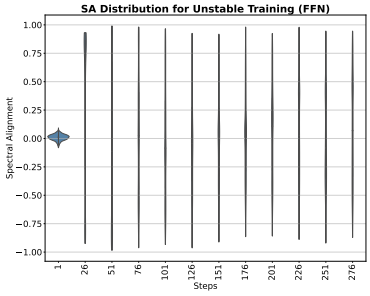
(c) Normal layer in the stable case (mlp gate_proj in layer0)



(d) Normal layer in the unstable case (mlp gate_proj in layer0)



(e) Normal layer in the stable case (mlp up_proj in layer0)



(f) Normal layer in the unstable case (mlp up_proj in layer0)

Figure 14: SA distributions of other layers in FFN experiment.

B Proofs and Additional Analysis

The preceding analysis establishes that one-sided spectral alignment can promote weight spectral norm growth. We now explain three properties that make the *distribution* of SA useful as a detector: its direct coupling to a mini-batch update, the cancellation that is lost when the distribution becomes one-sided, and its invariance to arbitrary positive rescaling. We also formalize why the distributional change can precede a detectable change in the spectral norm. The results below concern directional alignment collapse; a one-sided projection alone does not imply that the full representation has lost rank or all discriminative information. We reuse the notation of Section 3.2 (s_i, q_i, z, $\bar{q}$, κ, $M_\pm$) and recall it below where needed.

Consider a linear layer with weight $\mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$ and a mini-batch of nonzero row-vector inputs $\{\mathbf{h}_i\}_{i=1}^B$. Let $\mathbf{f}_i = \mathbf{h}_i\mathbf{W}$ and $\boldsymbol{\delta}_i = \partial \ell_i / \partial \mathbf{f}_i$, where ℓ_i is the loss on sample i and η is the learning rate. For the average mini-batch loss $\mathcal{L}_B = B^{-1}\sum_i \ell_i$, define

$$\mathbf{G} := \nabla_{\mathbf{W}} \mathcal{L}_B = \frac{1}{B}\sum\nolimits_{i=1}^{B} \mathbf{h}_i^\top \boldsymbol{\delta}_i. \tag{6}$$

Let $\sigma_1 > \sigma_2$ be the two largest singular values of $\mathbf{W}$, with corresponding principal singular vectors $\mathbf{u}_1$ and $\mathbf{v}_1$. We write

$$s_i := \mathrm{SA}_i = \frac{\langle \mathbf{h}_i^\top, \mathbf{u}_1 \rangle}{\|\mathbf{h}_i\|_2} \in [-1, 1], \tag{7}$$

where $\|\mathbf{u}_1\|_2 = 1$.

Proposition B.1 (Mini-batch spectral update decomposition)**.** *After one gradient-descent update $\mathbf{W}^+ = \mathbf{W} - \eta\mathbf{G}$ with learning rate η, assume that $\eta\|\mathbf{G}\|_2 \le c_0(\sigma_1 - \sigma_2)$ for a sufficiently small constant $c_0 < 1$, i.e., the update is small relative to the spectral gap $\sigma_1 - \sigma_2$. Then the leading singular value satisfies*

$$\sigma_1(\mathbf{W}^+) - \sigma_1(\mathbf{W}) = -\frac{\eta}{B}\sum\nolimits_{i=1}^{B} \|\mathbf{h}_i\|_2 s_i \langle \boldsymbol{\delta}_i^\top, \mathbf{v}_1 \rangle + R, \tag{8}$$

where, for a sufficiently small update,

$$|R| = O\left(\frac{\eta^2 \|\mathbf{G}\|_2^2}{\sigma_1 - \sigma_2}\right). \tag{9}$$

Thus, SA is one of the two sample-wise factors that determine the first-order update along the most amplifying direction of the layer. Equivalently, with $q_i = -\|\mathbf{h}_i\|_2 \langle \boldsymbol{\delta}_i^\top, \mathbf{v}_1 \rangle$ as in Section 3.2, (8) *reads $\Delta\sigma_1 = \frac{\eta}{B}\sum_i q_i s_i + R$, the form used in the main text.*

Proof. Because σ_1 is simple, its directional derivative at $\mathbf{W}$ in a direction $\mathbf{E}$, written $\mathrm{D}\sigma_1(\mathbf{W})[\mathbf{E}]$, is $\mathrm{D}\sigma_1(\mathbf{W})[\mathbf{E}] = \mathbf{u}_1^\top \mathbf{E} \mathbf{v}_1$. Applying this identity to $\mathbf{E} = -\eta\mathbf{G}$ gives

$$\sigma_1(\mathbf{W}^+) - \sigma_1(\mathbf{W}) = -\eta \mathbf{u}_1^\top \mathbf{G} \mathbf{v}_1 + R. \tag{10}$$

Under the stated small-update condition, the standard second-order singular-value perturbation bound gives (9); the denominator is the spectral gap that keeps the principal singular directions locally identifiable. Substituting (6) into (10) yields

$$\mathbf{u}_1^\top \mathbf{G} \mathbf{v}_1 = \frac{1}{B}\sum\nolimits_{i=1}^{B} \langle \mathbf{h}_i^\top, \mathbf{u}_1 \rangle \langle \boldsymbol{\delta}_i^\top, \mathbf{v}_1 \rangle \tag{11}$$

$$= \frac{1}{B}\sum\nolimits_{i=1}^{B} \|\mathbf{h}_i\|_2 s_i \langle \boldsymbol{\delta}_i^\top, \mathbf{v}_1 \rangle, \tag{12}$$

which proves (8). □

Proposition B.1 separates the observed forward factor s_i from the backpropagated factor. We now prove the compatibility result stated in the main text.

Proof of Proposition 3.2. Equation (3) gives $\Delta\sigma_1 = \eta B^{-1}\sum_i q_i s_i + R$. Let $\widetilde{q}_i = z q_i$ and $\widetilde{s}_i = z s_i$. Their product is unchanged, $q_i s_i = \widetilde{q}_i \widetilde{s}_i$, and the compatibility condition in Proposition 3.2 lets us write $\widetilde{q}_i = \bar{q} + e_i$, where $|e_i| \le \kappa$. By the choice of z,

$$\frac{1}{B}\sum\nolimits_i \widetilde{s}_i = M_d - M_o, \qquad \frac{1}{B}\sum\nolimits_i |\widetilde{s}_i| = M_d + M_o. \tag{13}$$

Therefore,

$$\frac{1}{B}\sum\nolimits_i q_i s_i = \bar{q}(M_d - M_o) + \frac{1}{B}\sum\nolimits_i e_i \widetilde{s}_i \tag{14}$$

$$\geq \bar{q}(M_d - M_o) - \kappa(M_d + M_o) \tag{15}$$

$$= (\bar{q} - \kappa)M_d - (\bar{q} + \kappa)M_o. \tag{16}$$

Combining this inequality with (3) proves (4). The triangle inequality also gives the useful complementary bound

$$|\Delta\sigma_1| \leq \eta\big[\bar{q}(M_d - M_o) + \kappa(M_d + M_o)\big] + |R|. \tag{17}$$

In particular, the lower bound in (4) is positive when $M_d > 0$,

$$\frac{M_o}{M_d} < \frac{\bar{q} - \kappa}{\bar{q} + \kappa}, \qquad \eta\big[(\bar{q} - \kappa)M_d - (\bar{q} + \kappa)M_o\big] > |R|. \tag{18}$$

□

Proposition B.2 (Necessity of downstream compatibility)**.** *Fix* $\mathbf{W}$ *and a nonzero one-sided SA batch: for some* $z \in \{-1, +1\}$*,* $zs_i \geq 0$ *for every* i *and* $S := B^{-1}\sum_i \|\mathbf{h}_i\|_2 z s_i > 0$*. Two collections of downstream gradients can have identical sample-wise norms and the same SA distribution but induce first-order changes* $+\eta S$ *and* $-\eta S$*. Hence, one-sided SA alone cannot determine whether the spectral norm grows or shrinks.*

Proof. Choose $\boldsymbol{\delta}_i^{(+)} = -z\mathbf{v}_1^\top$ and $\boldsymbol{\delta}_i^{(-)} = z\mathbf{v}_1^\top$. Both have unit norm and leave $\mathbf{W}$, the inputs, and all SA values unchanged, while the first-order term in (8) is respectively $+\eta S$ and $-\eta S$. □

The ratio

$$C_{\mathrm{SA}} := \frac{|M_+ - M_-|}{M_+ + M_-} \in [0, 1] \tag{19}$$

is a convenient theoretical measure of directional coherence: it is near zero for balanced signed alignment mass and near one for a one-sided distribution. We set $C_{\mathrm{SA}} = 0$ when $M_+ + M_- = 0$. It is unchanged if the SVD routine returns the jointly sign-flipped pair $(-\mathbf{u}_1, -\mathbf{v}_1)$. It measures the observable loss of cancellation but does not encode the unobserved dominant-side compatibility in Proposition 3.2. To connect these masses to the summaries in Section 3.3, let $p_+ = B^{-1}\sum_i \mathbb{1}(s_i > 0)$ and $p_- = B^{-1}\sum_i \mathbb{1}(s_i < 0)$. For nonempty sides, let $\mu_+ = \mathbb{E}[s_i \mid s_i > 0]$ and $\mu_- = \mathbb{E}[|s_i| \mid s_i < 0]$. Then

$$M_+ = p_+\mu_+, \qquad M_- = p_-\mu_-, \qquad \frac{M_-}{M_+} = \frac{p_-}{p_+}\frac{\mu_-}{\mu_+}. \tag{20}$$

Thus, the theoretical condition depends jointly on sign mass and conditional magnitude. The empirical mean, median, conditional-mean ratio, and persistence criterion are observable proxies for this sustained distributional shift, but are not algebraically equivalent to C_{SA}; their numerical cutoffs are not universal consequences of Proposition 3.2.

Temporal lead assumptions. Write $\mathbf{W}_t$ for the weights after t updates and $\sigma_{1,t} = \sigma_1(\mathbf{W}_t)$, with $q_{i,t}$, $s_{i,t}$, and R_t defined analogously at that step. For Corollary 3.3, assume that from t_0 onward the conditions of Proposition 3.2 hold and $0 < \gamma \leq B^{-1}\sum_i q_{i,t}s_{i,t} \leq \Gamma$, with $|R_t| \leq c_R\eta^2$; here γ, Γ, and c_R are constants of the analysis, bounding the per-step drift from below and above and the remainder term respectively. We also require $\eta\gamma > c_R\eta^2$, and we let $T > \sigma_{1,t_0}$ denote the user-chosen spectral-norm threshold of Section 3.2.

Proof of Corollary 3.3 and full delay bound. Let K_{norm} be the first delay for which $\sigma_{1,t_0+K_{\mathrm{norm}}} \geq T$. Every update lies between $\eta\gamma - c_R\eta^2$ and $\eta\Gamma + c_R\eta^2$. Summing for K consecutive steps gives

$$K(\eta\gamma - c_R\eta^2) \leq \sigma_{1,t_0+K} - \sigma_{1,t_0} \leq K(\eta\Gamma + c_R\eta^2). \tag{21}$$

Consequently, the norm-threshold delay satisfies the full two-sided bound

$$\left\lceil \frac{T - \sigma_{1,t_0}}{\eta\Gamma + c_R\eta^2} \right\rceil \leq K_{\mathrm{norm}} \leq \left\lceil \frac{T - \sigma_{1,t_0}}{\eta\gamma - c_R\eta^2} \right\rceil. \tag{22}$$

The lower and upper bounds on the first threshold-crossing time follow immediately after substituting the threshold gap $T - \sigma_{1,t_0}$. By the checkpoint assumption, $K_{\mathrm{SA}} = (P-1)\tau_m$. Comparing this delay with the lower bound gives the sufficient condition $(P-1)\tau_m < \lceil (T - \sigma_{1,t_0})/(\eta\Gamma + c_R\eta^2) \rceil$. □

The preceding results explain why SA can reveal the onset of an amplification mechanism before its scalar consequences become large. The next proposition explains why its scale is also easier to interpret across layers.

Proposition B.3 (Scale and reparameterization invariance)**.** *For any nonzero* $\mathbf{h}$*, any weight matrix* $\mathbf{W}$ *with a simple leading singular value, and scalars* $\alpha, \beta > 0$*, a consistent orientation of the principal singular vector gives*

$$\mathrm{SA}(\beta\mathbf{h}, \alpha\mathbf{W}) = \mathrm{SA}(\mathbf{h}, \mathbf{W}). \tag{23}$$

Without fixing this orientation, all SA values may instead undergo a common sign flip, which leaves their sign diversity and C_{SA} *unchanged. Moreover, consider two consecutive bias-free ReLU layers. The function-preserving reparameterization*

$$\widetilde{\mathbf{W}}_l = c\mathbf{W}_l, \qquad \widetilde{\mathbf{W}}_{l+1} = c^{-1}\mathbf{W}_{l+1}, \qquad c > 0, \tag{24}$$

leaves the SA values at both layers unchanged, while it rescales their weight, activation, and gradient norms. Hence, unlike these raw magnitudes, SA is invariant to a family of parameterizations that represent the same function.

Proof. Positive multiplication of $\mathbf{W}$ changes its singular values but not its singular vectors, and positive multiplication of $\mathbf{h}$ changes its norm and inner product with $\mathbf{u}_1$ by the same factor. This proves (23).

For the second claim, positive homogeneity of ReLU gives

$$\widetilde{\mathbf{h}}^{(l)} = \mathrm{ReLU}(\mathbf{h}^{(l-1)} c\mathbf{W}_l) = c\mathbf{h}^{(l)}. \tag{25}$$

Consequently, $\widetilde{\mathbf{h}}^{(l)}\widetilde{\mathbf{W}}_{l+1} = \mathbf{h}^{(l)}\mathbf{W}_{l+1}$, so the network function is unchanged. Equation (23) shows that SA is unchanged at layer $l+1$, and scaling $\mathbf{W}_l$ leaves SA unchanged at layer l. In contrast,

$$\|\widetilde{\mathbf{W}}_l\|_2 = c\|\mathbf{W}_l\|_2, \qquad \|\widetilde{\mathbf{W}}_{l+1}\|_2 = c^{-1}\|\mathbf{W}_{l+1}\|_2, \tag{26}$$

$$\|\widetilde{\mathbf{h}}^{(l)}\|_2 = c\|\mathbf{h}^{(l)}\|_2. \tag{27}$$

Backpropagation through the reparameterized pair also gives $\nabla_{\widetilde{\mathbf{W}}_l}\mathcal{L} = c^{-1}\nabla_{\mathbf{W}_l}\mathcal{L}$ and $\nabla_{\widetilde{\mathbf{W}}_{l+1}}\mathcal{L} = c\nabla_{\mathbf{W}_{l+1}}\mathcal{L}$. Thus, c can change all of these raw magnitudes without changing the computed function or SA. □

Scope. Proposition 3.2 is conditional on dominant-side compatibility; Proposition B.2 shows that the same one-sided SA distribution can otherwise accompany either sign of spectral drift. Because q_i contains both the input norm and gradient projection, this sufficient condition requires common sign and limited variation that cannot be inferred from SA alone. SA therefore warns that forward-side cancellation has been lost; it does not certify spectral growth or loss explosion.

When $\sigma_1 - \sigma_2$ is small, the principal direction can rotate sharply; the leading singular subspace is then a more stable object. The update results also describe gradient descent (or mini-batch SGD), while attention, normalization, optimizer state, and other nonlinear interactions may alter the dynamics. Our experiments use AdamW. Writing its update with preconditioned step $\mathbf{P}_t$ and weight decay λ as $\mathbf{W}^+ = (1-\eta\lambda)\mathbf{W} - \eta\mathbf{P}_t$, first-order perturbation gives

$$\Delta\sigma_1 = -\eta\mathbf{u}_1^\top\mathbf{P}_t\mathbf{v}_1 - \eta\lambda\sigma_1 + R_t. \tag{28}$$

The cancellation result extends only when the preconditioner approximately preserves the relevant principal-direction projection. Stable rank is also scale-invariant but, unlike SA, does not measure data–weight coupling. Spectral growth permits stronger amplification but does not guarantee realized activation or gradient growth. Finally, Proposition B.3 is limited to positive rescaling and bias-free homogeneous ReLU pairs; it does not imply a universal SA threshold across architectures.