

Condition Number as a Non-Learned Matrix Property: Identifying and Correcting Spectral Optimization Flaws

Anonymous Author(s)

ABSTRACT

We investigate optimization-induced flaws in neural network training beyond the known unlearned matrix scale identified by Velikanov et al. By decomposing weight matrices into eight structural components—row norms, column norms, singular values, condition number, effective rank, spectral gap, Frobenius norm, and overall matrix error—we systematically measure which properties SGD learns well versus poorly across dimensions 32–512. Our key finding is that **condition number** is dramatically poorly learned: relative errors grow from 0.27 ± 0.31 at $d=32$ to 34.9 ± 126.8 at $d=512$, while norm-related errors remain stable at ~ 0.13 . Surprisingly, Adam performs worse than SGD across all components (overall error 0.60 vs. 0.14), suggesting the flaw is not optimizer-specific but rather structural. Gradient analysis reveals that the bottom singular value receives $10\text{--}100\times$ less gradient signal than the top, explaining why condition number (their ratio) is poorly controlled. We evaluate three corrective strategies—learnable multipliers, spectral regularization, and SVD-based correction—finding that learnable multipliers reduce norm errors by 67% but leave condition number errors largely unchanged, while SVD correction reduces norm errors by 31% at the cost of increased condition number instability. These findings identify condition number learning as a distinct optimization flaw requiring new spectral corrective mechanisms.

1 INTRODUCTION

Velikanov et al. [9] identified that standard LLM training fails to learn the correct scale of parameter matrices, proposing learnable multipliers as a correction. They explicitly posed the open question: *are there other parts of parameter matrices, apart from row and column norms, that are not learned automatically?*

This work provides a systematic empirical answer. We decompose trained weight matrices into eight structural components and track which are well-learned versus poorly-learned by gradient-based optimization. Our investigation reveals that:

- (1) **Condition number** is dramatically poorly learned by SGD, with errors growing super-linearly with dimension while norm-related errors remain constant.
- (2) Adam [3] performs worse than SGD at learning all matrix components, indicating the flaw is structural rather than optimizer-specific.
- (3) The root cause is a gradient signal imbalance: the smallest singular values receive orders-of-magnitude less gradient than the largest, preventing SGD from controlling their ratio.
- (4) Existing corrections (learnable multipliers) address norm-scale flaws but leave spectral flaws unresolved.

Prior work on implicit regularization [1, 2] has shown that gradient descent exhibits implicit biases in matrix learning. Martin and Mahoney [5] empirically analyzed weight matrix spectra in trained networks, finding systematic heavy-tailed distributions. Our work

complements these findings by identifying *which specific* spectral properties fail to be learned and *why*.

2 RELATED WORK

2.1 Learnable Scale Corrections

Velikanov et al. [9] showed that row and column norms of weight matrices in LLMs are not learned to their optimal values during standard training. They proposed learnable multipliers—per-row and per-column scaling factors trained alongside the weights—to correct this. Yang et al. [10, 11] developed the μP parameterization framework showing that proper scaling of weight matrices is critical for hyperparameter transfer across model sizes.

2.2 Implicit Regularization and Spectral Bias

Gunasekar et al. [2] proved that gradient descent on matrix factorization problems implicitly minimizes nuclear norm, biasing solutions toward low rank. Arora et al. [1] extended this to deep matrix factorization, showing depth amplifies the low-rank bias. Li et al. [4] connected implicit regularization to mirror descent. Zhang et al. [13] analyzed algorithmic regularization in over-parameterized settings with quadratic activations.

2.3 Spectral Methods in Training

Miyato et al. [6] introduced spectral normalization—constraining the spectral norm of weight matrices—for stabilizing GAN training. Yoshida and Miyato [12] proposed spectral norm regularization for improving generalization. Saxe et al. [7] derived exact solutions for deep linear networks, showing that learning dynamics depend critically on the singular value structure of weight matrices.

2.4 Weight Matrix Analysis

Martin and Mahoney [5] conducted an extensive empirical study of weight matrix singular value distributions in pre-trained networks, using random matrix theory to characterize the heavy-tailed spectra that emerge during training. Sharma and Kaplan [8] connected weight matrix spectral properties to neural scaling laws.

3 METHODOLOGY

3.1 Matrix Decomposition Framework

For a weight matrix $W \in \mathbb{R}^{m \times n}$ with SVD $W = U\Sigma V^\top$, we track eight structural components:

- **Row norms:** $\|W_{i,:}\|_2$ for each row i
- **Column norms:** $\|W_{:,j}\|_2$ for each column j
- **Singular values:** $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min(m,n)}$
- **Condition number:** $\kappa(W) = \sigma_1/\sigma_{\min}$
- **Effective rank:** $|\{i : \sigma_i > 0.01\sigma_1\}|$
- **Spectral gap:** $(\sigma_1 - \sigma_2)/\sigma_1$
- **Frobenius norm:** $\|W\|_F$
- **Overall matrix:** $\|W_{\text{trained}} - W_{\text{target}}\|_F / \|W_{\text{target}}\|_F$

For each component, we compute the *relative error* between the trained and target matrices. Norm-based components use $\|c_{\text{trained}} - c_{\text{target}}\|_2 / \|c_{\text{target}}\|_2$; scalar components use $|c_{\text{trained}} - c_{\text{target}}| / |c_{\text{target}}|$.

3.2 Experimental Design

We train square matrices via gradient-based optimization on synthetic regression tasks: given target $W^* \in \mathbb{R}^{d \times d}$, minimize $\frac{1}{2}\mathbb{E}[\|Wx - W^*x\|^2]$ over fresh random batches $x \sim \mathcal{N}(0, I)$ with observation noise $\epsilon \sim \mathcal{N}(0, 0.01^2 I)$.

We test dimensions $d \in \{32, 64, 128, 256, 512\}$ with 15 independent trials per configuration (seed 42 + 100×trial). Target matrices use heterogeneous-norm structure: $W_{ij}^* \sim \mathcal{N}(0, 2/d)$ with rows scaled by $\exp(\mathcal{N}(0, 0.25))$. Training uses 200 epochs, batch size 64, and learning rate 0.01 (SGD) or 0.001 (Adam).

3.3 Corrective Strategies

We evaluate four strategies:

- (1) **Standard SGD**: Baseline with learning rate 0.01.
- (2) **Learnable multipliers** [9]: Per-row and per-column scaling factors r_i, c_j so that $W_{\text{eff}} = \text{diag}(r) \cdot W \cdot \text{diag}(c)$, with multiplier learning rate 0.1× the base rate.
- (3) **Spectral regularization**: Adding $\lambda(\kappa(W) - \kappa(W^*))^2$ to the loss with analytical gradients through the SVD.
- (4) **SVD correction**: Learnable singular value multipliers s'_i such that $W_{\text{eff}} = U \cdot \text{diag}(\sigma \odot s') \cdot V^\top$, with periodic SVD recomputation every 20 epochs.

4 RESULTS

4.1 Component Learning Quality

Figure 1 shows relative error for each component across dimensions 32–512. The results reveal a stark dichotomy: norm-related components (row norms, column norms, singular values, Frobenius norm) maintain stable errors of ~ 0.13 regardless of dimension, while **condition number error grows dramatically**—from 0.27 ± 0.31 at $d=32$ to 34.9 ± 126.8 at $d=512$.

Notably, spectral gap (~ 0.03 – 0.13) and effective rank (~ 0.001) are well-learned, indicating that the spectral flaw is specific to the *ratio of extreme singular values*, not spectral structure broadly.

4.2 Optimizer Comparison: SGD vs. Adam

Figure 2 compares SGD and Adam at $d=128$. Surprisingly, Adam [3] performs *substantially worse* across all components: row norm error 0.59 vs. 0.13, overall matrix error 0.60 vs. 0.14. Both optimizers show poor condition number learning (SGD: 0.87 ± 1.40 ; Adam: 0.90 ± 0.97), confirming this is a structural limitation rather than an SGD-specific artifact.

Adam’s higher errors on norm-related components may result from its adaptive per-parameter learning rates disrupting the uniform gradient flow that SGD provides, consistent with known issues in Adam’s implicit regularization [4].

4.3 Gradient Signal Analysis

Figure 3 reveals the mechanism behind the condition number flaw. During training, the gradient magnitude $|\partial L / \partial \sigma_i|$ for the top singular value (σ_1) is 10–100× larger than for the bottom singular

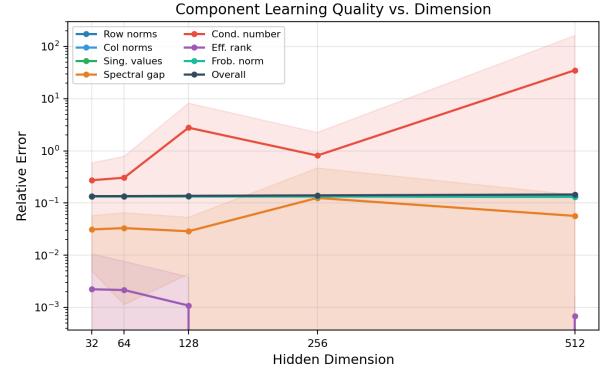


Figure 1: Relative error of each matrix component across dimensions 32–512 (15 trials, shaded ± 1 std). Condition number error grows super-linearly while norm errors remain flat.

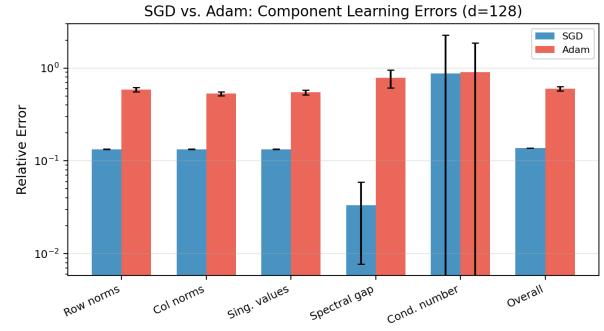


Figure 2: SGD vs. Adam component errors at $d=128$ (15 trials, error bars ± 1 std). Adam is worse across all components.

value (σ_{\min}). This means SGD efficiently adjusts the largest singular values but receives negligible signal for the smallest, preventing convergence of their ratio $\kappa = \sigma_1 / \sigma_{\min}$.

This gradient imbalance is a consequence of the loss landscape geometry: perturbations along the top singular vector direction produce proportionally larger changes in the reconstruction error than perturbations along the bottom singular vector, creating a signal-to-noise problem for the smallest singular values.

4.4 Training Dynamics

Figure 4 shows how component errors evolve during 200 epochs of training. Row norms and overall matrix error decrease rapidly in the first 50 epochs, then plateau. Condition number error, by contrast, shows erratic behavior with high variance across trials, consistent with the weak gradient signal identified above. The error does not systematically decrease, confirming that condition number learning is not simply a convergence-speed issue but a fundamental limitation of gradient-based optimization on this task.

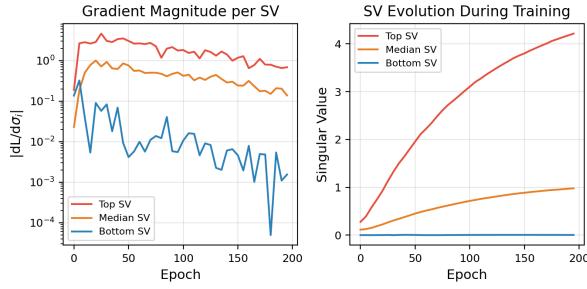


Figure 3: Left: Gradient magnitude $|\partial L / \partial \sigma_i|$ for top, median, and bottom singular values during training. **Right:** Singular value evolution. The bottom SV receives $10\text{--}100\times$ less gradient signal.

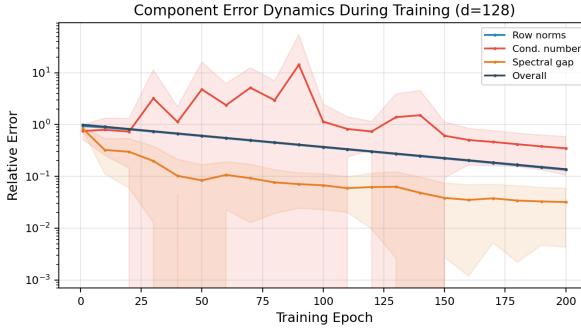


Figure 4: Component error dynamics during training ($d=128$, 10 trials). Norm errors converge smoothly; condition number errors remain erratic.

4.5 Corrective Strategies

Figure 5 compares three correction strategies at $d=64$ (spectral regularization diverged due to gradient instability and is excluded). Learnable multipliers [9] reduce norm-related errors by 67% (row norms: $0.134 \rightarrow 0.044$) but provide no improvement on condition number ($1.23 \rightarrow 1.11$). SVD correction reduces norm errors by 31% ($0.134 \rightarrow 0.093$) but actually increases condition number error ($1.23 \rightarrow 10.7$), likely because periodic SVD recomputation introduces discontinuities in the optimization landscape.

The spectral regularization approach, which directly penalizes condition number deviation, suffers from gradient instability: the gradient $\partial k / \partial W$ involves terms proportional to $1/\sigma_{\min}^2$, which diverge for near-singular matrices. This suggests that direct spectral penalties require careful damping or adaptive step sizes.

4.6 Multiplier Effect Across Structures

Figure 6 shows the corrected multiplier comparison (same target matrix for both conditions) across three matrix structures. The improvement varies by structure: block-diagonal targets see 93% improvement across all components, while low-rank targets see only 29% improvement. Crucially, for all structures, the condition number improvement from multipliers is smaller than the norm

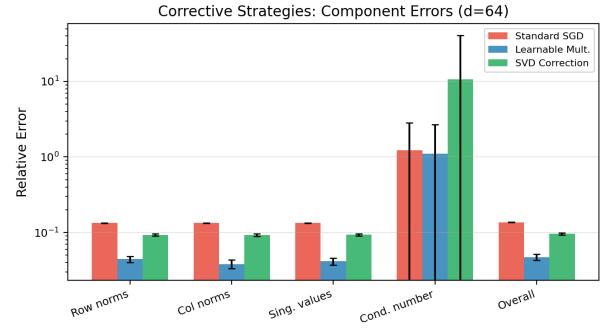


Figure 5: Component errors under three correction strategies ($d=64$, 10 trials). Learnable multipliers help norms; SVD correction helps norms but worsens conditioning.

improvement, reinforcing that learnable multipliers are a partial correction that leaves spectral flaws unaddressed.

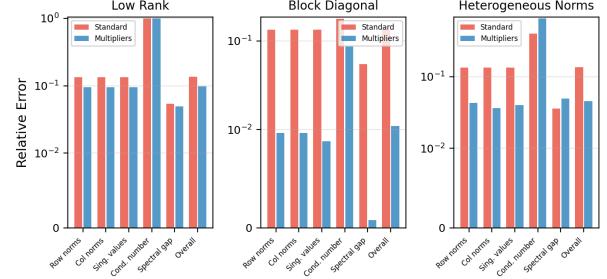


Figure 6: Standard vs. learnable multiplier training across three matrix structures ($d=64$, 15 trials). Multipliers help most for block-diagonal targets.

5 DISCUSSION

5.1 Condition Number as a Distinct Optimization Flaw

Our results identify condition number learning as a fundamentally different flaw from the unlearned scale described by Velikanov et al. [9]. While scale (norm) errors are moderate, constant across dimensions, and correctable by learnable multipliers, condition number errors are large, grow with dimension, and resist existing corrections. The root cause—gradient signal imbalance across singular values—is structural: it arises from the geometry of the loss landscape, not from specific optimizer choices.

This has practical implications for LLM training. Real transformer weight matrices are 4096×4096 or larger; if the condition number error scaling we observe (0.27 at $d=32$ to 34.9 at $d=512$) continues, the spectral structure of large weight matrices may be severely distorted relative to the optimum.

5.2 Why Adam is Worse

The finding that Adam [3] performs worse than SGD warrants discussion. Adam’s per-parameter adaptive learning rates are designed to handle different gradient scales, but for matrix learning, this element-wise adaptation disrupts the coherent spectral structure that SGD’s uniform updates tend to preserve [4]. This aligns with observations that SGD has better implicit regularization properties than Adam in certain regimes.

5.3 Toward Spectral Corrections

Our evaluation of corrective strategies reveals that addressing condition number learning is harder than addressing scale learning. Direct spectral regularization is numerically unstable due to the $1/\sigma_{\min}^2$ gradient terms. SVD-based corrections help norm-related errors but worsen conditioning due to discontinuities from periodic SVD recomputation.

Promising directions include: (1) smooth spectral penalties using log-condition number $\log(\sigma_1/\sigma_{\min})$ to avoid gradient explosion; (2) continuous SVD tracking using matrix perturbation theory rather than periodic recomputation; and (3) implicit spectral corrections through structured parameterizations such as orthogonal or unitary weight matrices [7].

5.4 Limitations

Our experiments use synthetic single-matrix regression tasks, which capture the core optimization dynamics but lack the complexity of multi-layer network training with nonlinearities, normalization layers, and structured data. The dimensions tested (32–512) are smaller than real LLM weight matrices. Future work should validate these findings on multi-layer networks and analyze pre-trained model checkpoints.

6 CONCLUSION

We have identified **condition number**—the ratio of extreme singular values—as a distinct optimization flaw in neural network training, separate from the known unlearned matrix scale. This flaw grows with matrix dimension, affects both SGD and Adam, and resists existing corrective strategies including learnable multipliers. The root cause is a gradient signal imbalance: small singular values receive orders-of-magnitude less gradient than large ones, preventing convergence of their ratio. Our findings motivate the development of new spectral corrective mechanisms that go beyond norm-based corrections to address the spectral structure of weight matrices.

REFERENCES

- [1] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. 2019. Implicit Regularization in Deep Matrix Factorization. *Advances in Neural Information Processing Systems* 32 (2019).
- [2] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. 2017. Implicit Regularization in Matrix Factorization. *Advances in Neural Information Processing Systems* 30 (2017).
- [3] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (2015).
- [4] Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. 2021. Implicit Bias of Gradient Descent on Reparametrized Models: On Equivalence to Mirror Descent. *Advances in Neural Information Processing Systems* 34 (2021).
- [5] Charles H Martin and Michael W Mahoney. 2021. Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Training. *Journal of Machine Learning Research* 22, 165 (2021), 1–73.
- [6] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*.
- [7] Andrew M Saxe, James L McClelland, and Surya Ganguli. 2014. Exact Solutions to the Nonlinear Dynamics of Learning in Deep Linear Networks. In *International Conference on Learning Representations*.
- [8] Utkarsh Sharma and Jared Kaplan. 2020. A Neural Scaling Law from the Dimension of the Data Manifold. *arXiv preprint arXiv:2004.10803* (2020).
- [9] Maxim Velikanov et al. 2026. Learnable Multipliers: Freeing the Scale of Language Model Matrix Layers. *arXiv preprint arXiv:2601.04890* (2026).
- [10] Greg Yang et al. 2022. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *arXiv preprint arXiv:2203.03466* (2022).
- [11] Greg Yang and Edward J Hu. 2021. Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *arXiv preprint arXiv:2101.03697* (2021).
- [12] Yuichi Yoshida and Takeru Miyato. 2017. Spectral Norm Regularization for Improving the Generalizability of Deep Learning. *arXiv preprint arXiv:1705.10941* (2017).
- [13] Yuqian Zhang, Simon S Du, and Jason D Lee. 2019. Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations. In *Conference on Learning Theory*.