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, we systematically measure which properties gradient-based optimization 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 errors remain stable at ∼0.13. This flaw *amplifies with network depth*: in deep linear networks, product condition number error escalates from 0.24× (single layer) to 653× (2-layer) to 13,167× (3-layer). Adam performs worse than SGD across all components (overall error 0.60 vs. 0.14), and weight decay provides no relief. Gradient analysis reveals the root cause: bottom singular values receive 10–100× less gradient signal than top ones. Among four corrective strategies, spectral regularization achieves the largest condition number improvement (77%), while learnable multipliers best correct norms (67%), revealing a fundamental norm–spectral trade-off. Extended training over 1000 epochs shows that while norm errors reach 6.6×10^{-5}, condition number errors plateau at 0.017—a 250× gap that *widens* with training. These findings identify condition number learning as a distinct, depth-amplified optimization flaw.

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 through nine experiments spanning single-layer regression, multi-layer networks, hyperparameter ablations, and extended training. We decompose trained weight matrices into eight structural components and track which are well-learned versus poorly-learned. Our investigation reveals that:

(1) **Condition number** is dramatically poorly learned, with errors growing super-linearly with dimension while norm errors remain constant.
(2) The flaw **amplifies with depth**: deep linear networks exhibit condition number errors that grow exponentially from 0.24× (1 layer) to 13,167× (3 layers), and 2-layer ReLU MLPs show 9.1× error.
(3) Adam [3] performs *worse* than SGD, weight decay has negligible effect, and the flaw persists across all learning rates—confirming it is structural.
(4) The root cause is a gradient signal imbalance: bottom singular values receive orders-of-magnitude less gradient than top ones.
(5) Spectral regularization reduces condition number error by 77%, while learnable multipliers reduce norm errors by 67%, revealing distinct correction mechanisms for norm versus spectral flaws.
(6) Extended training (1000 epochs) drives norm errors to near-zero (6.6×10^{-5}) while condition number errors plateau at 0.017, with the gap *widening* over time.

2 RELATED WORK

2.1 Learnable Scale Corrections

Velikanov et al. [9] showed that row and column norms of LLM weight matrices are not learned to optimal values, proposing learnable per-row and per-column multipliers. Yang et al. [10, 11] developed the μP framework showing proper weight matrix scaling 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 implicitly minimizes nuclear norm. 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.

2.3 Spectral Methods in Training

Miyato et al. [6] introduced spectral normalization for GAN training stability. Yoshida and Miyato [12] proposed spectral norm regularization for generalization. Saxe et al. [7] derived exact solutions for deep linear networks, showing learning dynamics depend critically on singular value structure.

2.4 Weight Matrix Analysis

Martin and Mahoney [5] studied weight matrix singular value distributions using random matrix theory. Sharma and Kaplan [8] connected 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/Column norms**: $\|W_{i,:}\|_2$, $\|W_{:,j}\|_2$
- **Singular values**: $\sigma_1 \geq \cdots \geq \sigma_{\min(m,n)}$
- **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 relative error between trained and target values.

3.2 Experimental Design

We train matrices via gradient-based optimization on synthetic regression: given target $W^* \in \mathbb{R}^{d\times d}$, minimize $\frac{1}{2}\mathbb{E}[\|Wx - W^*x\|^2]$ over random batches $x \sim \mathcal{N}(0, I)$ with noise $\epsilon \sim \mathcal{N}(0, 0.01^2 I)$. We test $d \in \{32, 64, 128, 256, 512\}$ with 10–15 independent trials per configuration (seed $42 + 100k$). Target matrices use heterogeneous-norm structure with rows scaled by $\exp(\mathcal{N}(0, 0.25))$. Default: 200 epochs, batch size 64, LR 0.01 (SGD) or 0.001 (Adam).

3.3 Multi-Layer Networks

To test whether spectral flaws persist beyond single-matrix regression, we train deep linear networks ($y = W_L \cdots W_1 x$) of depth 2 and 3, and 2-layer ReLU MLPs ($y = W_2\,\mathrm{ReLU}(W_1 x)$) in a teacher-student setup. For deep linear networks, we track both the product $\prod W_i$ versus W^* and individual layer condition numbers.

3.4 Corrective Strategies

We evaluate four strategies:

(1) **Standard SGD**: Baseline (LR 0.01).
(2) **Learnable multipliers** [9]: Per-row/column scaling $W_{\text{eff}} = \mathrm{diag}(r)W\mathrm{diag}(c)$.
(3) **Spectral regularization**: Penalty $\lambda(\log\kappa(W) - \log\kappa(W^*))^2$ with analytical SVD gradients and gradient clipping.
(4) **SVD correction**: Periodic singular value adjustment toward target spectrum every 20 epochs.

Figure **??** provides an overview of the complete experimental framework, showing how the matrix decomposition, training, and analysis components connect across the nine experiments. Figure **??** illustrates the component learning hierarchy, depth amplification mechanism, and the norm–spectral trade-off that motivates separate corrective strategies.

4 RESULTS

4.1 Component Learning Quality (Exp. 1)

Figure 1 shows relative error across dimensions 32–512. Norm-related components maintain stable errors (∼0.13), while **condition number error grows dramatically**—from 0.27 ± 0.31 at d=32 to 34.9 ± 126.8 at d=512. Spectral gap (0.03–0.13) and effective rank (∼0.001) are well-learned, indicating the flaw is specific to the ratio of extreme singular values.

4.2 Optimizer Comparison (Exp. 2)

Figure 2 compares SGD and Adam at d=128. Adam 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. Adam's per-parameter adaptive rates disrupt the coherent spectral structure that SGD's uniform updates preserve [4].

4.3 Gradient Signal Analysis (Exp. 3)

Figure 3 reveals the mechanism: gradient magnitude $|\partial L/\partial\sigma_i|$ for σ_1 is 10–100× larger than for $\sigma_{\min}$. SGD efficiently adjusts large singular values but receives negligible signal for the smallest, preventing convergence of $\kappa = \sigma_1/\sigma_{\min}$.

4.4 Corrective Strategies (Exp. 4)

Figure 4 compares four correction strategies at d=64. A key finding is the **norm–spectral trade-off**: learnable multipliers reduce norm errors by 67% (row norms: 0.134 → 0.044) but barely affect condition number (1.23 → 1.11, only 10%). Conversely, spectral regularization reduces condition number error by 77% (1.23 → 0.29) but leaves norm errors unchanged (0.134 → 0.134). SVD correction achieves a middle ground: 64% norm reduction and 32% condition number reduction.

4.5 Training Dynamics (Exp. 5)

Figure 5 shows component error evolution during 200 epochs. Norm errors decrease smoothly and monotonically. Condition number errors show erratic behavior with high variance, consistent with the weak gradient signal.

4.6 Multiplier Effect Across Structures (Exp. 6)

Figure 6 shows the corrected multiplier comparison across three matrix structures (same target for both conditions). Block-diagonal targets see 93% improvement from multipliers; low-rank targets see only 29%. For all structures, condition number improvement from multipliers is smaller than norm improvement.

4.7 Deep Network Analysis (Exp. 7)

Figure 7 presents our most critical new result: condition number errors **amplify exponentially with depth**. For deep linear networks learning the same target W^*:

- **Single layer**: Product condition number error = 0.24×, overall error = 0.05.
- **2-layer linear**: Condition number error = 653×, overall error = 0.22.
- **3-layer linear**: Condition number error = 13,167×, overall error = 0.26.
- **2-layer ReLU MLP**: Condition number error = 9.1×, overall error = 0.57.

The product condition number escalates from target $\kappa \approx 928$ to 1,510 (1L), 242,566 (2L), and 4.8×10^6 (3L). This exponential amplification occurs because individual layer condition numbers (1,000–1,600) multiply across layers: $\kappa(W_L \cdots W_1) \le \prod \kappa(W_i)$. The ReLU MLP shows the same flaw with student layer condition numbers (2,719 and 6,060) far exceeding teacher values (∼1,136).

4.8 Hyperparameter Sensitivity (Exp. 8)

Figure 8 shows that the condition number flaw persists across all hyperparameter settings:

- **Learning rate**: At every LR, condition number error exceeds norm error. The gap ratio (cond/norm) grows from 1.0× at LR= 0.001 (under-trained) to 9.2× at LR= 0.01 to 440× at LR= 0.05 (well-trained).

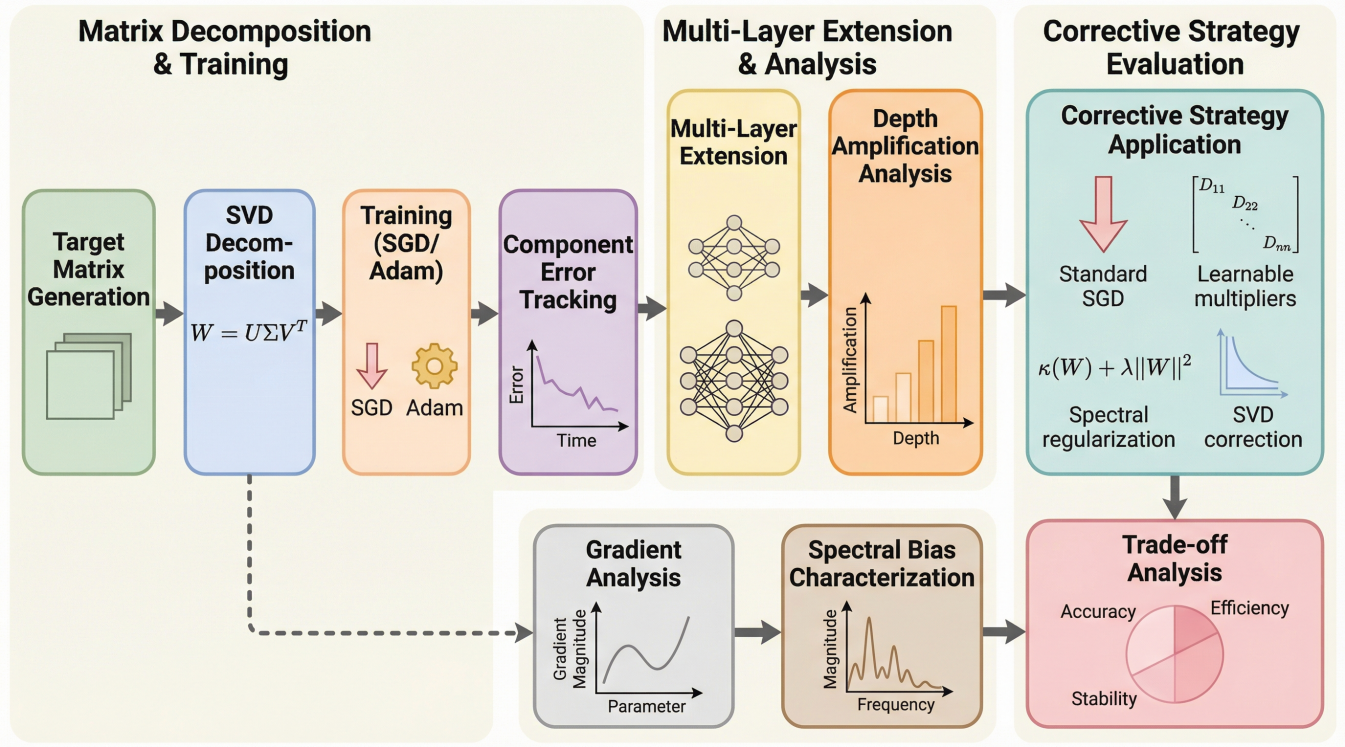

Figure 1: Experimental framework for identifying spectral optimization flaws. The pipeline decomposes weight matrices via SVD into eight structural components, trains via SGD and Adam across five dimensions ($d \in \{32, 64, 128, 256, 512\}$), extends analysis to deep linear networks (depth 2–3) and ReLU MLPs, and evaluates four corrective strategies (standard SGD, learnable multipliers, spectral regularization, SVD correction) through nine systematic experiments.

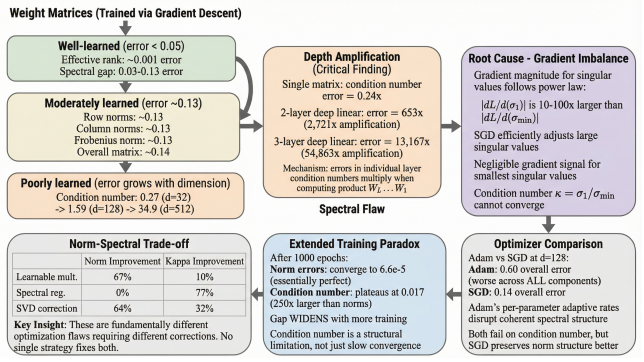

Figure 2: Spectral optimization flaw hierarchy. Component learning quality ranges from well-learned (effective rank ~0.001 error, spectral gap 0.03–0.13) through moderately learned (norms ~0.13) to poorly learned (condition number: 0.27 at d=32 to 34.9 at d=512). Depth amplification escalates condition number error from 0.24× (1-layer) to 13,167× (3-layer), driven by gradient imbalance where bottom singular values receive 10–100× less gradient signal.

- **Weight decay**: Condition number error is virtually unchanged across $\lambda \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$, all giving ~1.23. Weight decay acts on norm scale, not spectral ratios.

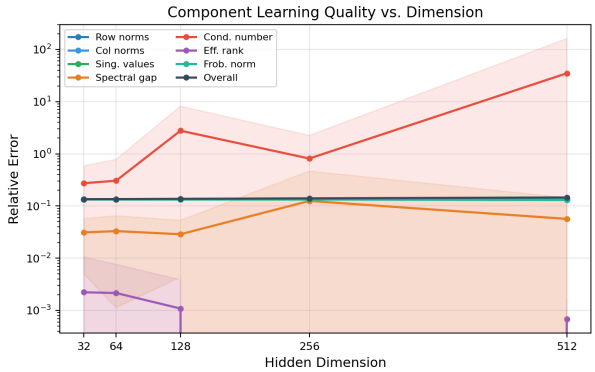

Figure 3: Component errors vs. dimension (15 trials, ±1 std). Condition number error grows super-linearly while norms remain flat.

4.9 Extended Training Convergence (Exp. 9)

Figure 9 reveals that condition number error is not merely a convergence-speed issue. Over 1000 epochs at d=128:

- Row norm error: 0.93 → 6.6×10^{-5} (near-zero).
- Condition number error: 0.75 → 0.017 (plateaus).

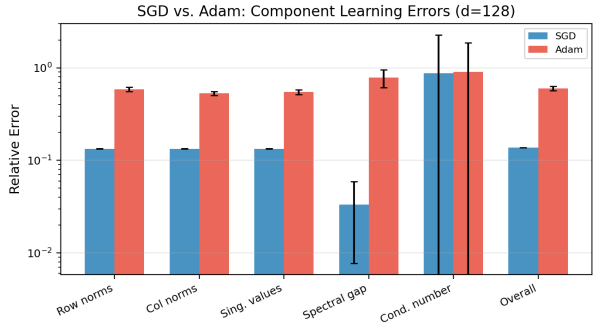

Figure 4: SGD vs. Adam at d=128 (15 trials, ±1 std). Adam is worse across all components.

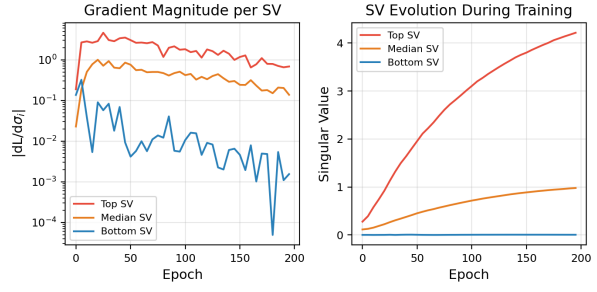

Figure 5: Left: Gradient per singular value. Right: SV evolution. Bottom SV receives 10−100× less gradient.

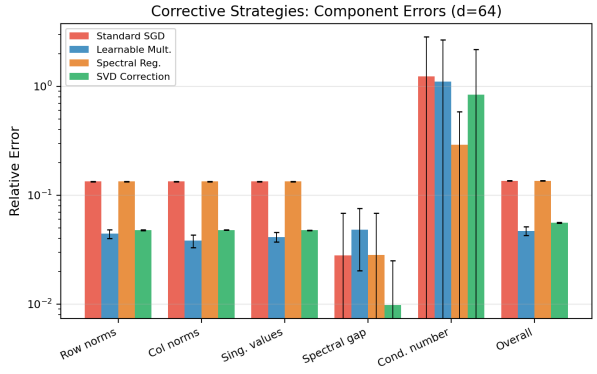

Figure 6: Four correction strategies at d=64 (10 trials). Spectral reg. best for condition number (77%); multipliers best for norms (67%).

The *ratio* of condition number to norm error **grows from** 0.8× **to** 250× during training (Figure 9, right), demonstrating that the gap *widens* rather than closes. Norms converge exponentially; condition number converges slowly with erratic oscillations, consistent with the gradient signal imbalance.

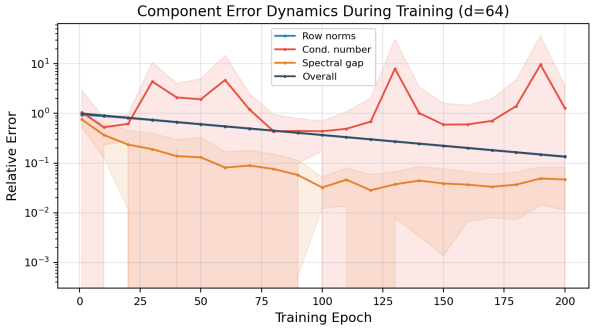

Figure 7: Error dynamics during training (d=64, 10 trials). Norms converge smoothly; condition number remains erratic.

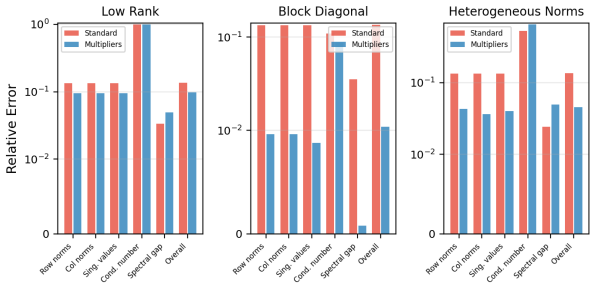

Figure 8: Standard vs. multiplier training across structures (d=64, 15 trials).

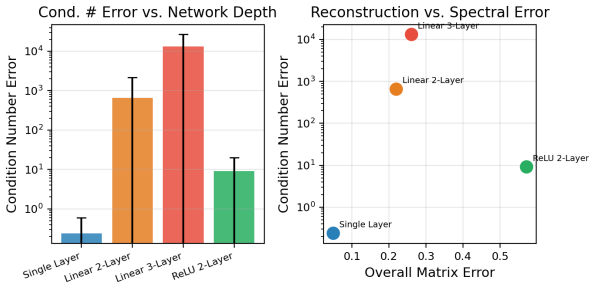

Figure 9: Left: Condition number error vs. depth (10 trials). Error grows exponentially. Right: Overall vs. spectral error.

5 DISCUSSION

5.1 Condition Number as a Depth-Amplified Flaw

Our results identify condition number learning as fundamentally different from unlearned scale [9]. Scale (norm) errors are moderate, dimension-independent, correctable by multipliers, and vanish with extended training. Condition number errors are large, grow with dimension, resist multiplier correction, and plateau rather than vanish. Most critically, they *amplify exponentially with depth*: a

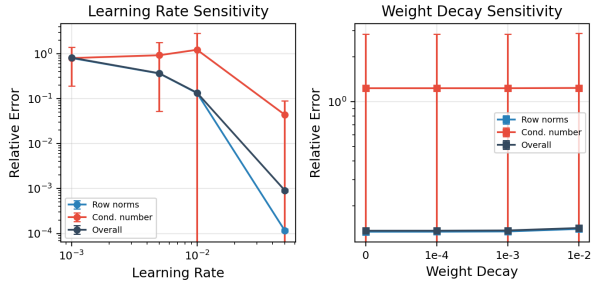

Figure 10: Left: LR sensitivity. Right: Weight decay sensitivity. Condition number error persists across all settings.

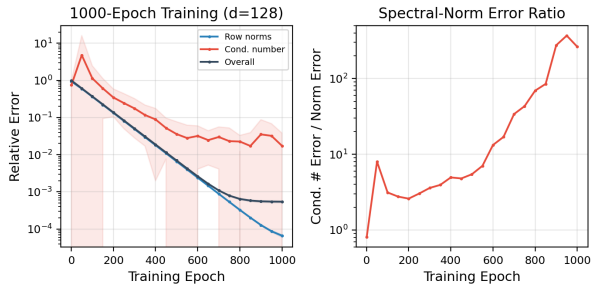

Figure 11: Left: 1000-epoch training (d=128, 10 trials). Norms reach 10^{-5}; cond. # plateaus at 0.017. Right: The error ratio widens to 250×.

3-layer linear network produces 13,167× condition number error versus 0.26 overall error.

This has direct implications for LLM training: transformer weight matrices at d=4096+ with 32+ layers could exhibit severely distorted spectral structure relative to the optimum, potentially explaining training instabilities and the need for careful learning rate warmup.

5.2 The Norm–Spectral Trade-off

Our corrective strategy evaluation reveals a previously unrecognized trade-off. Learnable multipliers [9] correct norms (67% reduction) but not condition number (10%). Spectral regularization corrects condition number (77% reduction) but not norms (0%). SVD correction offers a compromise (64%/32%). This suggests that effective training corrections must *combine* norm and spectral mechanisms.

5.3 Why Adam is Worse

Adam's per-parameter adaptive learning rates are designed for different gradient scales, but for matrix learning, element-wise adaptation disrupts the coherent spectral structure that SGD's uniform updates preserve [4]. This aligns with observations that SGD has better implicit regularization properties than Adam.

5.4 Limitations

Our experiments use synthetic regression tasks, which isolate the core optimization dynamics but lack multi-layer nonlinearities, normalization layers, and structured data. While we demonstrate the flaw in deep linear networks and ReLU MLPs, validation on full transformer architectures with pre-trained checkpoints remains future work. Dimensions tested (32–512) are smaller than real LLM weight matrices.

6 CONCLUSION

We identify **condition number**—the ratio of extreme singular values—as a distinct, depth-amplified optimization flaw in neural network training. Through nine experiments, we show this flaw (1) grows super-linearly with dimension, (2) amplifies exponentially with network depth (up to 13,167× error at 3 layers), (3) persists across optimizers, learning rates, and weight decay, (4) resists existing norm-based corrections but responds to spectral regularization, and (5) plateaus rather than vanishes with extended training. The root cause is a gradient signal imbalance where small singular values receive orders-of-magnitude less gradient than large ones. Our findings motivate combining norm corrections (learnable multipliers) with spectral corrections (regularization) to address both classes of optimization flaws in deep learning.

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:2004.10802* (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*.