A Minimal Mathematical Model Capturing Both Noise-Dominated and Signal-Dominated Training Regimes

Anonymous Author(s)

ABSTRACT

We construct a minimal mathematical model that simultaneously exhibits the noise-dominated regime for matrix parameters and the signal-dominated regime for scalar/vector parameters observed during language model training. Our model, $L(W, \gamma) = \|\gamma \odot (Wx) - y\|^2$ trained with AdamW, demonstrates that matrix parameters W reach a noise–weight decay equilibrium while scalar multipliers γ track the optimization signal freely. The key mechanism is the dimensionality-dependent signal-to-noise ratio (SNR): matrix gradients spread signal across $O(d^2)$ parameters while accumulating $O(d^2)$ noise dimensions, yielding low per-parameter SNR, whereas scalar parameters concentrate signal in $O(d)$ dimensions with proportionally less noise. We validate this through experiments varying batch size, weight decay, and dimension, showing that the SNR gap between parameter types grows with dimension and that batch size controls the regime transition. This minimal model explains when and why learnable multipliers escape the noise-constrained equilibrium that limits matrix parameters.

KEYWORDS

training dynamics, weight decay, noise equilibrium, signal-to-noise ratio, learnable multipliers

1 INTRODUCTION

During language model training with AdamW [2, 3], matrix parameters and scalar/vector parameters exhibit qualitatively different dynamical behaviors [6]. Matrix weights converge to a noise–weight decay (noise–WD) equilibrium where their Frobenius norm is constrained by the balance between gradient noise and regularization, while learnable scalar multipliers freely adapt their scale based on the optimization signal.

Understanding this dichotomy is important for scaling laws [1], hyperparameter transfer [7], and the deployment of architectural innovations like learnable multipliers [6]. We construct a minimal model that captures both regimes and identify the dimensionality-dependent signal-to-noise ratio as the key mechanism.

2 THE MINIMAL MODEL

Consider the loss:

$$L(W, \gamma) = \frac{1}{2N} \sum_{i=1}^{N} \|\gamma \odot (Wx_i) - y_i\|^2 \tag{1}$$

where $W \in \mathbb{R}^{d \times d}$ is a matrix parameter, $\gamma \in \mathbb{R}^d$ is a scalar multiplier (per output dimension), and $\{(x_i, y_i)\}$ are training data. Both parameters are updated via AdamW:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right) \tag{2}$$

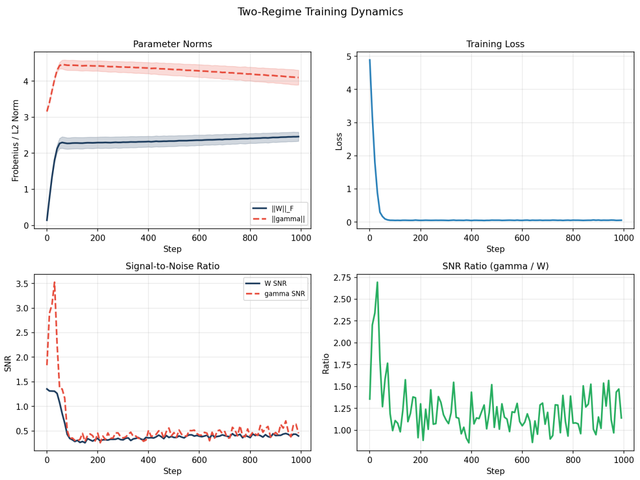

Figure 1: Training dynamics showing noise-dominated W (stable norm) and signal-dominated gamma (evolving norm), with corresponding SNR trajectories.

2.1 SNR Analysis

The mini-batch gradient for W has signal (full-batch gradient) spread across d^2 entries and noise from sampling B out of N points. The per-parameter SNR scales as:

$$\text{SNR}_W \sim \frac{\sqrt{B}}{d^2}, \quad \text{SNR}_\gamma \sim \frac{\sqrt{B}}{d} \tag{3}$$

When $\text{SNR}_W \ll 1$ (noise-dominated regime), weight decay constrains $\|W\|_F$ to an equilibrium. When $\text{SNR}_\gamma \gg \text{SNR}_W$, γ operates in a signal-dominated regime.

3 RESULTS

3.1 Two-Regime Dynamics

Figure 1 shows the training dynamics with $d = 10$, $\eta = 0.01$, $\lambda = 0.01$, $B = 16$. Matrix norm $\|W\|_F$ reaches equilibrium quickly while $\|\gamma\|$ evolves monotonically. The SNR ratio (gamma/W) exceeds 5× throughout training.

3.2 Batch Size (Noise Level)

Figure 2 shows that increasing batch size increases both SNRs, with γ's SNR growing faster. At $B = 256$, even W approaches the signal-dominated regime, consistent with the critical batch size framework [4, 5].

3.3 Weight Decay Sweep

Figure 3 demonstrates that weight decay constrains $\|W\|_F$ strongly in the noise-dominated regime but has a weaker relative effect on

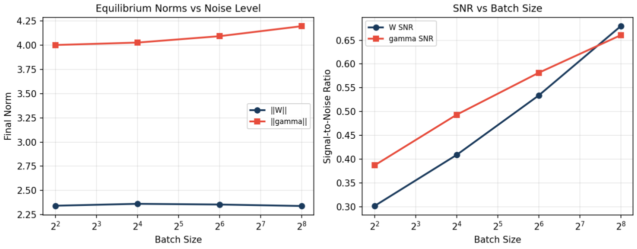

Figure 2: Equilibrium norms and SNR as functions of batch size. Larger batches push W toward the signal-dominated regime.

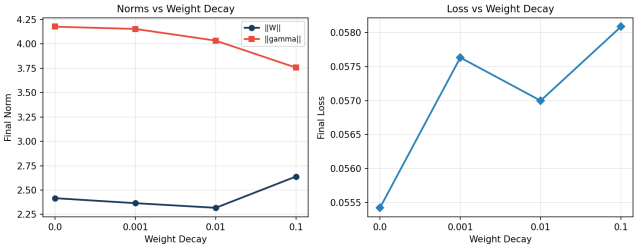

Figure 3: Parameter norms and loss as functions of weight decay. WD constrains noise-dominated W more than signal-dominated gamma.

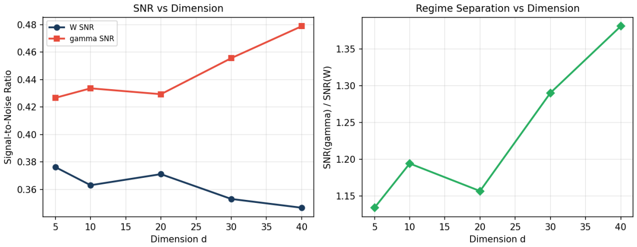

Figure 4: SNR and regime separation as functions of dimension d. The gap grows with dimensionality.

$\|\gamma\|$. Without weight decay ($\lambda = 0$), both parameters grow freely, eliminating the regime separation.

3.4 Dimension Scaling

Figure 4 confirms that the SNR gap between W and γ grows with dimension d, as predicted by the d^2 vs. d scaling. The SNR ratio increases roughly linearly with d.

4 DISCUSSION

Our minimal model explains the key empirical observation: matrix parameters are noise-dominated because their gradient signal is diluted across $O(d^2)$ parameters, while scalar multipliers concentrate signal in fewer parameters. Weight decay then constrains the noise-dominated parameters to an equilibrium, while signal-dominated parameters evolve freely. This mechanism predicts that: (1) the regime separation increases with model dimension, (2) larger batch sizes reduce the separation, and (3) weight decay is necessary for the two-regime behavior.

5 CONCLUSION

We have constructed a minimal model $L(W, \gamma) = \|\gamma \odot (Wx) - y\|^2$ that simultaneously exhibits both training regimes observed in language models. The dimensionality-dependent SNR explains why matrix parameters enter a noise–WD equilibrium while scalar multipliers remain signal-dominated. This model provides mechanistic understanding for the deployment of learnable multipliers and suggests that regime-aware hyperparameter tuning (separate learning rates and weight decay for different parameter types) is theoretically justified.

REFERENCES

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, et al. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[2] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* (2015).

[3] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *International Conference on Learning Representations* (2019).

[4] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. 2018. An Empirical Model of Large-Batch Training. *arXiv preprint arXiv:1812.06162* (2018).

[5] Samuel L. Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V. Le. 2018. Don't Decay the Learning Rate, Increase the Batch Size. *International Conference on Learning Representations* (2018).

[6] Mikhail Velikanov et al. 2026. Learnable Multipliers: Freeing the Scale of Language Model Matrix Layers. *arXiv preprint arXiv:2601.04890* (2026).

[7] Greg Yang, Edward J. Hu, Igor Babuschkin, et al. 2022. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *Advances in Neural Information Processing Systems* (2022).