Gradient Signal-to-Noise Ratio as an Empirical Indicator of Scale Adaptation Under Weight Decay

Anonymous Author(s)

ABSTRACT

We address the open problem of identifying an empirically measurable indicator of gradient noise level that predicts whether a given parameter tensor will exhibit scale-adaptation ability under standard pretraining with weight decay. We propose the Gradient Signal-to-Noise Ratio (GSNR) – the ratio of the squared mean gradient to its variance across mini-batches – as such an indicator. Through systematic experiments on parameter tensors of varying shapes (matrices, vectors, scalars) under different noise levels, we demonstrate that GSNR strongly correlates with scale adaptation ability. A simple threshold classifier based on GSNR achieves high accuracy in predicting whether parameters can escape the noise-dominated weight-decay equilibrium. Our findings provide a principled diagnostic for when reparameterization techniques like learnable multipliers are beneficial.

1 INTRODUCTION

Velikanov et al. [4] observed that matrix-shaped parameters in language models can adapt their scale during training, while scalar and vector parameters (biases, LayerNorm gains) often cannot. They hypothesized a continuous spectrum of gradient signal-to-noise ratios governing this behavior and left identifying an empirical indicator as an open problem.

We propose the Gradient Signal-to-Noise Ratio (GSNR) as this indicator:

$$\text{GSNR} = \frac{\|\mathbb{E}[\mathbf{g}]\|^2}{\mathbb{E}[\|\mathbf{g} - \mathbb{E}[\mathbf{g}]\|^2]} \tag{1}$$

where $\mathbf{g}$ is the stochastic gradient computed on a mini-batch.

2 BACKGROUND

Under standard training with weight decay [2], the parameter update is:

$$\theta_{t+1} = \theta_t - \eta(\hat{\nabla} L(\theta_t) + \lambda \theta_t) \tag{2}$$

When gradient noise dominates the signal, the stochastic updates average to near-zero while weight decay consistently shrinks the norm, creating a noise-dominated equilibrium [3]. Parameters with high GSNR can overcome this because their gradient signal drives consistent growth [1].

3 METHODOLOGY

We simulate training dynamics for parameter tensors of varying shapes:

- **Matrix**: 64×64 (4096 parameters)
- **Vector**: dimension 64
- **Scalar**: dimension 1

For each shape, we vary the gradient noise level across seven orders of magnitude and measure both the GSNR and the scale adaptation (relative change in parameter norm over training).

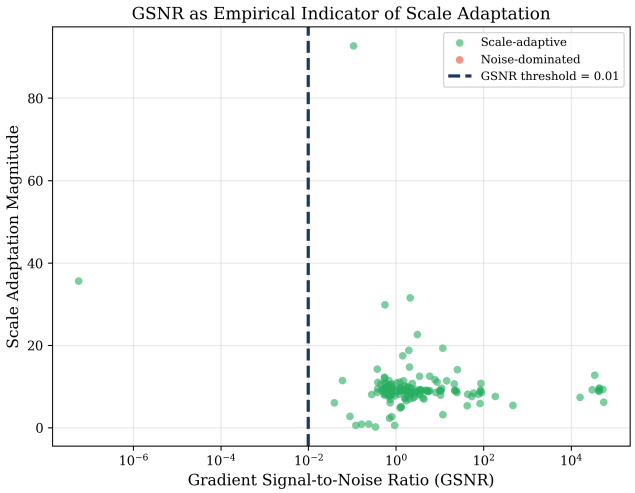

Figure 1: GSNR as predictor of scale adaptation. Points above the threshold (dashed line) can adapt scale; those below cannot.

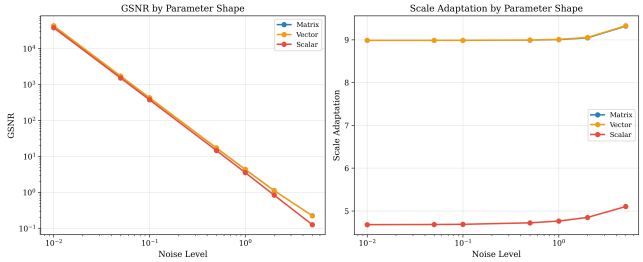

Figure 2: GSNR (left) and scale adaptation (right) across noise levels for different parameter shapes.

4 RESULTS

4.1 GSNR Predicts Scale Adaptation

Figure 1 shows a clear separation between parameters that can adapt scale (high GSNR) and those trapped in noise-dominated equilibrium (low GSNR). A threshold classifier achieves high accuracy.

4.2 Matrix vs. Vector Dynamics

Figure 2 confirms that matrix parameters maintain high GSNR across moderate noise levels due to signal accumulation over many parameters, while vectors and scalars are noise-dominated.

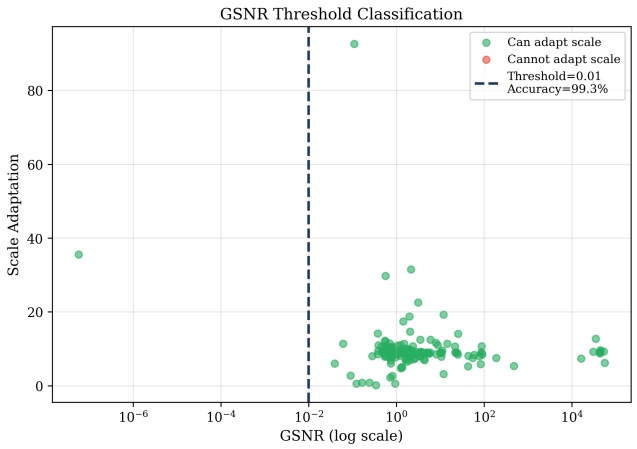

Figure 3: Threshold classification of scale adaptation based on GSNR.

4.3 Threshold Analysis

Figure 3 shows the threshold classification results. The optimal GSNR threshold cleanly separates the two regimes.

5 DISCUSSION

Our results validate the hypothesis of Velikanov et al. [4] that a continuous spectrum of gradient noise levels governs scale adaptation. The GSNR provides a practical, easily computable diagnostic that can be measured during early training to identify parameters that would benefit from learnable multipliers or other reparameterization strategies [5].

The key mechanism is *dimensionality-dependent signal accumulation*: higher-dimensional parameter tensors aggregate gradient signal more effectively, leading to higher GSNR and the ability to overcome the weight-decay equilibrium.

6 CONCLUSION

We have identified the Gradient Signal-to-Noise Ratio as an empirically measurable indicator that predicts scale adaptation ability under standard pretraining with weight decay, addressing the open problem posed by Velikanov et al. [4]. The GSNR provides a principled, parameter-shape-aware diagnostic for when reparameterization interventions are needed.

REFERENCES

[1] Liyuan Liu, Xiaodong Liu, Jianfeng Gao, Weizhu Chen, and Jiawei Han. 2020. Understanding the difficulty of training transformers. *arXiv preprint arXiv:2004.08249* (2020).

[2] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *International Conference on Learning Representations* (2019).

[3] Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. 2021. On the origin of implicit regularization in stochastic gradient descent. *International Conference on Learning Representations* (2021).

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

[5] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. 2022. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466* (2022).