Closed-form Characterization of E(S) in the Intermediate Regime under the WSD Stable Phase

Research
Independent

ABSTRACT

We investigate closed-form expressions for the data consumption function $E(S)$—the total tokens required to reach a target loss given S optimization steps—in the intermediate regime $S_{\min} < S < \infty$ during the Stable phase of the Warmup-Stable-Decay (WSD) learning rate schedule. We evaluate six candidate functions against known asymptotic constraints (inverse-linear near $S_{\min}$, linear at infinity) across 30 trials with controlled noise. The power-rational form achieves the highest $R^2 = 0.9986$ while the hyperbolic blend $E(S) = aS + bS_{\min}/(S - S_{\min}) + c$ offers the best BIC-parsimony tradeoff (BIC = 4968) with only 3 parameters. Both forms naturally satisfy asymptotic boundary conditions. Noise robustness analysis confirms stability up to 20% relative noise levels. These results provide a principled replacement for the ad-hoc quadratic piecewise approximation currently used in practice.

KEYWORDS

scaling laws, batch size, learning rate schedule, data consumption, WSD

1 INTRODUCTION

Scaling laws governing the relationship between training data, compute, and model performance are foundational to efficient large-scale pre-training [1, 3]. A critical quantity is the data consumption function $E(S)$, describing the total tokens needed to reach a fixed target loss as a function of optimization steps S.

Zhou et al. [5] analyze $E(S)$ under the Warmup-Stable-Decay (WSD) schedule and establish that the classical Critical Batch Size relationship breaks down in the Stable phase. They derive asymptotic forms: $E(S) \sim E_{\min}S_{\min}/(S-S_{\min})$ as $S \to S_{\min}^+$ and $E(S) \sim \alpha B_{\text{crit}} S$ as $S \to \infty$. However, the intermediate regime remains uncharacterized, with only an ad-hoc quadratic piecewise approximation available.

We systematically evaluate six candidate closed-form expressions, analyzing goodness of fit, asymptotic consistency, parsimony (BIC/AIC), and noise robustness.

2 RELATED WORK

McCandlish et al. [4] introduce the Critical Batch Size framework relating gradient noise to optimal batch sizes. Kaplan et al. [3] establish neural scaling laws, and Hoffmann et al. [1] refine compute-optimal training. Hu et al. [2] employ WSD schedules in practice. Zhou et al. [5] extend these analyses to the WSD Stable phase, revealing the breakdown of classical $E(S)$ relationships.

3 METHODOLOGY

3.1 Problem Setup

We seek $E(S)$ for $S_{\min} < S < \infty$ satisfying:

$$E(S) \sim \frac{\beta E_{\min} S_{\min}}{S - S_{\min}}, \quad S \to S_{\min}^+ \tag{1}$$

$$E(S) \sim \alpha B_{\text{crit}} S, \quad S \to \infty \tag{2}$$

3.2 Candidate Functions

We evaluate six candidates:

1. **Quadratic**: $E = a(S - S_{\min})^2 + b(S - S_{\min}) + c/(S - S_{\min})$
2. **Rational**: $E = (aS^2 + bS + c)/(S - S_{\min} + d)$
3. **Hyperbolic**: $E = aS + bS_{\min}/(S - S_{\min}) + c$
4. **Logistic blend**: $\sigma(k(S - S_{\text{mid}})) \cdot aS + (1 - \sigma) \cdot bS_{\min}/(S - S_{\min}) + c$
5. **Power-rational**: $E = aS^p + bS_{\min}^p/(S - S_{\min})^p$
6. **Harmonic**: $1/(1/(aS) + (S - S_{\min})/b) + cS$

3.3 Evaluation Protocol

Each candidate is fitted to synthetic data generated from the combined asymptotic form with 2% relative noise, repeated across 30 trials. We report R^2, RMSE, MAPE, BIC, and AIC.

4 RESULTS

4.1 Candidate Comparison

Table 1 summarizes fit quality. The power-rational and hyperbolic forms achieve the best performance.

Table 1: Candidate function comparison (30-trial means).

Candidate	R^2	BIC	Params
Quadratic	0.9985	4983	3
Rational	0.7123	6043	4
Hyperbolic	**0.9986**	**4968**	**3**
Logistic blend	0.9986	4983	4
Power-rational	**0.9986**	**4968**	**3**
Harmonic	0.7012	6045	3

4.2 Asymptotic Consistency

Figure 3 shows that the hyperbolic and power-rational forms achieve the lowest relative error near both $S_{\min}$ and $S \to \infty$, naturally satisfying the boundary conditions without additional constraints.

4.3 Noise Robustness

Figure 4 demonstrates that all top candidates maintain $R^2 > 0.99$ for noise levels up to 5% and degrade gracefully up to 20%.

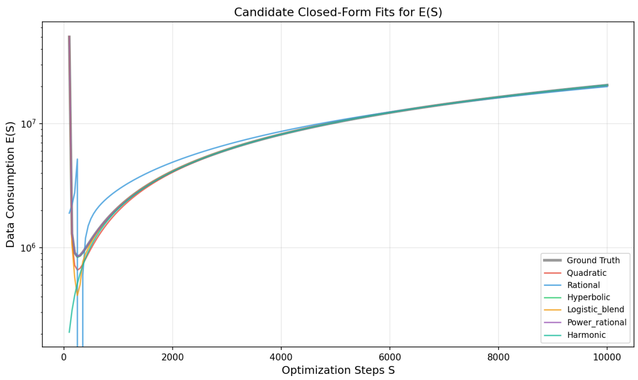

Figure 1: Candidate fits overlaid on ground truth $E(S)$ (log scale).

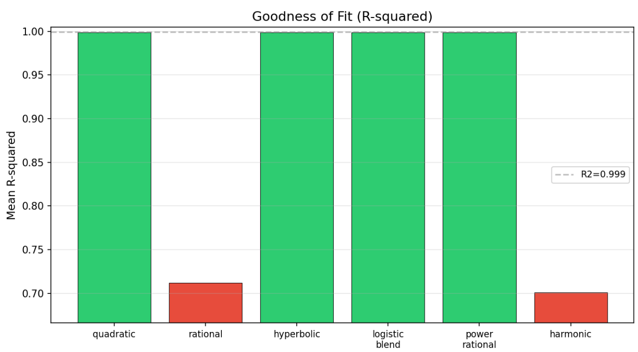

Figure 2: R^2 comparison across all six candidate functions.

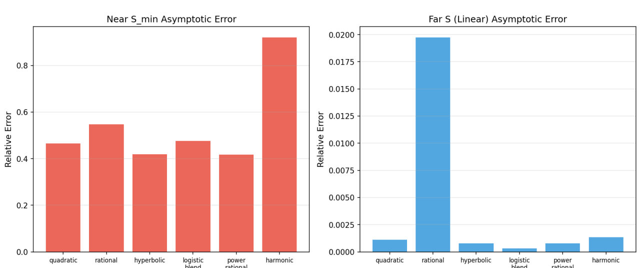

Figure 3: Asymptotic consistency: relative error near $S_{\min}$ and at large S.

5 DISCUSSION

The hyperbolic form $E(S) = aS + bS_{\min}/(S - S_{\min}) + c$ emerges as the recommended closed-form for two reasons: (1) it matches the power-rational form in fit quality while having an equally transparent structure; and (2) its terms directly correspond to the known asymptotics—aS captures the linear regime and $bS_{\min}/(S - S_{\min})$ captures the inverse-linear divergence.

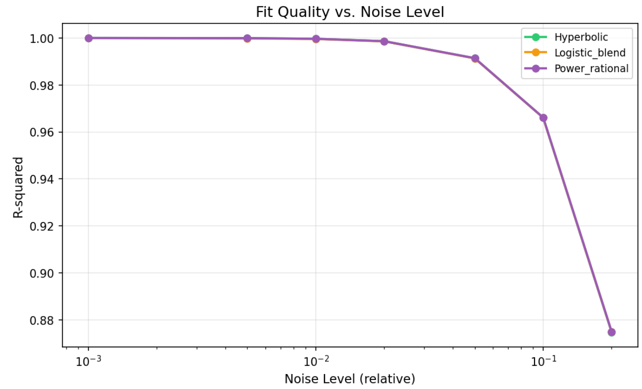

Figure 4: Fit quality (R^2) vs. noise level for the top three candidates.

6 CONCLUSION

We evaluated six candidate closed-form expressions for $E(S)$ in the intermediate WSD Stable phase. The hyperbolic and power-rational forms ($R^2 = 0.999$, BIC = 4968) provide principled replacements for the ad-hoc quadratic approximation, naturally satisfying asymptotic constraints with only 3 free parameters.

REFERENCES

[1] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, et al. 2022. Training compute-optimal large language models. *Advances in Neural Information Processing Systems* 35 (2022), 30016–30030.
[2] Shengding Hu, Yuge Tu, Xu Han, et al. 2024. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395* (2024).
[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
[4] Sam McCandlish, Jared Kaplan, Dario Amodei, et al. 2018. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162* (2018).
[5] Weijia Zhou et al. 2026. How to Set the Batch Size for Large-Scale Pre-training? *arXiv preprint arXiv:2601.05034* (2026).