Optimal Learning Rate Schedules under Functional Scaling Laws: Power Decay and Warmup–Stable–Decay

Binghui Li[*] LIBINGHUI@PKU.EDU.CN
Center for Machine Learning Research, Peking University
Zilin Wang[*] WANGZILIN@STU.PKU.EDU.CN
School of Mathematical Sciences, Peking University
Fengling Chen FENGLINGCHEN@STU.PKU.EDU.CN
School of Mathematical Sciences, Peking University
Shiyang Zhao ZHAOSHIYANG@STU.PKU.EDU.CN
School of Mathematical Sciences, Peking University
Ruiheng Zheng RUIHENGZHENG@STU.PKU.EDU.CN
School of Mathematical Sciences, Peking University
Lei Wu[†] LEIWU@MATH.PKU.EDU.CN
School of Mathematical Sciences, Peking University

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study optimal learning rate (LR) schedules under the functional scaling law (FSL) framework (Li et al., 2025), where loss dynamics are controlled by a source exponent $s > 0$ for signal learning and a capacity exponent $\beta > 1$ for noise forgetting. For a fixed training horizon N, we characterize the schedules that minimize the final-step loss under natural stability constraints and reveal a sharp *phase transition*. In the *easy-task regime* $s \geq 1 - 1/\beta$, the optimal schedule takes the power-decay form $\eta^*(z) = \eta_{\text{peak}}(1 - z/N)^{2\beta-1}$ with $\eta_{\text{peak}} \asymp N^{-(s-1+1/\beta)/(s+1/\beta)}$. In contrast, in the *hard-task regime* $s < 1 - 1/\beta$, the optimal schedule exhibits a warmup–stable–decay (WSD)-like (Hu et al., 2024) structure: it maintains the largest admissible LR for most of training and decays only near the end, with the decay phase occupying a vanishing fraction of the horizon.

We next study the practical setting where the decay shape is fixed and only the peak LR is tuned. To separate these two design choices, we introduce a family of fractional LR schedules that decouple peak-LR tuning from decay-shape design. We prove that fixed-shape schedules suffer from *capacity saturation*: each shape can adapt to the capacity exponent only up to a shape-dependent threshold, beyond which the achievable convergence rate no longer improves. This yields a principled criterion for evaluating commonly used schedules such as cosine and linear decay, revealing both their strengths and limitations.

We then apply the FSL-optimal power-decay schedule to one-pass stochastic gradient descent (SGD) for kernel regression and show that the last iterate attains the *exact* minimax-optimal convergence rate, eliminating the logarithmic gap in prior analyses. Finally, experiments validate our theoretical predictions in controlled settings and illustrate their usefulness for practical LR-schedule design in neural network training. [1]

Keywords: Learning rate schedule; stochastic gradient descent; functional scaling law; kernel regression; warmup–stable–decay

∗. Equal contribution.

†. Corresponding author. Lei Wu is also affiliated with the Center for Machine Learning Research, Peking University, and AI for Science Institute, Beijing.

1. Extended abstract. Full version appears at arXiv:2602.06797.

© 2026 B. Li[*], Z. Wang[*], F. Chen, S. Zhao, R. Zheng & L. Wu[†].

References

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Binghui Li, Fengling Chen, Zixun Huang, Lean Wang, and Lei Wu. Functional scaling laws in kernel regression: Loss dynamics and learning rate schedules. *arXiv preprint arXiv:2509.19189*, 2025.