Learning Rates Do Not Transfer Across Double Descent

Dayal Singh Kalra [2]

Maissam Barkeshli [1, 3,4]

Abstract

A central characteristic of Maximal Update Parameterization (μP) is *hyperparameter transfer*—the optimal hyperparameters (e.g., learning rate) found on small models continue to be optimal at large scales. This allows practitioners to tune hyperparameters cheaply on small models and reuse them at scale, avoiding the prohibitive cost of direct tuning on large models. In its original formulation, μP was derived under a set of analytical assumptions to ensure $\Theta(1)$ feature updates for a finite number of training steps in the large width limit $n \to \infty$. Although $\Theta(1)$ feature updates do not formally imply that the optimal learning rate transfers across widths, such a transfer has been widely observed empirically in practice. In this work, we identify a regime in which the optimal learning rate fails to transfer as the model is scaled. We find that the optimal learning rate for the test performance sharply changes across the double descent transition, yet remains fairly consistent within both the under-parameterized and over-parameterized regimes. We further show that weight decay and data augmentation can each improve the reliability of learning rate transfer, however, through different mechanisms. Our findings clarify the practical boundaries of hyperparameter transfer and highlight regimes where optimal learning rates are unlikely to transfer reliably.

1. Introduction

Identifying the optimal learning η^* rate at large scale is challenging as directly searching for it becomes computationally prohibitive, motivating strategies that estimate η^* from cheaper proxies at small scales [2]. *Maximal Update Parameterization* (μP) [13] is one such strategy: it parameterizes the model so that the optimal learning rate found on a smaller model remains optimal as width is scaled up, a property known as *learning rate transfer* [13, 14]. Since its introduction, learning rate transfer has been observed across a range of architectures, parameterizations, and optimizers [1, 4, 8, 11].

μP is derived to ensure $\Theta(1)$ updates under three analytical assumptions: (1) a finite number of training steps as the width $n \to \infty$, (2) a fixed dataset as the width scales, and (3) full alignment between weight updates and the activations they act on [13]. This Desiderata, however, do not formally imply that the optimal learning rate transfers across widths, and its underlying assumptions are routinely violated in practice: alignment between weight updates and activations does not reach the value assumed by μP [4], and modern training regimes scale data and training steps with model width. Moreover, transfer is observed in other, simpler parameterizations [4, 6, 9, 11], suggesting the full set of assumptions might not be needed for transfer.

1. Department of Physics, University of Maryland, College Park
2. Department of Computer Science, University of Maryland, College Park
3. Joint Quantum Institute, University of Maryland, College Park
4. Meta Superintelligence Labs, Fundamental AI Research

© D.S. Kalra & M. Barkeshli.

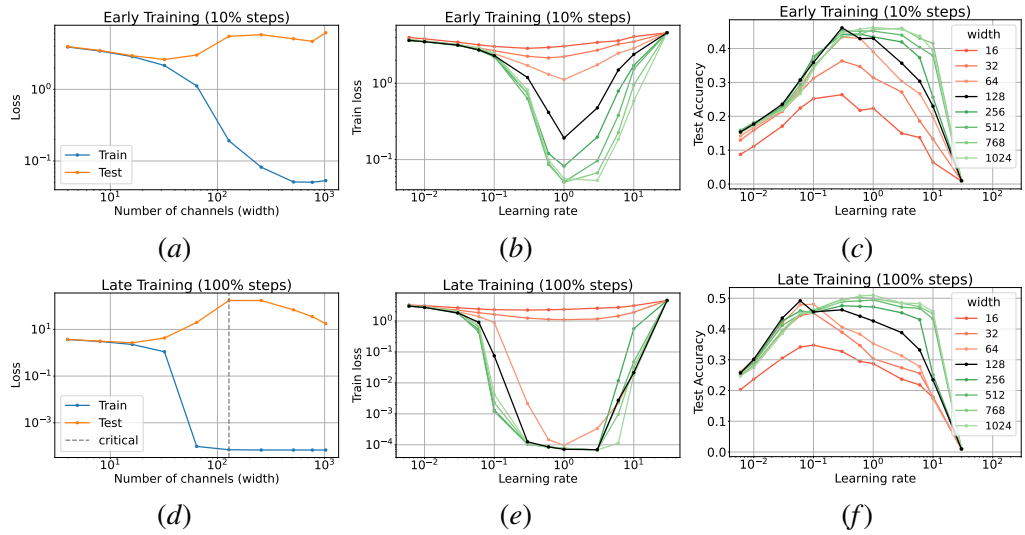


Figure 1: *Optimal learning rate does not transfer across double descent.* CNNs trained on CIFAR-100 using SGD, with under-parameterized widths shown in red, over-parameterized widths in green, and the interpolating width is shown in black. *Top row (early training, 10% of total steps)*: (a) training and test loss as a function of width, (b) optimal learning rate for the loss, (c) optimal learning rate for the test accuracy. *Bottom row (end of training)*: (d) training and test loss as a function of width, showing the double descent curve with the interpolation threshold at $n \approx 128$, (e) optimal learning rate for the training loss remains consistent across widths, with a small decrease at the smallest widths, (f) optimal learning rate for the test accuracy exhibits a sharp transition at the interpolation threshold, differing by roughly $10\times$ between the under- and over-parameterized regimes.

Despite these gaps, learning rate transfer continues to be observed empirically [14]. Yet it remains poorly understood when transfer is expected to hold and when it can break down. In this work, we identify one such regime in which transfer breaks down—the optimal learning rate for test performance changes sharply across the double-descent transition, while remaining consistent within the under- and over-parameterized regimes.

2. Optimal Learning Rates Across the Double Descent Transition

To systematically examine whether learning rates transfer across double descent, we consider the CIFAR image classification [7] task, where the interpolation threshold can be precisely identified. We train vanilla CNNs on the CIFAR-100 dataset using SGD with batch size 512 and no momentum. We clip the gradients when their L2 norm exceeds 1.0. The learning rate schedule follows a warmup of 4000 steps and a cosine decay to $1/10$ of the peak value η. The networks are parameterized using spectral parameterization [15], which generalizes μP to settings with non-uniform layer widths. We sweep over widths $n \in \{16, 32, \ldots, 1024\}$ and peak learning rates $\eta \in [6 \times 10^{-3}, 10^{2}]$, and identify the peak optimal learning rate η^* at the end of training.

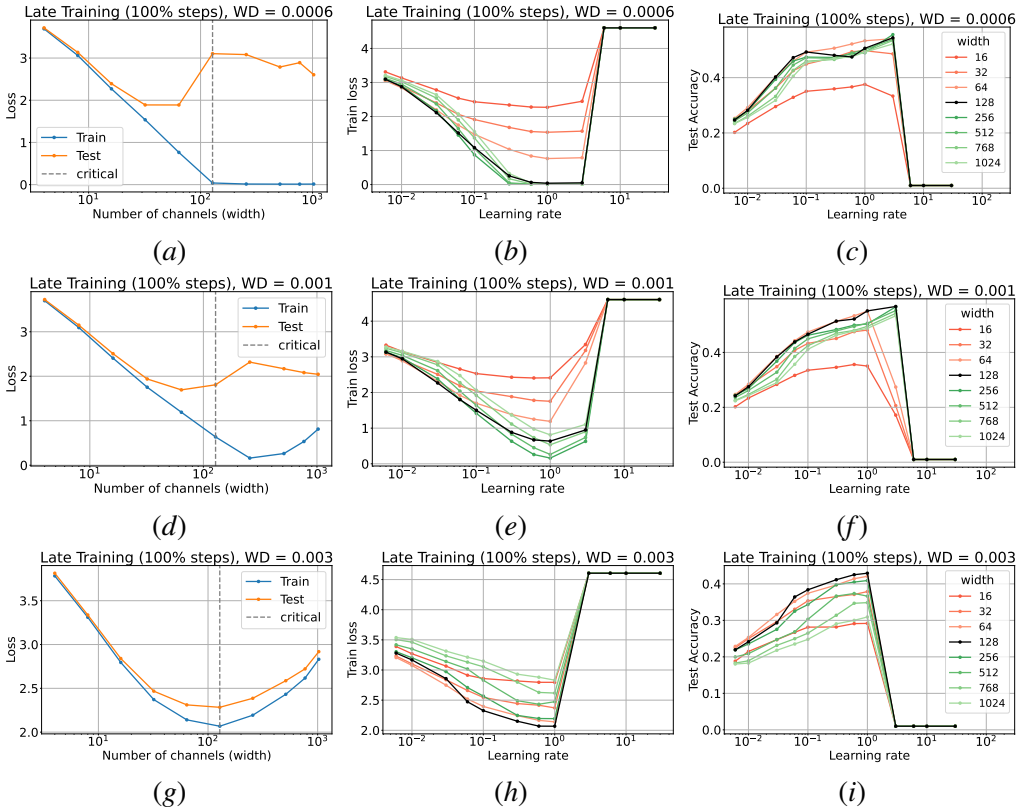
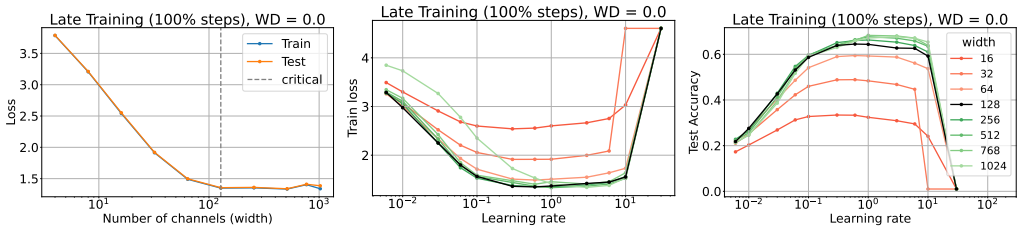


Figure 2: *Regularization improves learning rate transfer.* Same setup as Figure 1, with weight decay (WD) set to 0.0006 (top), 0.001 (middle), and 0.003 (bottom). Columns show training and test loss at the end of training vs width, training loss, and test accuracy as a function of learning rate. As weight decay increases, the double descent peak in the test loss is progressively suppressed, and the sharp transition in η^* for the test accuracy at the interpolation threshold is reduced. At large weight decay values, η^* transfers across widths, but comes at a cost of performance.

Figure 1(d) shows the classic double descent curve, with the test loss peaking at the interpolation threshold around width $n \approx 128$. The optimal learning rate for the training loss remains fairly consistent across widths (Figure 1(e)), with a small decrease at the smallest widths. In the overparameterized regime, the training loss converges to a near-zero value, so a range of learning rates achieve comparable performance. In contrast, the optimal learning rate for the test accuracy changes sharply with width (Figure 1(f)): underparameterized models prefer $\eta \sim 0.1$, while overparameterized models prefer $\eta \sim 1.0$, which is a $\sim 10\times$ jump at the interpolation threshold. This gap is mild early in training (Figure 1(c)) and widens as training proceeds. Within each regime, however, the optimal learning rate is consistent, indicating that the failure of transfer is localized to the transition between regimes rather than a general width-dependence.



(*a*) (*b*) (*c*)

Figure 3: *Data augmentation pushes the interpolation threshold beyond the widths considered.* Same setup as Figure 1 with augmentation enabled. (a) Training and test loss decrease monotonically with width, indicating that all models are in the under-parameterized regime. (b, c) Training loss and test accuracy curves have a consistent shape, but the optimal learning rate slowly but continuously increases toward larger values with width, in contrast to the sharp transition observed without augmentation.

3. Regularization and Data Augmentation Improve Transfer

Our results so far suggest that the failure of learning rate transfer is tied to double descent. Prior work has shown that regularization and data augmentation can suppress double descent [10]. We therefore ask whether either is sufficient to restore the transfer of the optimal learning rate for the test accuracy.

3.1. Effect of Weight Decay

We next ask whether weight decay can restore transfer by suppressing double descent. Figure 2 shows the effect of three weight decay strengths. Since the per-layer learning rate in μP scales with width, the effective decay strength $\eta\lambda$ in weight decay would scale with width as well. We omit the μP width scaling from the learning rate when computing the decay update so that the effective weight decay is consistent across widths. At small weight decay values (WD $= 0.0006$, top row), the double descent peak in the test loss is still clearly visible, but transfer of η^* for the test accuracy is improved: the sudden transition in η^* at the interpolation threshold is replaced with a smoother increase with width. At large weight decay values (WD $= 0.003$, bottom row), the peak is suppressed, and the test loss recovers a classical bias-variance shape; however, this comes at the cost of overall test performance. Notably, at this strong weight decay, larger widths no longer result in better test accuracy, undermining the practical motivation of scaling up. Taken together, weight decay can restore transfer, but at the cost of performance and the benefit of scale itself.

3.2. Effect of Data Augmentation

We next ask whether data augmentation can also improve transfer. We apply data augmentation in the following order: random horizontal flips, random cropping, and mixup [16]. Figure 3(a) shows that the training and test loss decrease monotonically with width, and the double descent peak has disappeared from the range of widths we consider. This is consistent with augmentation increasing the effective dataset size and thereby pushing the interpolation threshold beyond our largest model. With all models in the under-parameterized regime, η^* for the test accuracy (Figure 3(c)) no longer

exhibits a sharp transition; it instead drifts slowly but continuously toward larger values as width increases. This residual drift is much smaller than the order-of-magnitude jump observed without augmentation, but shows that the width dependence of η^* is not entirely eliminated.

Taken together, both weight decay and data augmentation improve learning rate transfer for the test accuracy, but through different mechanisms: weight decay suppresses the double descent peak in place, while data augmentation moves the interpolation threshold past the widths considered.

4. Related Works

μP and its variants have been shown to enable reliable learning rate transfer across a range of architectures, optimizers, and scaling dimensions [1, 3, 4, 6, 11, 13, 14]. Most of this evidence, however, comes from language model pre-training, where the number of training tokens typically exceeds the number of parameters, placing it far from the interpolation threshold and deep in the underparameterized regime [12]. In contrast, few works examine where transfer breaks down. Most relevant to us, Ghosh et al. [5] construct synthetic settings—two-layer ReLU networks learning a k-index target and a ball-indicator function—in which the optimal learning rate drifts or shifts abruptly with width even under μP. By comparison, we identify transfer breakdown in a standard image-classification setup, where the optimal learning rate shifts sharply at the double descent interpolation threshold, and examine how regularization and data augmentation can improve transfer.

5. Discussion

In this work, we identified a regime where learning rate transfer breaks down: while η^* for the training loss remains fairly consistent across widths, η^* for the test accuracy jumps by roughly an order of magnitude across the double descent transition. This finding suggests that learning rate transfer depends not only on the parameterization, but also on the loss landscape being optimized and the metric being used to define the optimal learning rate η^*. Weight decay and data augmentation can each improve transfer, but through different mechanisms: weight decay suppresses the double descent peak, while data augmentation pushes the interpolation threshold beyond the widths of interest. Yet neither restores transfer fully, as either a residual width dependence of η^* remains, or test performance degrades.

More broadly, our results raise the possibility that learning rate transfer may break down across other learning transitions (e.g., grokking, emergent capabilities, or task generalization) wherever the test metric landscape scales differently from the training loss being optimized. Understanding which conditions are necessary for transfer, and which transitions cause it to fail, is an exciting direction for future work.

References

[1] Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KZJehvRKGD.

[2] DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang,

Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. Deepseek llm: Scaling open-source language models with longtermism, 2024. URL https://arxiv.org/abs/2401.02954.

[3] Nolan Dey, Bin Claire Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers, 2025. URL https://arxiv.org/abs/2505.01618.

[4] Katie E Everett, Lechao Xiao, Mitchell Wortsman, Alexander A Alemi, Peter J Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=0ksNeD1SJT.

[5] Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer, 2025. URL https://arxiv.org/abs/2512.22768.

[6] Dayal Singh Kalra and Maissam Barkeshli. Quantifying hyperparameter transfer and the importance of embedding layer learning rate, 2026. URL https://arxiv.org/abs/2605.21486.

[7] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

[8] Etai Littwin and Greg Yang. Adaptive optimization in the ∞-width limit. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zgVDqw9ZUES.

[9] Sean McLeish, John Kirchenbauer, David Yu Miller, Siddharth Singh, Abhinav Bhatele, Micah Goldblum, Ashwinee Panda, and Tom Goldstein. Gemstones: A model suite for multi-faceted scaling laws, 2025. URL https://arxiv.org/abs/2502.06857.

[10] Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=7R7fAoUygoa.

[11] Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie E Everett, Alexander A Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d8w0pmvXbZ.

[12] Lechao Xiao. Rethinking conventional wisdom in machine learning: From generalization to scaling, 2024. URL https://arxiv.org/abs/2409.15156.

[13] Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yang21c.html.

[14] Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=Bx6qKuBM2AD.

[15] Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

[16] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.