A Proof of Learning Rate Transfer Under μP

Soufiane Hayou
Department of Applied Mathematics and Statistics
Johns Hopkins University

Abstract

We provide the first proof of learning rate transfer with width in a linear multi-layer perceptron (MLP) parametrized with μP, a neural network parameterization designed to "maximize" feature learning in the infinite-width limit. We show that under μP, the optimal learning rate converges to a *non-zero constant* as width goes to infinity, providing a theoretical explanation to learning rate transfer. In contrast, we show that this property fails to hold under alternative parametrizations such as Standard Parameterization (SP) and Neural Tangent Parametrization (NTP). We provide intuitive proofs and support the theoretical findings with extensive empirical results.

1 INTRODUCTION

The recent successes in AI are mostly fueled by scale: large neural networks trained on large corpuses of data. Given a fixed training dataset, the size of a neural network can be scaled by increasing the width (hidden dimension) and/or depth (number of layers). As we scale these dimensions, several hyperparameters (HPs) must be adjusted with scale to avoid numerical overflows. Motivated by this empirical observation, several works have explored the large-width limit of neural networks and its impact on optimal HPs. He et al. (2016) introduced the "1/fan-in" initialization which normalizes the weights to achieve order one activations as width grows (Note that Neal (1996) was the first to introduce the "1/fan-in" initialization in the context of Bayesian neural networks). The Neural Tangent Kernel (NTK, (Jacot et al., 2018)) was one of the first attempts to understand training dynamics of large-width neural networks. The authors showed that under the neural tangent parametrization, training dynamics converge to a kernel regime in the infinite-width limit, a phenomenon known as lazy training (Chizat et al., 2020). In this regime, neural features are almost identical to their values at initialization and training dynamics can be linearized around initialization. It quickly became clear that NTK regime does not represent practical training of neural network, which exhibit significant feature learning. Yang and Hu (2021) reverse-engineered this problem by investigating neural parametrizations that result in feature learning in the infinite-width limit and introduced the Maximal Update Parametrization (μP) which sets precise scaling exponents for the initialization and learning rate. A nice by-product of μP is HP transfer, or where optimal HPs seem to converge as width increases, a very useful property since it allows tuning HPs on relatively small models and using them for larger models with no additional tuning cost (see Fig. 1 for a conceptual illustration). The authors conjectured that HP transfer resulted from the fact that μP achieves "maximal" feature learning, and therefore the limiting dynamics are "optimal" in the sense that no other limit (corresponding to other parametrizations) is better in terms of training loss, thus leading to the convergence of the optimal HPs as width grows. While this intuition is valid to some extent, to the best of our knowledge, no rigorous proof of HP transfer exists in the literature.

Perhaps the most important hyperparameter is the learning rate, which generally requires some tuning in practice. Motivated by this, we focus on *learning rate transfer* in this work and present the first proof for this phenomenon in deep linear networks parametrized with μP. Specifically, we consider a linear Multi-Layer Perceptron (MLP) and show that at training steps t, the optimal learning rate converges to a non-zero constant as width goes to infinity, providing a theoretical proof for learning rate transfer observed in practice. Our proof is based on the observation that with linear MLPs, the loss function at any training step can

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

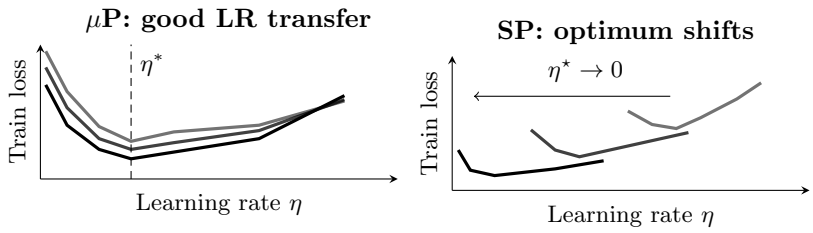

Figure 1: **Conceptual illustration of learning-rate transfer.** Left: Under μP, loss curves across widths share (approximately) the same optimal learning rate η^*. Right: Under SP, the optimal learning rate $\eta_n^\star$ shifts toward 0 as width grows. Curves illustrating different widths (darker $\Rightarrow$ wider).

be expressed as a polynomial function of the learning rate. We study convergence dynamics of these polynomials and their roots and conclude on the convergence of the optimal learning rate as width goes to infinity. We further show that other parametrizations such as Standard Parametrization (SP) (and Neural Tangent Parametrization (NTP)) lead to significant shift in optimal learning rate as width grows, thus requiring expensive tuning.

The paper is structured as follows. In Section 2, we introduce notation and definitions. In Section 3, we provide a full characterization of LR transfer after one step and study the convergence rate of the optimal LR. In Section 4, we provide a proof for LR transfer for general step t. In both Section 3 and Section 4, extensive simulations are provided to support the theoretical results. In Section 5, we provide additional empirical results with varying setups: activation function, optimizer, depth, training time.

1.1 Related work

Infinite-width analysis. There is a rich literature on the theory of infinite-width neural networks. The first works on infinite-width theory are related to approximation results showing that neural networks are universal approximators when the width to infinity (see e.g. Hornik et al. (1989); Cybenko (1989)). Perhaps the first methodological work on infinite-width neural networks was a study of priors in large-width Bayesian neural network by Neal (1996), where the author studied how Gaussian prior should be scaled as network width increases, and showed that single-layer Bayesian networks converge to a Gaussian process in the infinite-width limit, a result that was later used in Williams (1996) to compute infinite-width posteriors, and was later generalized to multi-layer networks in (Lee et al., 2018; de G. Matthews et al., 2018). Subsequent research has examined the impact of initialization (Schoenholz et al., 2017; Hayou et al., 2019; Li et al., 2021; Daniely et al., 2017), the activation functions (Hayou et al., 2019), learning rate (Yang et al., 2022), batch size (Zhang et al., 2025), etc. Others works studied how these HPs should scale with depth (assuming large-width) (Hayou et al., 2021; Yang et al., 2023; Bordelon et al., 2023). There is also a rich literature on training dynamics of infinite-width neural networks, including the literature on the neural tangent kernel (Jacot et al., 2018; Hayou et al., 2022; Arora et al., 2019; Chizat et al., 2020; Allen-Zhu et al., 2019), and the literature on mean-field neural networks (Sirignano and Spiliopoulos, 2019; Mei et al., 2019; Mignacco et al., 2021; Chizat et al., 2022).

Hyperparameter transfer. Yang and Hu (2021) introduced μP, a parametrization that specifies how initialization and learning rate should scale with width n. The authors derived this parametrization by searching for HPs that yield feature learning in the infinite-width limit, in contrast to neural tangent parametrization which leads to a kernel regime in the limit (Jacot et al., 2018). In particular, the authors observed that μP leads to an interesting phenomenon: HP transfer with width, where optimal HPs tend to stabilize as width increases. It was conjectured that feature learning properties of the infinite-width limit under μP is the main factor behind HP transfer. In (Yang et al., 2022), the authors showed that μP yields HP transfer in Large Language Models (LLMs) of GPT-3 scale. However, other works showed mixed results on the efficacy of μP with LLMs and Diffusion model (Team, 2023; Blake et al., 2025; Lingle, 2025; Everett et al., 2024; Hayou and Liu, 2025; Kosson et al., 2025; Zheng et al., 2025). Other works include (Noci et al., 2024) which studied learning rate transfer studied from the angle of Hessian geometry and its connection to the edge of stability (Cohen et al., 2022), (Bordelon and Pehlevan, 2025) which studied learning rate transfer in linear networks, (Chizat and Netrapalli, 2025) which studied feature learning and learning rate transfer, and other works that extended HP transfer to cover other optimizers (Jordan et al., 2024; Ahn et al., 2025; Pethick et al., 2025), depth scaling (Yang et al., 2023; Bordelon et al., 2023; Dey et al., 2025), etc.

2 SETUP AND DEFINITIONS

Consider a linear Multi-Layer Perceptron (MLP)

$$f(x) = V^\top W_L W_{L-1} \dots W_1 W_0 x, \tag{1}$$

where $x \in \mathbb{R}^d$ is the input, $W_0 \in \mathbb{R}^{n\times d}$, $W_\ell \in \mathbb{R}^{n\times n}$ for $\ell \in \{1, 2, \dots, L\}$, and $V \in \mathbb{R}^n$, are the weights. While we consider one-dimensional output, our results can be generalized to neural networks with multi-dimensional outputs.

Model Eq. (1) is trained by minimizing the quadratic loss $\mathcal{L} = \frac{1}{2m}\sum_{i=1}^m (f(x_i) - y_i)^2$, where $\mathcal{D} = \{(x_i, y_i), i = 1 \dots m\}$ is the training dataset. For the sake of simplicity, we only train the weight matrices $W_1, W_2, \dots, W_L$, and fix W_0 and V to their initialization values.[1] For weight updates, we use gradient descent (GD)

$$W_\ell^{(t+1)} = W_\ell^{(t)} - \eta \nabla_{W_\ell^{(t)}} \mathcal{L}, \tag{2}$$

where $t \in \{1, 2, \dots, T\}$ is the step, η is the learning rate, and $W_\ell^{(0)}$ is randomly initialized.

When training a neural network, we should first set the hyperparameters (HPs) such as initialization and learning rate. Generally speaking, as width grows, it should be expected that optimal HPs shift with width, indicating dependence on width n. Therefore, it makes sense to explicitly parametrize HPs as a function of width. For instance, He initialization (He et al., 2016) sets the initialization weights as centred gaussian random variables with "1/fan_in" variance, where "fan_in" refers to the dimension of the previous layer, e.g. n for $\ell \in \{1, 2, \dots, L\}$, and d for $\ell = 0$. For the learning rate, μP scaling parametrizes the learning rate as ηn^{-1} for Adam (Yang et al., 2022) and η for gradient descent. We call these *neural parametrizations*, a notion that we formalize in the next definition.

Definition 1 (Neural Parametrization)**.** *A neural parametrization for model Eq. (1) specifies the constants $(\alpha_\ell)_{0\le\ell\le L}, \alpha_V$, and α_η:*

- *Initialization: $W_0 \sim \mathcal{N}(0, d^{-\alpha_0})$, $W_\ell \sim \mathcal{N}(0, n^{-\alpha_\ell})$, and $V \sim \mathcal{N}(0, n^{-\alpha_V})$.*
- *Learning rate: $\eta \times n^{-c}$.*

While a neural parametrization should in-principle cover all HPs (initialization, learning rate, batch size, Adam's (β_1, β_2), etc), we consider only the initialization and learning rate in this work. Here are two examples of such neural parametrizations:

- Standard Parametrization (SP): $\alpha_\ell = 1$ for $\ell \in \{0, \dots, L\}$, $\alpha_V = 1$, and $c = 0$. SP does not specify width exponent for the learning rate, hence the choice of $c = 0$. [2]
- Maximal Update Parametrization (μP): $\alpha_\ell = 1$ for $\ell \in \{0, \dots, L\}$, $\alpha_V = 2$, and $c = 0$. Notice that the only difference with SP is the choice of $\alpha_V = 2$. For the learning rate, μP coincides with SP when the training algorithm is GD, however, when considering Adam (Kingma and Ba, 2017), the learning rate exponent becomes $c = 1$.

2.1 What is Learning Rate (LR) Transfer?

In the context of μP, LR transfer refers to the *stability of optimal LR as model width grows*. Let η_n be the optimal learning rate for neural network Eq. (1) of width n; LR transfer occurs if η_n converges to a constant $\eta_\infty > 0$. As a result of this convergence, we can expect the optimal learning rate to remain stable for $n \gg 1$, i.e. increasing model beyond some base width $n_0 \gg 1$ does not significantly affect optimal LR. This is a highly desirable property as it implies that optimal LR can be tuned on model width n_0 and used for models of widths $n \gg n_0$, thus reducing tuning costs. However, for such property to be useful, η_n should converge fast enough so that considering $|\eta_n - \eta_\infty|$ is small enough for practical model widths (e.g. $n = 10^3$). A recent concurrent work by Ghosh et al. (2025) studied the mechanisms of fast HP transfer and connects it to the geometry of the gradients.

Learning rate transfer as described in Yang and Hu (2021). The authors showed empirically that learning rate transfer occurs under μP. They justified this observation with the intuition that μP is associated with "maximal" feature learning. Specifically, μP is the only parametrization that achieves $\Delta z = \Theta(1)$ asymptotically in width n for any activation z in the neural network, while other parametrizations such as Standard Parametrization (SP) and Neural Tangent Parametrization (NTP) lead to suboptimal learning dynamics as model width n grows (e.g. vanishing feature updates $\Delta z = \mathcal{O}(n^{-\beta})$ or exploding feature updates $\Delta z = \Omega(n^\alpha)$ for some $\alpha, \beta > 0$). While heuristic arguments were provided as to why learning rate transfer occurs under μP, to the best of our knowledge, no formal proof was provided showing the convergence of η_n in the case of multi-layer neural networks.

[1] Our results can be extended to the case where W_0 and V are trainable. For μP, the learning rate for W_0 should be parametrized as $\eta \times n$.

[2] While some works introduce a learning rate scaling for SP (see e.g. Everett et al. (2024)), the standard parametrization represents common practice (e.g. PyTorch defaults) which do not set default scaling rules for the learning rate.

Proving learning rate transfer is non-trivial. From a mathematical perspective, proving learning rate transfer requires proving the convergence of the optimal learning rate η_n to a non-zero constant as width goes to infinity. Optimal learning rate is (naturally) defined as the argmin of the training loss over a some set of possible values for the learning rate η. Since the loss is a random variable (from the random initialization), proving convergence of optimal learning rate requires proving convergence of the argmin of a stochastic process.

We provide the *first proof to LR transfer* with width in linear MLPs of any depth (model 1). We further show that with other parameterizations such as SP (or NTP), learning rate doesn't transfer. Let us first introduce some notation that will be consistently be used throughout the paper.

Notation. Hereafter, n will always denote model width. As n grows, given sequences $c_n \in \mathbb{R}$ and $d_n \in \mathbb{R}^+$, we write $c_n = \mathcal{O}(d_n)$ when $c_n < \kappa d_n$ for n large enough, for some constant $\kappa > 0$. We write $c_n = \Theta(d_n)$ if we have $\kappa_1 d_n \leq c_n \leq \kappa_2 d_n$ for some $\kappa_1, \kappa_2 > 0$. For vector sequences $c_n = (c_n^i)_{1 \leq i \leq k} \in \mathbb{R}^k$ (for some $k > 0$), we write $c_n = \mathcal{O}(d_n)$ when $c_n^i = \mathcal{O}(d_n^i)$ for all $i \in [k]$, and same holds for other asymptotic notation. Finally, when the sequence c_n is a vector of random variables, asymptotics are defined in the sense of the second moment (L_2 norm). For a vector $z \in \mathbb{R}^n$, we will use the following norms: $\|z\| = \left(\sum_{i=1}^n z_i^2\right)^{1/2}$ (euclidean norm), and $\|z\|_1 = \sum_{i=1}^n |z_i|$ (ℓ_1 norm). For two vectors $z, z' \in \mathbb{R}^n$, $z' \otimes z$ denotes the outer product. Finally, all expectations in our analysis are taken with respect to random initialization weights.

The training dataset $\mathcal{D}$ is considered fixed, and the weights $(W_\ell)_{1 \leq \ell \leq L}$ are updated with GD (Eq. (2)). We use superscript (t) for $t \in \{0, 1, \ldots, T\}$ to denote the gradient step, e.g. $W_\ell^{(t)}$ is the weight matrix at the ℓth layer at training step t. Finally, since our goal is to study the asymptotics of the optimal learning rate, we abuse the notation and write $\mathcal{L}_n^{(t)}(\eta)$ for the loss function of a neural network of width n trained for t steps with GD with learning rate η. Given width n and training step t, an optimal LR can be defined as $\eta_n^{(t)} \in \operatorname{argmin}_{\eta > 0} \mathcal{L}_n^{(t)}(\eta)$. Note that the loss function $\mathcal{L}_n^{(t)}$ depends on the random initialization weights, and therefore is a random variable itself. As a result, the optimal learning rate $\eta_n^{(t)}$ is also a random variable that is measurable with respect to the sigma-algebra generated by the initialization weights. When $\eta_n^{(t)}$ converges to non-zero deterministic constant $\eta_\infty^{(t)}$ as width n goes to infinity, we say that LR transfer occurs.

Definition 2 (LR Transfer)**.** *Let $t \in \{1, 2, \ldots, T\}$. We say that LR transfers with width n if there exists a deterministic constant $\eta_\infty^{(t)} > 0$ such that the optimal learning rate $\eta_n^{(t)}$ converges in probability to a $\eta_\infty^{(t)}$ as n goes to infinity.*

The condition $\eta_\infty^{(t)} > 0$ is crucial for LR transfer. In the case where $\eta_\infty^{(t)} = 0$, all we can say is that $\eta_n^{(t)}$ converges to 0 but setting the learning rate to 0 results in no training. When $\eta_\infty^{(t)} > 0$, the limiting training loss is different by a $\Theta(1)$ factor in width n, i.e. achieving non-trivial feature updates.

Note that we consider convergence in probability for the definition of LR transfer, but it is equivalent to convergence in distribution since convergence in distribution to a constant implies convergence in probability. In the next section, we provide a comprehensive analysis of LR transfer for $t = 1$ with convergence rates. We later prove LR transfer for general t.

3 LEARNING RATE TRANSFER: FULL CHARACTERIZATION AT $t = 1$

We characterize the asymptotic behavior of the optimal learning rate after one gradient step. We show that under μP, LR transfer occurs. For other parametrizations such as SP and NTP, the optimal learning rate converges to zero or diverges, respectively, which implies that LR transfer doesn't occur in these cases. Here, we only study μP and SP, the result for NTP is straightforward.

3.1 Learning Rate Transfer Under μP

We assume that initialization and learning rate exponents are set according to μP, namely

- Initialization: $W_0 \sim \mathcal{N}(0, d^{-1})$, $W_\ell \sim \mathcal{N}(0, n^{-1})$, and $V \sim \mathcal{N}(0, n^{-2})$.
- Learning rate: constant $\eta > 0$.

Intuitive analysis. Consider the simple case where the dataset consists of a single datapoint (x, y). We will later state the result for general dataset size. The loss function at step $t = 1$ is given by $\mathcal{L}_n^{(1)}(\eta) = \frac{1}{2}(f^{(1)}(x) - y)^2$, and the gradients are given by rank-1 matrices

$$\nabla_{W_\ell} \mathcal{L}_n^{(0)} = \chi \, b_{\ell+1} \otimes a_{\ell-1}$$

where

$$\begin{cases} b_\ell = (W_\ell^{(0)})^\top (W_{\ell+1}^{(0)})^\top \ldots (W_L^{(0)})^\top V, \\ a_\ell = W_\ell^{(0)} \ldots W_1^{(0)} W_0 x, \\ \chi = f^{(0)}(x) - y. \end{cases}$$

At $t=1$, model output for input x is given by

$$f^{(1)}(x) = V^\top \left[\prod_{\ell=1}^{L}(W_\ell^{(0)} - \eta\,\chi\, b_{\ell+1}\otimes a_{\ell-1})\right] W_0 x,$$

which can be expressed as a polynomial in η. For integers $p_2 \geq p_1$, define the products

$$J_{p_2:p_1} = W_{p_2}^{(0)} W_{p_2-1}^{(0)} \dots W_{p_1}^{(0)},$$

and $J_{p_2:p_1} = I_n$ for $p_2 < p_1$. We can write

$$f^{(1)}(x) = f^{(0)}(x) + \sum_{\ell=1}^{L} \phi_\ell \eta^\ell,$$

where for $k \in \{1,\dots,L\}$,

$$\phi_k = (-\chi)^k \sum_{1\leq \ell_1 < \dots < \ell_k \leq L} \|b_{\ell_k+1}\|^2 \|a_{\ell_1-1}\|^2 \Psi(\ell_1,\dots,\ell_k),$$

with Ψ given by

$$\Psi(\ell_1,\ell_2,\dots,\ell_k) = \prod_{j=1}^{k} a_{\ell_j-1}^\top J_{\ell_j-1:\ell_{j-1}+1}\, b_{\ell_{j-1}+1}$$

for $k \geq 2$, and $\Psi(\ell) = 1$ for all ℓ by definition when $k=1$.

Now assume that the optimal learning rate for width n at step $t=1$, $\eta_n^{(1)} = \text{argmin}_{\eta>0}\, \frac{1}{2}(f^{(1)}(x)-y)^2$, is unique for convenience. The asymptotic behavior of $\eta_n^{(1)}$ w.r.t n depends mainly on the coefficients ϕ_ℓ:

- $\ell = 1$ (the coefficient of degree 1 monomial):

$$\phi_1 = (-\chi)\sum_{\ell=1}^{L} \|b_{\ell+1}\|^2 \|a_{\ell-1}\|^2.$$

  Strong Law of Large Numbers (SLLN) as $n\to\infty$ yields convergence to $y\,L\|x\|^2\,d^{-1}$ almost surely.

- $\ell \geq 2$: we prove that ϕ_ℓ converges to 0 in $\mathbb{L}_2$ for $\ell \geq 2$. The convergence of ϕ_ℓ to 0 is a result of the fact that $f^{(0)}(x)$ converges to zero because of the Mean-field-type initialization of the projection layer $V \sim \mathcal{N}(0, n^{-2})$. We now state these results below for general dataset size m.

Results. Recall the training dataset consisting of m samples $\mathcal{D} = \{(x_i,y_i), i=1,\dots,m\}$. Similar to the notation above, define

$$\begin{cases} a_{\ell,i} := W_\ell W_{\ell-1}\cdots W_0 x_i, \\ b_\ell := W_\ell^\top W_{\ell+1}^\top \cdots W_L^\top V, \\ \chi_i := f^{(0)}(x_i) - y_i, \text{ for } i\in[m], \end{cases}$$

with $a_{-1,i} := x_i$ and $b_{L+1} := V$ by definition. The loss at step $t=1$ is given by $\mathcal{L}_n^{(1)}(\eta) = \frac{1}{2m}\sum_{i=1}^m (f^{(1)}(x_i) - y_i)^2$ and the gradients are weighted sums of rank-1 matrices

$$\nabla_{W_\ell}\mathcal{L}_n^{(0)} = \frac{1}{m}\sum_{i=1}^{m} \chi_i\, b_{\ell+1}\otimes a_{\ell-1}^{(i)}. \tag{3}$$

Model output $f^{(1)}(x)$ can be expressed as a polynomial function in learning rate η. The next result characterizes the asymptotic behavior of its coefficients.

Lemma 1 (Asymptotic coefficients)**.** *Fix $x\in\mathbb{R}^d$. Then, there exists random scalars $(\phi_\ell)_{1\leq\ell\leq L}$ such that $f^{(1)}(x) = f^{(0)}(x) + \sum_{\ell=1}^L \phi_\ell\eta^\ell$, and for $\ell\in\{2,\dots,L\}$, $\|\phi_\ell\|_{L_2} = \mathcal{O}\big(n^{-(\ell-1)/2}\big)$. Moreover, we have*

$$\phi_1 \xrightarrow[n\to\infty]{a.s.} \frac{L}{m}\sum_{i=1}^{m} y_i\,\frac{\langle x, x_i\rangle}{d}.$$

The proof of Lemma 1 is provided in Appendix A and is based on the intuition developed above. The result shows that coefficients of degree $\ell\geq 2$ vanish as $n\to\infty$ with a rate of $n^{-(\ell-1)/2}$ in width. Interestingly, only the monomial of degree one does not vanish in the limit, and converges to a deterministic constant. As a result, asymptotically, the loss is quasi-quadratic in η. This allows us to fully characterize the convergence of the optimal learning rate $\eta_n^{(1)}$ at $t=1$.

For the remainder of the paper, we define the normalized input Gram matrix $K = \big(d^{-1}\langle x_i, x_j\rangle\big)_{1\leq i,j\leq m} \in \mathbb{R}^{m\times m}$,, and the vector containing all outputs $y = (y_1,\dots,y_m)^\top \in \mathbb{R}^m$. The next result shows LR transfer at $t=1$ and characterizes the limiting optimal learning rate and the convergence rate.

Theorem 1 (LR transfer at $t=1$)**.** *Assume that $Ky \neq 0$ and define*

$$\eta_\infty^{(1)} = \frac{m}{L}\frac{y^\top K y}{\|Ky\|^2}.$$

Then, for any compact interval $I\subset[0,\infty)$ containing $\eta_\infty^{(1)}$, and any $\eta_n^{(1)}\in \text{argmin}_{\eta\in I}\mathcal{L}_n^{(1)}(\eta)$, we have

$$\eta_n^{(1)} - \eta_\infty^{(1)} = O_{\mathbb{P}}(n^{-1/2}).$$

Theorem 1 shows convergence of the optimal LR to a deterministic limit $\eta_\infty^{(1)} > 0$, thus proving learning rate transfer at $t=1$. The convergence rate is $\mathcal{O}(n^{-1/2})$ which is expected with large-width asymptotics. The compact interval I can be arbitrarily large as long as it contains $\eta_\infty^{(1)}$. The proof is provided in Appendix A and is based on several technical lemmas used to control large-width deviations.

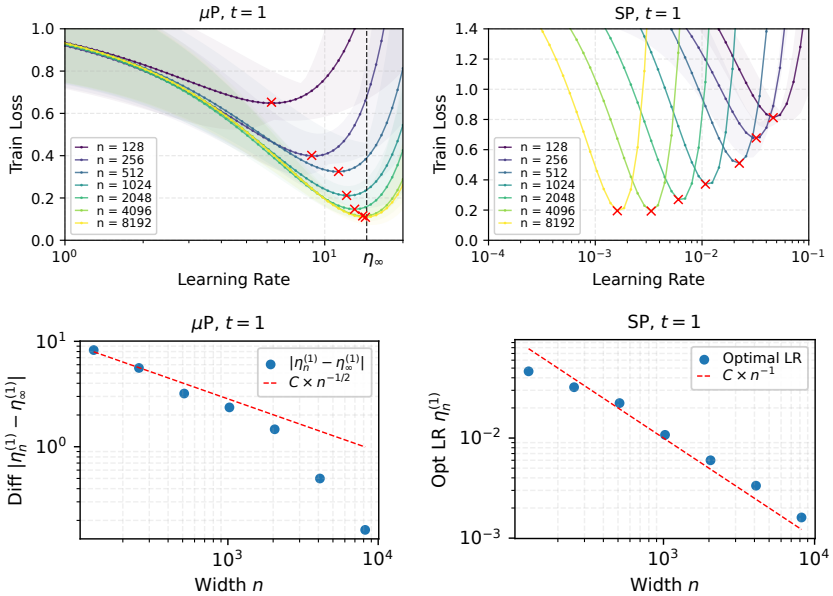

Figure 2: Optimal LR as a function of model width with 3 random seeds. **(Top)** Train loss as function of LR $\eta_n^{(1)}$ at $t = 1$ for both μP and SP. **(Bottom)** Convergence of optimal LR $\eta_n^{(1)}$ as width grows.

To verify LR transfer empirically, we trained a three layers linear MLP parametrized with μP with varying widths $n \in \{2^k, k = 7, \ldots, 13\}$ with GD. Training data consists of synthetically generated data $y = w^\top x + \epsilon$ where $x \sim \mathcal{N}(0, I_d)$ and $w \sim \mathcal{N}(0, d^{-1} I_d)$ $(d = 1)$, and $\epsilon \sim \mathcal{N}(0, 0.01)$. We use $N = 1000$ samples for training (see Section 5 for more details about experimental setup). Fig. 2 (top left) shows optimal learning rate with μP as a function of width. Convergence analysis is displayed in the bottom left figure. We observe convergence of the optimal LR $\eta_n^{(1)}$ to the theoretical value $\eta_\infty^{(1)}$ as n grows which confirms the theoretical findings. Interestingly, the empirical convergence rate seems to match the theoretical prediction of $n^{-1/2}$ up to width $n = 1024$ then becomes faster for larger widths. This indicates that our upperbound $\mathcal{O}(n^{-1/2})$ is likely not tight for large widths and we currently do not have an explanation for this sudden change in convergence rate.[3]

3.2 Failure of LR Transfer Under SP/NTP

With standard parametrization, the only difference with μP lies in how the projection layer weight V is initialized: $V \sim \mathcal{N}(0, n^{-1})$ for SP, instead n^{-2} variance with μP. Other weights are initialized as $W_0 \sim \mathcal{N}(0, d^{-1})$ and $W_\ell \sim \mathcal{N}(0, n^{-1})$ for $\ell = 1, \ldots, L$, and the learning rate is a constant η that is not parametrized with width. Note that this is only true for GD (and SGD). For Adam (Kingma and Ba, 2017), SP and μP also differ in the learning rate exponent ($c = 1$ for μP and $c = 0$ for SP).

The next result shows that optimal learning rate with SP converges to 0 as width grows, suggesting that LR transfer cannot occur under this parametrization.

Theorem 2 (No LR transfer under SP)**.** *Let $\bar{\eta} > 0$ be an arbitrary constant, and $\eta_n^{(1)} \in \arg\min_{\eta \in [0, \bar{\eta}]} \mathcal{L}_n^{(1)}(\eta)$ for the one-step loss, and assume $Ky \neq 0$. Then $\eta_n^{(1)} \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$.*

Intuitively, because of the n^{-1} variance in V initialization, all coefficients are amplified by a factor $\sqrt{n}$ compared to μP, so the optimal one-step LR compensates for that growth. The proof of Theorem 2 is provided in Appendix A.

With NTP (Jacot et al., 2018), the opposite occurs. To see this, recall that NTP involves multipliers in front of the weights. Specifically, we take $\widetilde{W}_\ell, \widetilde{V}$ with i.i.d. $\mathcal{N}(0, 1)$ entries and define

$$W_0 = \frac{1}{\sqrt{d}} \widetilde{W}_0, \quad W_\ell = \frac{1}{\sqrt{n}} \widetilde{W}_\ell, \quad V = \frac{1}{\sqrt{n}} \widetilde{V}.$$

This is distributionally identical to $W_\ell \sim \mathcal{N}(0, n^{-1})$ and $V \sim \mathcal{N}(0, n^{-1})$. However, the "effective" learning rate is now scaled by the $n^{-1/2}$ factor in front of the weights, which leads to a kernel regime in the limit (no feature learning). Hence, optimal learning rate

[3] Note that LR transfer is most usefull when convergence is fast.

tends to compensate for this down-scaling by blowing-up with width.

Fig. 2 (right) shows the optimal LR as a function of width n under SP. Unlike with μP, the optimal LR $\eta_n^{(1)}$ does not exhibit convergence to a non-zero constant, but rather shifts significantly with width, converging to zero. Therefore, LR transfer does not occur with SP. The bottom right figure shows the empirical convergence rate which seems to be faster than $n^{-1/2}$ and closer to n^{-1}.

4 LEARNING RATE TRANSFER AT ANY STEP

We generalize the results from the previous section and prove LR transfer for general gradient step t under mild conditions. The proof relies on the fact that for any step t and input x, model output $f^{(t)}(x)$ can be expressed as a polynomial function in η, similar to the previous section, although with coefficients that depend on initialization in a more complex way. By studying the behavior of this polynomial for η small/large enough, we show that optimal η converges almost surely to a non-zero deterministic constant under μP; hence proving LR transfer for general t.

4.1 Understanding The Difficulty At $t \geq 2$

In the previous section, we showed that after one step the network output becomes asymptotically linear in η. This significantly simplified the asymptotic analysis of $\eta_n^{(1)}$ and allowed derivation of a closed-form expression for the limit $\eta_\infty^{(1)}$. For $t \geq 2$, such analysis is nontrivial since the linear asymptotics no longer hold. Indeed, for $t \geq 2$, higher-order monomials in η are no longer negligible when n is large. For instance, for $t = 2$, we show that a coefficient of order $3L - 1$ in $f^{(2)}(x)$ converges to a non-zero constant as $n \to \infty$. Recall model output for a given input x

$$f^{(2)}(x) = V^\top \left(\prod_{\ell=1}^{L} W_\ell^{(2)} \right) W_0 x,$$

where

$$W_\ell^{(2)} = W_\ell^{(1)} - \eta m^{-1} \sum_{i=1}^{m} \chi_i^{(1)} \, b_{\ell+1}^{(1)} (a_{\ell-1,i}^{(1)})^\top,$$

and, extending the notation from previous section,

$$\begin{cases} b_\ell^{(t)} = (W_\ell^{(t)})^\top (W_{\ell+1}^{(t)})^\top \dots (W_L^{(t)})^\top V, \\ a_{\ell,i}^{(t)} = W_\ell^{(t)} W_{\ell-1}^{(t)} \dots W_1^{(t)} W_0 x_i, \\ \chi_i^{(t)} = f^{(t)}(x_i) - y_i. \end{cases}$$

Unlike in the one-step analysis, model output at $t = 2$ depends on the terms $b_\ell^{(1)}$, $a_\ell^{(1)}$, and $\chi^{(1)}$, which are all functions of the learning rate η. The leading monomial in $b_\ell^{(1)}$ is of degree $L - \ell + 1$ while in $a_\ell^{(1)}$ is of degree ℓ. $\chi^{(1)}$ is a polynomial of degree L in η. As a result, the leading monomial in $f^{(2)}(x)$ is of degree $L \times (1 + L + (L - \ell + 1) + \ell) = 2L(L + 1)$ in η. However, as in the analysis of the first step, the limiting polynomial as n goes to infinity may not be of degree $2L(L + 1)$. Expanding the product in $f^{(2)}(x)$ yields

$$f^{(2)}(x) = f^{(1)}(x) + \sum_{\ell=1}^{L} \phi_\ell(\eta) \eta^L,$$

where $\phi_L(\eta) = (-1)^L V^\top \left(\prod_{\ell=1}^{L} \gamma_\ell \right) W_0 x$, and $\gamma_\ell = m^{-1} \sum_{i=1}^{m} \chi_i^{(1)} \, b_{\ell+1}^{(1)} (a_{\ell-1,i}^{(1)})^\top$.

Note that we emphasized the dependence of ϕ_L on learning rate η in the notation. In the next result, we show that $\phi_L(\eta)$ converges to a non-zero constant as width goes to infinity, which is different from what we saw in the one-step loss.

Lemma 2 (Non-linear asymptotics at $t = 2$)**.** *The limit of the coefficient $\phi_L(\eta)$ can be expressed as*

$$\lim_{n \to \infty} \phi_L(\eta) = (-m)^L \sum_{i=1}^{m} \gamma_i \frac{\langle x_i, x \rangle}{d},$$

where,

$$\begin{cases} \gamma_i = \sum_{1 \leq i_2, \dots, i_L \leq m} \zeta_{i, i_2, \dots, i_L}, \\ \zeta_{i_1, i_2, \dots, i_L} = \prod_{j=2}^{L} f_\infty^{(1)}(x_{i_j}) \prod_{j=1}^{L} \left(f_\infty^{(1)}(x_{i_j}) - y_{i_j} \right), \end{cases}$$

with $f_\infty^{(1)}(x) = \eta \frac{L}{m} \sum_{i=1}^{m} y_i \frac{\langle x_i, x \rangle}{d}$.

Lemma 2 shows that $\phi_L(\eta)$ converges to a polynomial of degree $2L - 1$ in η as n goes to infinity.[4] Adding the η^L term in $f^{(2)}(x)$, we obtain that $f^{(2)}(x)$ converges to a polynomial that has a non-zero term of order $3L - 1$. Therefore, in contrast to step 1, step 2 involves more complex dependencies in η, and a full characterization of the minimum is highly non-trivial in this case. This complexity should be expected to "increase" with step t as gradient dependencies on η become more complex with t.

However, under an additional mild condition, we show that optimal LR converges to a non-zero constant for any step t, proving LR transfer for general t. Similar to the previous section, let $K = \left(d^{-1} \langle x_i, x_j \rangle \right)_{1 \leq i,j \leq m}$ be the input Gram matrix and $y = (y_1, y_2, \dots, y_m)^\top \in \mathbb{R}^m$ be the vector containing all inputs from the training dataset.

[4] Note that here, we are implicitly assuming that $f_\infty^{(1)}(x_i) \neq y_i$ for all i, which is a realistic assumption since it is highly unlikely to interpolate the data after one gradient step.

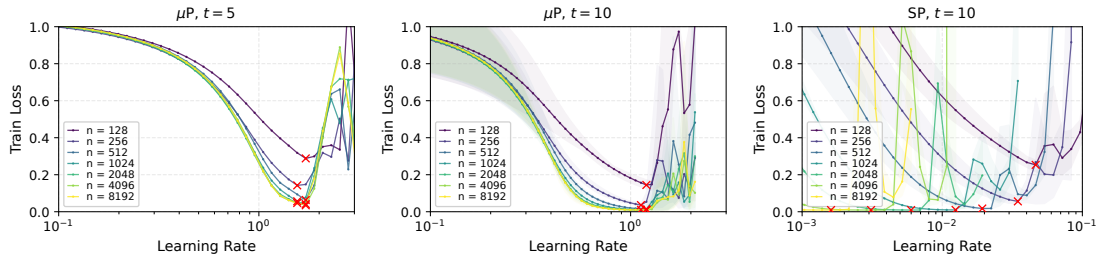

Figure 3: Train loss as function of LR at $t = 5$ and $t = 10$ for both μP and SP. Results are shown with 3 random seeds

Theorem 3 (LR transfer at step t)**.** *Assume that $Ky \neq 0$. Then the following holds:*

1. *Given a fixed input x, the t-step model output $f^{(t)}(x)$ can be expressed as a polynomial function in η where the coefficients depend only on initialization. As $n \to \infty$, all the coefficients converge almost surely to deterministic constants. We denote the limit by $f_\infty^{(t)}$.*

2. *The t-step loss $\mathcal{L}_n^{(t)}(\eta)$ converges almost surely to $\mathcal{L}_\infty^{(t)}(\eta) = \frac{1}{2m}\sum_{i=1}^m (f_\infty^{(t)}(\eta) - y_i)^2$ uniformly over η on any compact set. Moreover, there exists $\underline{\eta}, \bar{\eta} > 0$ such that $\operatorname{argmin}_{\eta\in[0,\infty)} \mathcal{L}_\infty^{(t)} \subset [\underline{\eta}, \bar{\eta}]$.*

3. *Assume that $\mathcal{L}_\infty^{(t)}$ has a unique minimizer $\eta_\infty^{(t)}$, let I be an arbitrary compact set containing $\eta_\infty^{(t)}$, and let $\eta_n^{(t)} \in \operatorname{argmin}_{\eta\in I} \mathcal{L}_n^{(t)}$. Then, as $n \to \infty$,*
$$\eta_n^{(t)} \to \eta_\infty^{(t)}, \quad a.s.$$

The proof of Theorem 3 is provided in Appendix B. The following sketch summarizes the proof machinery: the fact that $f^{(t)}(x)$ is a polynomial in η is straightforward. The convergence of the coefficients to deterministic limit follows from the "Master Theorem" in Yang and Hu (2021). This convergence implies that $\mathcal{L}_\infty^{(t)}$ is a polynomial with the leading monomial having a positive coefficient (quadratic loss). Therefore, the minimizer $\eta_\infty^{(t)}$ of $\mathcal{L}_\infty^{(t)}$ is finite which yields a probabilistic bound on $\eta_n^{(t)}$ for n large enough. We further show that the derivative of $\mathcal{L}_n^{(t)}(\eta)$ at $\eta = 0$ converges to a negative real number which bounds the minimizer (in η) away from 0. We conclude by observing that bounded roots of a converging sequence of polynomials converge to the roots of the limiting polynomial. Note that we show almost sure convergence, a much stronger convergence than convergence in probability or in $\mathbb{L}_2$ (almost sure convergence yields $\mathbb{L}_2$ convergence by Dominated Convergence Theorem). This stems from using almost sure convergence of scalar quantities from the Tensor Programs framework.

Theorem 3 shows that under the mild assumption that the limiting loss has a unique minimizer, LR transfer occurs under μP. This assumption is realistic as it is commonly observed in practice that training loss has a unique minimizer at any training step t.

Fig. 3 shows the same results of Fig. 2 at different training steps. With μP, we observe that optimal LR $\eta_n^{(1)}$ converges as width n grows for different training steps $t \in \{5, 10\}$, confirming the result of Theorem 4. Note that we consider small number of steps here because training converges after 10 to 15 iterations since the dataset is relatively simple (linear) and we use full batch GD. With SP, we observe a similar pattern to the one-step analysis; the optimal LR vanishes with width, and therefore optimal LR doesn't transfer with width in this case.

In the next section, we provide additional experiments with more challenging setups, including non-linear synthetic data, networks with ReLU activation function, varying depth, and varying optimizers.

5 ADDITIONAL EXPERIMENTS

We provide additional experiments to assess learning transfer with μP under several setups that are not necessarily covered by our theory. Our results shed light on the impact of the following factors: non-linearity (ReLU), network depth, training step, and optimizer.

Training data. We fix input dimension $d = 100$ in all experiments. We generate a ground truth vector $\omega \sim \mathcal{N}(0, d^{-1}I_d)$ and generate N inputs $x \sim \mathcal{N}(0, I_d)$ where $N = 1000$ is fixed. We generate N noise terms $\epsilon \sim \mathcal{N}(0, 0.01)$ and consider two output generating processes:

- *Linear*: the outputs are generated as $y = \omega^\top x + \epsilon$. This setup is used for the linear networks (no activation function).

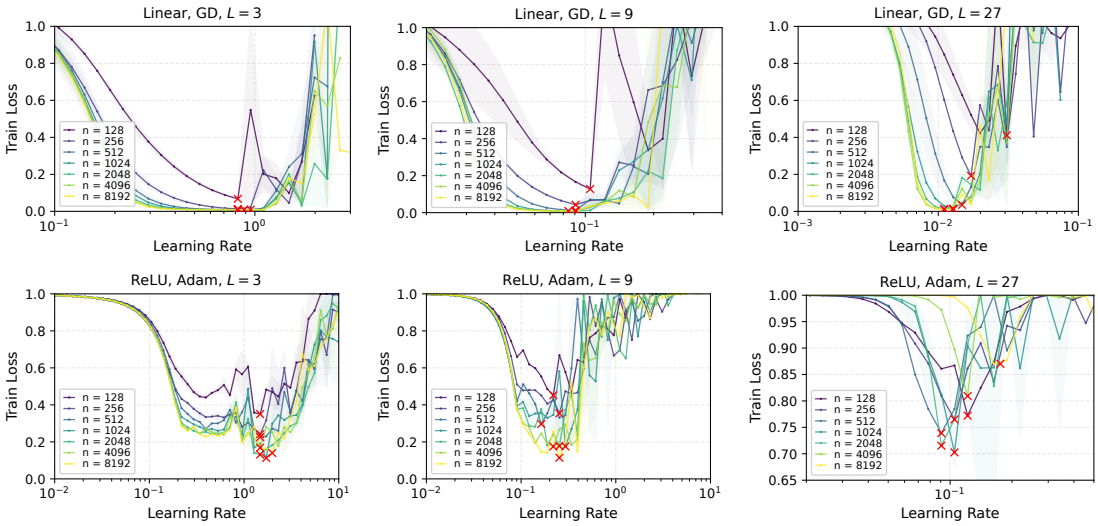

Figure 4: Train loss as a function of learning rate at $t = 20$ with 3 random seeds. Red crosses highlight the optimal LR for each width. **(Top)** Linear MLP of varying depth trained with SGD. **(Bottom)** MLP with ReLU activation of varying depth trained with Adam.

- *Non-linear*: the outputs are generated as $y = \text{Sign}(\omega^\top x + \epsilon)$, where $Sign(.)$ is the sign function ($+1$ if non-negative and -1 otherwise). This setup is used for neural networks with ReLU activation function.

We train MLPs with varying depths $L \in \{3, 9, 27\}$ and discuss the results below.

Impact of Depth. From Fig. 4, we observe that LR transfer occurs at different depths, confirming the result of Theorem 4 which holds for any depth. Interestingly, the optimal LR seems to decrease with depth, which confirms depth-dependency predicted by the result of Theorem 1 (see expression of $\eta_\infty^{(1)}$).[5]

ReLU and Adam. Fig. 4 shows that LR transfer holds for non-linear MLPs (with ReLU) trained with Adam. While our theory does not cover this case, empirical results suggest that LR transfer remains valid for non-linear architectures and more advanced training algorithms.

Impact of Training Step. Fig. 5 shows LR transfer also holds near convergence. Interestingly, the range of close-to optimal learning rates widens with the number of steps, suggesting that when the number of training steps is large enough, optimal LR has low resolution in the sense that choosing the right order of magnitude for the LR should be enough to obtain near-best performance.

[5]There a depth version of μP called Depth-μP, see Yang et al. (2023).

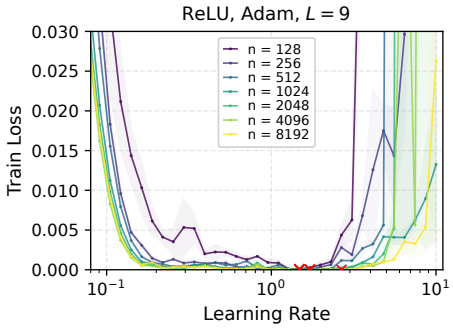

Figure 5: Train loss as a function of learning rate at $t = 100$ with 3 random seeds. MLP of depth $L = 9$ with ReLU activation trained with Adam.

6 DISCUSSION AND LIMITATIONS

We presented the first of learning rate transfer under μP. Our theoretical results rely on expressing the training loss of a deep linear network as a polynomial function of the learning rate. By studying the infinite-width limit, we derived convergence results for the optimal LR. While our results are limited to linear networks trained with GD, we believe they can be extended to non-linear MLPs and different optimizers. However, this will likely require different proof machinery especially when dealing when large-width deviations. We leave this question for future work.

ACKONWLEDGMENT

The author would like to thank Ruijia Zhang for their help with fixing some typos in this paper.

References

Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates, 2025. URL https://arxiv.org/abs/2504.05295.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization, 2019. URL https://arxiv.org/abs/1811.03962.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net, 2019. URL https://arxiv.org/abs/1904.11955.

Charlie Blake, Constantin Eichenberg, Josef Dean, Lukas Balles, Luke Y. Prince, Björn Deiseroth, Andres Felipe Cruz-Salinas, Carlo Luschi, Samuel Weinbach, and Douglas Orr. u-μp: The unit-scaled maximal update parametrization, 2025. URL https://arxiv.org/abs/2407.17465.

Blake Bordelon and Cengiz Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. In *Forty-second International Conference on Machine Learning*, 2025.

Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit, 2023. URL https://arxiv.org/abs/2309.16620.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming, 2020. URL https://arxiv.org/abs/1812.07956.

Lénaïc Chizat and Praneeth Netrapalli. The feature speed formula: a flexible approach to scale hyper-parameters of deep neural networks, 2025. URL https://arxiv.org/abs/2311.18718.

Lénaïc Chizat, Maria Colombo, Xavier Fernández-Real, and Alessio Figalli. Infinite-width limit of deep linear neural networks, 2022. URL https://arxiv.org/abs/2211.16980.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability, 2022. URL https://arxiv.org/abs/2103.00065.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, Dec 1989. ISSN 1435-568X. doi: 10.1007/BF02551274. URL https://doi.org/10.1007/BF02551274.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity, 2017. URL https://arxiv.org/abs/1602.05897.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks, 2018. URL https://arxiv.org/abs/1804.11271.

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.

Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A. Alemi, Roman Novak, Peter J. Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, and Jeffrey Pennington. Scaling exponents across parameterizations and optimizers, 2024. URL https://arxiv.org/abs/2407.05872.

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

Soufiane Hayou and Liyuan Liu. Optimal embedding learning rate in llms: The effect of vocabulary size, 2025. URL https://arxiv.org/abs/2506.15025.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2672–2680. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hayou19a.html.

Soufiane Hayou, Eugenio Clerico, Bobby He, George Deligiannidis, Arnaud Doucet, and Judith Rousseau. Stable resnet. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1324–1332. PMLR, 13–15 Apr 2021.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. Exact convergence rates of the neural tangent kernel in the large depth limit, 2022. URL https://arxiv.org/abs/1905.13654.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. ISSN 0893-6080. doi: https://doi.org/10.1016/0893-6080(89)90020-8. URL https://www.sciencedirect.com/science/article/pii/0893608089900208.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Atli Kosson, Jeremy Welborn, Yang Liu, Martin Jaggi, and Xi Chen. Weight decay may matter more than mup for learning rate transfer in practice, 2025. URL https://arxiv.org/abs/2510.19093.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes, 2018. URL https://arxiv.org/abs/1711.00165.

Mufan Li, Mihai Nica, and Dan Roy. The future is log-gaussian: Resnets and their infinite-depth-and-width limit at initialization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7852–7864. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/412758d043dd247bddea07c7ec558c31-Paper.pdf.

Lucas Lingle. An empirical study of μp learning rate transfer, 2025. URL https://arxiv.org/abs/2404.05728.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit, 2019. URL https://arxiv.org/abs/1902.06015.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12): 124008, December 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3a80. URL http://dx.doi.org/10.1088/1742-5468/ac3a80.

Radford M. Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*, pages 29–53. Springer New York, 1996. ISBN 978-0-387-94724-2. doi: 10.1007/978-1-4612-0745-0_2.

Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer, 2024. URL https://arxiv.org/abs/2402.17457.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos, 2025. URL https://arxiv.org/abs/2502.07529.

S.S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep information propagation. In *International Conference on Learning Representations*, 2017.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A law of large numbers, 2019. URL https://arxiv.org/abs/1805.01053.

Falcon Team. The falcon series of open language models, 2023. URL https://arxiv.org/abs/2311.16867.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Second edition, 2026. Pre-publication PDF dated February 19, 2026.

Christopher K. I. Williams. Computing with infinite networks. In *Neural Information Processing Systems*, 1996. URL https://api.semanticscholar.org/CorpusID:16883702.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

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

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite-depth neural networks, 2023. URL https://arxiv.org/abs/2310.02244.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training?, 2025. URL https://arxiv.org/abs/2410.21676.

Chenyu Zheng, Xinyu Zhang, Rongzhen Wang, Wei Huang, Zhi Tian, Weilin Huang, Jun Zhu, and Chongxuan Li. Scaling diffusion transformers efficiently via μp, 2025. URL https://arxiv.org/abs/2505.15270.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Proofs

A.1 Proof of Lemma 1

We prove the result for $m = 1$ (single sample dataset). Extending the result to general m is straightforward.

Lemma 3. *Assume $m = 1$. Then, for all $\ell \in \{2, 3, \dots, L\}$, we have $\|\phi_\ell\|_{L_2} = \mathcal{O}(n^{-(\ell-1)/2})$.*

Proof. Let $k \in \{2, \dots, L\}$. We show that all the terms inside ϕ_k are $\left(n^{-1/2}\right)$ which concludes the proof. Let $1 \le \ell_1 < \ell_2 < \dots < \ell_k \le L$. Then, we can write the summand as

$$\begin{aligned} &V^\top J_{L:\ell_k+1} b_{\ell_k+1} a_{\ell_k-1}^\top J_{\ell_k-1:\ell_{k-1}+1} \dots b_{\ell_1+1} a_{\ell_1-1}^\top J_{\ell_1-1:1} W_0 x \\ &\qquad = \|b_{\ell_k+1}\|^2 \|a_{\ell_1-1}\|^2 \prod_{j=2}^{k} a_{\ell_j-1}^\top J_{\ell_j-1:\ell_{j-1}+1} b_{\ell_{j-1}+1}. \end{aligned}$$

For some $j \in \{2, \dots, k\}$, let $J_j := J_{\ell_j-1:\ell_{j-1}+1}$. We have

$$a_{\ell_j-1}^\top J_{\ell_j-1:\ell_{j-1}+1} b_{\ell_{j-1}+1} = u^\top J_j^\top J_j J_j^\top v,$$

where $u = a_{\ell_{j-1}}$ and $v = b_{\ell_j}$.

Using Holder's inequality (product of k random variables) and Lemma 11, we obtain that

$$\mathbb{E}(V^\top J_{L:\ell_k+1} b_{\ell_k+1} a_{\ell_k-1}^\top J_{\ell_k-1:\ell_{k-1}+1} \dots b_{\ell_1+1} a_{\ell_1-1}^\top J_{\ell_1-1:1} W_0 x)^2 = \mathcal{O}(n^{-k+1}).$$

We conclude by observing that $\lim_{n\to\infty} \chi = -y$.

□

Proof for Lemma 1. Identical to Lemma 3: each inner product block has second moment $\Theta(n^{-1})$ by Lemma 11. Products of $k-1$ such factors contribute $\Theta(n^{-(k-1)})$ to the second moment; the extra sum over $i_r \in [m]$ only changes constants, not the n-scaling. The convergence of ϕ_1 is straightforward by Strong Law of Large Numbers (SLLN), and is a consequence of Lemma 4 below, which proves convergence of a kernel matrix to the Gram matrix K of input data.

A.2 Proof of Theorem 1

The proof proceeds as follows: we first characterize the infinite-width limit of ϕ_1, then we study the asymptotics of the loss function and conclude on the convergence of the optimal learning rate.

First-order term and a layerwise Gram matrix. Fox (x_j, y_j) in the training dataset, the degree one coefficient ϕ_1 in the expression of $f^{(1)}(x_j)$ as a polynomial in η is given by

$$\phi_1 \;=\; -\frac{1}{m} \sum_{\ell=1}^{L} \sum_{i=1}^{m} \chi_i \, \|b_{\ell+1}\|^2 \left\langle a_{\ell-1,i},\, a_{\ell-1,j} \right\rangle. \tag{4}$$

Let $G_{\ell-1} \in \mathbb{R}^{m\times m}$ be the layerwise Gram with $(G_{\ell-1})_{ij} = \langle a_{\ell-1,i}, a_{\ell-1,j}\rangle$, and define the normalized input Gram $K \in \mathbb{R}^{m\times m}, \quad K_{ij} = \langle x_i, x_j\rangle/d$. The next results characterizes the infinite-width limit of a kernel matrix from which the limit of ϕ_1 follows.

Lemma 4 (Layerwise Gram limit; m points)**.** *As $n \to \infty$,*

$$\frac{1}{L} \sum_{\ell=1}^{L} \|b_{\ell+1}\|^2 \, G_{\ell-1} \;\xrightarrow{a.s.}\; K.$$

Proof. For $\ell \in \{1, \dots, L\}$, we have $\mathbb{E}\|b_{\ell+1}\|^2 = 1/n$. The vectors $a_{\ell-1,i}$ are jointly Gaussian with per-coordinate covariance $\langle x_i, x_j\rangle/d$. Independence between $b_{\ell+1}$ and $(a_{\ell-1,i})_{i=1}^m$ gives $\mathbb{E}[\|b_{\ell+1}\|^2 G_{\ell-1}] = K$. A simple application of the SLLN implies the a.s. convergence of the layerwise average to K. □

Limiting one-step loss and optimal step size. Let $\chi = (\chi_1^{(1)}, \dots, \chi_m^{(1)})^\top$, $y = (y_1, \dots, y_m)^\top$. Using Lemma 1 and (4), uniformly for η on compact intervals,

$$\mathcal{L}_n(\eta) = \frac{1}{2m}\big\|\chi - \eta\, H_n\, \chi\big\|^2 + o_{\mathbb{L}_2}(1), \qquad H_n \;=\; \sum_{\ell=1}^{L} \frac{1}{m}\, \|b_{\ell+1}\|^2\, G_{\ell-1}. \tag{5}$$

By Lemma 4, $H_n \xrightarrow{\text{a.s.}} \frac{L}{m} K$, and since $\chi \to -y$ in $\mathbb{L}_2$ (as $f^{(0)}(x_i) \to 0$ in $\mathbb{L}_2$), we obtain the deterministic limit

$$\mathcal{L}_\infty^{(1)}(\eta) \overset{def}{=} \lim_{n\to\infty} \mathcal{L}_n^{(1)}(\eta) = \frac{1}{2m}\big\| - y + \eta\, \tfrac{L}{m}\, K\, y\big\|^2. \quad \text{a.s.} \tag{6}$$

The next result shows convergence of the optimal learning rate $\eta_n^{(1)}$.

Lemma 5 (LR transfer; limiting minimizer)**.** *Assume $Ky \neq 0$, then $\mathcal{L}_\infty^{(1)}(\eta)$ is strictly convex quadratic with the unique minimizer*

$$\eta_\infty^{(1)} \;=\; \frac{m}{L}\, \frac{y^\top K y}{\|Ky\|^2}. \tag{7}$$

Moreover, for any compact set $I \subset [0, \infty)$ containing $\eta_\infty^{(1)}$, we have for any $\eta_n^{(1)} \in \operatorname{argmin}_{\eta\in I} \mathcal{L}_n^{(1)}(\eta)$, $\eta_n^{(1)} \to \eta_\infty^{(1)}$ in $\mathbb{L}_2$.

Proof. The limiting loss (6) is a strictly convex quadratic in η whenever $Ky \neq 0$. Differentiating yields (7). Uniform convergence in $\mathbb{L}_2$ of $\mathcal{L}_n^{(1)} \to \mathcal{L}_\infty^{(1)}$ on compacts (in η) plus strict convexity implies convergence of minimizers. □

Particular case. When the inputs are orthogonal, i.e. if $\langle x_i, x_j\rangle = 0$ for $i \neq j$, then $K = \operatorname{diag}(k_1, \dots, k_m)$ with $k_i = \|x_i\|^2/d$, and

$$\eta_\infty^{(1)} = \frac{m}{L} \cdot \frac{\sum_{i=1}^m y_i^2 k_i}{\sum_{i=1}^m y_i^2 k_i^2}.$$

A.3 Convergence rate

As above, we assume $Ky \neq 0$ and work with the one–step loss

$$\mathcal{L}_n^{(1)}(\eta) \;=\; \frac{1}{2m} \sum_{j=1}^m \big(f^{(1)}(x_j) - y_j\big)^2$$

We also recall the limiting quadratic $\mathcal{L}_\infty^{(1)}(\eta) = \frac{1}{2m}\big\| - y + \eta\, \frac{L}{m} K y\big\|^2$ with unique minimizer $\eta_\infty^{(1)} = \frac{m}{L}\frac{y^\top K y}{\|Ky\|^2}$.

Let $\chi_\infty = (-y_1, \dots, -y_m)^\top$ and recall

$$H_n = \sum_{\ell=1}^{L} \frac{1}{m}\, \|b_{\ell+1}\|^2\, G_{\ell-1} \in \mathbb{R}^{m\times m}, \qquad (G_{\ell-1})_{ij} = \big\langle a_{\ell-1,i}, a_{\ell-1,j}\big\rangle.$$

Let us explicitly state the bounds (instead of $o(1)$ in the previous section) as these are needed to characterize the convergence rate.

Lemma 6 (One-step decomposition with uniform remainders)**.** *Fix any compact interval $I \subset (0, \infty)$. Then, uniformly in $\eta \in I$,*

$$\mathcal{L}_n(\eta) = \frac{1}{2m}\big\|\chi - \eta H_n \chi\big\|^2 + R_n(\eta), \tag{8}$$

where the remainder satisfies

$$\sup_{\eta\in I} \big|R_n(\eta)\big| = O_{\mathbb{L}_2}(n^{-1/2}), \qquad \sup_{\eta\in I} \big|R_n'(\eta)\big| = O_{\mathbb{L}_2}(n^{-1/2}).$$

Proof. The results follows Lemma 1. The term R_n collects all terms containing coefficients of monomial η^k with $k \geq 2$. By Lemma 1, for each $k \geq 2$ and j, $\|\phi_k\|_{L_2} = O(n^{-(k-1)/2})$; thus for fixed L and $\eta \in I$, $R_n(\eta)$ and $R_n'(\eta)$ are dominated by the $k = 2$ contribution and are $O_{\mathbb{L}_2}(n^{-1/2})$ uniformly on I. □

The next result characterizes the convergence rate of the effective kernel H_n to the infinite-width kernel K.

Lemma 7 (Convergence rates for χ and H_n)**.** *As $n \to \infty$,*

$$\max_{1\leq i\leq m} \left|f^{(0)}(x_i)\right|^2 = O_{\mathbb{L}_2}(n^{-1}), \qquad H_n = \tfrac{L}{m}K + O_{\mathbb{L}_2}(n^{-1/2}),$$

where the last equality holds element-wise.

Proof. First claim. For each i, conditionally on $a_{L,i}$, $f^{(0)}(x_i) = V^\top a_{L,i}$ is Gaussian with mean 0 and variance $\frac{1}{n^2}\|a_{L,i}\|^2$ since $V \sim \mathcal{N}(0, n^{-2}I_n)$ is independent of $a_{L,i}$. Taking expectations and using isotropy of the W_ℓ (so $\mathbb{E}\|a_{L,i}\|^2 = \|x_i\|^2$), we obtain $\mathbb{E}[f^{(0)}(x_i)^2] = \|x_i\|^2/n^2$, hence $|f^{(0)}(x_i)|^2 = O_{\mathbb{L}_2}(n^{-1})$. Since m is fixed, we can take the max over i.

Second claim. For $T_\ell \overset{def}{=} m^{-1}\|b_{\ell+1}\|^2 G_{\ell-1}$, independence of the "top" block ($b_{\ell+1}$) and the "bottom" block ($G_{\ell-1}$) implies $\mathbb{E}[T_\ell] = (1/m)K$ (as in Lemma 4). For any fixed (i, j),

$$(T_\ell)_{ij} = \frac{1}{m}\|b_{\ell+1}\|^2 \langle a_{\ell-1,i}, a_{\ell-1,j}\rangle.$$

Conditionally on the weights $W_{\ell-2}...W_0$, $\langle a_{\ell-1,i}, a_{\ell-1,j}\rangle$ is a sum of iid random variables with mean $n^{-1}\langle a_{\ell-2,i}, a_{\ell-2,j}\rangle$. Therefore,

$$\mathbb{E}\left[(n^{-1}\langle a_{\ell-1,i}, a_{\ell-1,j}\rangle - n^{-1}\langle a_{\ell-2,i}, a_{\ell-2,j}\rangle)^2 \mid W_{\ell-2}...W_0\right] = \mathcal{O}(n^{-1}).$$

Doing this recursively yields

$$\mathbb{E}\left[(n^{-1}\langle a_{\ell-1,i}, a_{\ell-1,j}\rangle - K_{ij})^2\right] = \mathcal{O}(n^{-1}),$$

which concludes the proof.

□

Lemma 8 (Uniform convergence and strong convexity)**.** *Fix compact $I \subset [0, \infty)$. Then*

$$\sup_{\eta\in I}\left|\mathcal{L}_n(\eta) - \mathcal{L}_\infty(\eta)\right| = O_{\mathbb{L}_2}(n^{-1/2}), \quad \sup_{\eta\in I}\left|\partial_\eta\mathcal{L}_n(\eta) - \partial_\eta\mathcal{L}_\infty(\eta)\right| = O_{\mathbb{L}_2}(n^{-1/2}),$$

and

$$\inf_{\eta\in I}\partial^2_{\eta\eta}\mathcal{L}_n(\eta) \xrightarrow{\mathbb{L}_2} \mu := \frac{L^2}{m^3}\, y^\top K^2 y > 0.$$

Proof. Using (8) and expanding the quadratic part,

$$\mathcal{L}_n(\eta) - \mathcal{L}_\infty(\eta) = \frac{1}{2m}\Big(\|\chi\|^2 - \|y\|^2 - 2\eta\big[\chi^\top H_n\chi - y^\top \tfrac{L}{m}Ky\big] + \eta^2\big[\chi^\top H_n^2\chi - y^\top \tfrac{L^2}{m^2}K^2y\big]\Big) + R_n(\eta).$$

By Lemma 7, $\mathbb{E}\max_i |f^0(x_i)|^2 = \mathcal{O}(n^{-1})$, hence $\chi = -y + O_{\mathbb{L}_2}(n^{-1/2})$. Also $H_n = (L/m)K + O_{\mathbb{L}_2}(n^{-1/2})$ coordinate wise (and thus in operator norm). Therefore each bracketed term above is $O_{\mathbb{L}_2}(n^{-1/2})$ uniformly on I, and $R_n(\eta) = O_{\mathbb{L}_2}(n^{-1/2})$ by Lemma 6, which proves the first result. Differentiating the decomposition gives the derivative bound by the same argument. Finally,

$$\partial^2_{\eta\eta}\mathcal{L}_n(\eta) = \frac{1}{m}\,\chi^\top H_n^2\chi + R_n''(\eta),$$

and the right-hand side converges in $\mathbb{L}_2$ to $(1/m)\, y^\top((L/m)K)^2 y$, uniformly on I. □

Lemma 9 (Rates for the argmin and for the loss at the argmin)**.** *Let $I \subset (0,\infty)$ be any compact interval containing $\eta_\infty^{(1)}$. Let $\eta_n^{(1)} \in \arg\min_{\eta\in I} \mathcal{L}_n(\eta)$. Then, as $n \to \infty$,*

$$\eta_n^{(1)} - \eta_\infty^{(1)} = O_{\mathbb{P}}(n^{-1/2}), \qquad \mathcal{L}_n(\eta_n^{(1)}) - \mathcal{L}_\infty(\eta_\infty^{(1)}) = O_{\mathbb{P}}(n^{-1/2}),$$

and

$$\mathcal{L}_\infty(\eta_n^{(1)}) - \mathcal{L}_\infty(\eta_\infty^{(1)}) = \frac{\mu}{2}\,(\eta_n^{(1)} - \eta_\infty^{(1)})^2 = \mathcal{O}_{\mathbb{P}}(n^{-1}).$$

Consequently, the loss gap at the argmin is dominated by the uniform $n^{-1/2}$ *error of $\mathcal{L}_n$ (the shift of the minimizer contributes only $O_{\mathbb{P}}(n^{-1})$).*

Proof. By Lemma 8, there exists (with high probability) a constant $c > 0$ such that $\inf_{\eta\in I} \mathcal{L}_n''(\eta) \geq c$ for all large n. Using the mean-value form of the optimality condition,

$$0 = \mathcal{L}_n'(\eta_n^{(1)}) = \mathcal{L}_n'(\eta_\infty^{(1)}) + \mathcal{L}_n''(\tilde{\eta}_n)\,(\eta_n^{(1)} - \eta_\infty^{(1)})$$

for some $\tilde{\eta}_n$ between $\eta_\infty^{(1)}$ and $\eta_n^{(1)}$. Hence

$$|\eta_n^{(1)} - \eta_\infty^{(1)}| \;\leq\; \frac{1}{c}\,|\mathcal{L}_n'(\eta_\infty^{(1)})| \;\leq\; \frac{1}{c}\Big(\sup_{\eta\in I}\big|\mathcal{L}_n'(\eta) - \mathcal{L}_\infty'(\eta)\big|\Big).$$

Using the fact that $\sup_{\eta\in I} |\mathcal{L}_n'(\eta) - \mathcal{L}_\infty'(\eta)| = O_{\mathbb{L}_2}(n^{-1/2})$ by Lemma 8 yields $\eta_n^{(1)} - \eta_\infty^{(1)} = O_{\mathbb{P}}(n^{-1/2})$.

For the loss at the argmin, write

$$\mathcal{L}_n(\eta_n^{(1)}) - \mathcal{L}_\infty(\eta_\infty^{(1)}) = \underbrace{\big(\mathcal{L}_n(\eta_\infty^{(1)}) - \mathcal{L}_\infty(\eta_\infty^{(1)})\big)}_{O_{\mathbb{P}}(n^{-1/2})} + \underbrace{\big(\mathcal{L}_\infty(\eta_n^{(1)}) - \mathcal{L}_\infty(\eta_\infty^{(1)})\big)}_{\text{shift term}}.$$

The first term is $O_{\mathbb{P}}(n^{-1/2})$ by Lemma 8. For the shift term, a Taylor expansion of $\mathcal{L}_\infty$ around $\eta_\infty^{(1)}$ gives

$$\mathcal{L}_\infty(\eta_n^{(1)}) - \mathcal{L}_\infty(\eta_\infty^{(1)}) = \tfrac{1}{2}\mathcal{L}_\infty''(\eta_\infty^{(1)})\,(\eta_n^{(1)} - \eta_\infty^{(1)})^2 = \frac{\mu}{2}\,(\eta_n^{(1)} - \eta_\infty^{(1)})^2,$$

and since $\eta_n^{(1)} - \eta_\infty^{(1)} = O_{\mathbb{P}}(n^{-1/2})$, this is $O_{\mathbb{P}}(n^{-1})$. So the dominant term is the $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$ above, which concludes the proof. □

A.4 Failure of LR Transfer under Standard Parametrizations

We consider Standard Parametrization where the different with μP lies only in how the head V is initialized: $V \sim \mathcal{N}(0, n^{-1})$, while $W_0 \sim \mathcal{N}(0, d^{-1})$ and $W_\ell \sim \mathcal{N}(0, n^{-1})$ for $\ell = 1, \ldots, L$. For the learning rate, we assume $c = 0$, i.e. the learning rate is parametrized as a constant $\eta > 0$.

We provide the proof for $m = 1$. Extending the result to $m \geq 1$ is straightforward. Let (x, y) be the training datapoint. At $t = 1$, the output is given by

$$f^{(1)}(x) = V^\top \left[\prod_{\ell=1}^{L}\Big(W_\ell^{(0)} - \eta\,\chi\, b_{\ell+1} a_{\ell-1}^\top\Big)\right] W_0 x,$$

where $\chi = f^{(0)}(x) - y$, which can be written as $f^{(1)}(x) = f^{(0)}(x) + \sum_{\ell=1}^{L} \phi_l \eta^\ell$.

With SP, it is straightforward to see that all coefficients ϕ_ℓ are of order $\sqrt{n}$ in L_2. It suffices to normalize V by $\sqrt{n}$ and we're essentially back to the case of μP with the same asymptotic analysis (Lemma 11).

Expressing the loss function as $\mathcal{L}_n^{(1)}(\eta) = (f^{(1)}(x) - y)^2 = (a_0 + a_1\eta + \cdots + a_L\eta^L)^2$, it is easy to check that this polynomial satisfies the conditions in Lemma 12, which yields the result.

B Proofs for Section 4

Lemma 2. [Non-linear behavior after step $t = 2$] *The limit of the coefficient $\phi_L(\eta)$ can be expressed as*

$$\lim_{n\to\infty} \phi_L(\eta) = (-m)^L \sum_{1\le i_1,i_2,\dots,i_L\le m} \zeta(i_1,i_2,\dots,i_L)\frac{\langle x_{i_1},x\rangle}{d},$$

where

$$\zeta(i_1,i_2,\dots,i_L) = \left(\prod_{j=1}^{L}\left(f_\infty^{(1)}(x_{i_j}) - y_{i_j}\right)\right)\left(\prod_{j=2}^{L} f_\infty^{(1)}(x_{i_j})\right),$$

with $f_\infty^{(1)}(x) = \eta\,\frac{L}{m}\sum_{i=1}^m y_i \frac{\langle x_i,x\rangle}{d}$.

The proof of Lemma 2 is straightforward by taking the infinite-width limit.

From Lemma 2, we obtain that $\phi_L(\eta)$ converges to a polynomial of degree $2L-1$ in η as n goes to infinity. Adding the η^L term in $f^{(2)}$, we obtain that $f^{(2)}$ converges to a polynomial that has a non-zero term of degree $3L-1$. Therefore, in contrast to step 1, step 2 involves more complex dependencies in η, and a full analysis of the minimum is non-trivial in this case. This complexity should be expected to increase with step t as gradient dependencies on η become more complex with t.

The next result shows convergence of $f^{(t)}(x)$ to a limiting polynomial $P^{(t)}$, with deterministic coefficients. This is a straightforward result from the convergence of constants in a Tensor Program.

Theorem 4. *Let $t \ge 1$ and $x \in \mathbb{R}^d$. Then, for any $K > 0$, there exists a polynomial $f_\infty^{(t)}$ with deterministic coefficients such that*

$$\lim_{n\to\infty} \sup_{\eta\in[0,K]} |f^{(t)}(x) - f_\infty^{(t)}(\eta)| = 0. \quad a.s.$$

Proof. Let $t \ge 1$ and $x \in \mathbb{R}^d$. $f^{(t)}(x)$ is a polynomial in η with coefficients that can be expressed via the Tensor Program framework. The convergence follows from Theorem 7.4 in Yang and Hu (2021). □

Note that the convergence can also be made uniform in input x living in compact sets. This is not useful here since we consider a finite training dataset.

We now state the formal LR transfer result and prove it.

Theorem 5 (HP Transfer for general t)**.** *Let $K = \left(\frac{\langle x_i,x_j\rangle}{d}\right)_{1\le i,j\le m}$ and $y = (y_1,y_2,\dots,y_m)^\top \in \mathbb{R}^m$, and assume that $Ky \ne 0$. Let $f_\infty^{(t)}$ be the limiting polynomial (in η) of $f^{(t)}(x)$ from the result above. Then, $\mathcal{L}_n^{(t)}(\eta)$ converges almost surely to $\mathcal{L}_\infty^{(t)}(\eta) = \frac{1}{2m}\sum_{i=1}^m (f_\infty^{(t)}(\eta) - y_i)^2$ uniformly over η in some arbitrary compact set. Moreover, there exists $\underline{\eta}$, $\bar{\eta} > 0$ such that $\operatorname{argmin}_{\eta\in[0,\infty)} f_\infty^{(t)} \subset [\underline{\eta},\bar{\eta}]$.*

Moreover, assume that $\mathcal{L}_\infty^{(t)}$ has a unique minimizer $\eta_\infty^{(t)}$, let $\gamma \gg \eta_\infty^{(t)}$ be an arbitrarily large constant, and let $\eta_n^{(t)} \in \operatorname{argmin}_{\eta\in[0,\gamma]} \mathcal{L}_n^{(t)}$. We have that

$$\lim_{n\to\infty} \eta_n^{(t)} = \eta_\infty^{(t)}, \quad a.s.$$

Proof. From Theorem 4, we know that $f^{(t)}(x)$ converges almost surely to $f_\infty^{(t)}$ on any compact set. The convergence of $\mathcal{L}^{(t)}$ follows.

Now looking at the limiting loss $\mathcal{L}_\infty^{(t)}$ as a polynomial in η, the leading monomial has positive coefficient because of the squared loss. Therefore $\lim_{\eta\to\infty} \mathcal{L}_\infty^{(t)}(\infty) = \infty$ which implies that there exists $\bar{\eta} > 0$ such that $\operatorname{argmin}_{\eta\in[0,\infty)]} \mathcal{L}_\infty^{(t)} \subset [0,\bar{\eta}]$.

Now, let us prove the existence of $\underline{\eta}$. Observe that $\mathcal{L}_\infty^{(t)}(0) = \frac{1}{2m}\sum_{i=1}^m y_i^2 > 0$. Moreover, from Lemma 10, we have that

$$\frac{\partial \mathcal{L}_\infty^{(t)}}{\partial \eta}\Big|_{\eta=0} = \frac{1}{m}\sum_{i=1}^m \frac{t\,L}{m}\sum_{j=1}^m y_j \frac{\langle x_j, x_i\rangle}{d}(-y_i) = -\frac{t\,L}{m^2}\, y^\top K y.$$

Under the assumption that $Ky \neq 0$, we have $\frac{\partial \mathcal{L}_\infty^{(t)}}{\partial \eta}\Big|_{\eta=0} < 0$. As a result, by continuity of $\mathcal{L}_\infty^{(t)}$ with respect to η, there exists a neighborhood of $\eta = 0$ that does not contain the minimizer of $\mathcal{L}_\infty^{(t)}$. In other words, there exists $\underline{\eta} > 0$ such that $(\operatorname{argmin}_{\eta\in[0,\infty)} \mathcal{L}_\infty^{(t)}) \cap [0,\underline{\eta}) = \emptyset$.

Finally, under the assumption that $\mathcal{L}_\infty^{(t)}$ has a unique minimizer in $(0,\infty)$, the convergence result follows from Theorem 6.

□

The next lemma characterizes the derivative of the infinite-width polynomial limit $f_\infty^{(t)}$ at $\eta = 0$. It is used in the proof of LR transfer for general t.

Lemma 10 (Derivative of $f^{(t)}$ at $\eta = 0$)**.** *Let $x \in \mathbb{R}^d$ and $t \geq 1$. We have the following*

$$\frac{\partial f_\infty^{(t)}}{\partial \eta}\Big|_{\eta=0} = \lim_{n\to\infty} \frac{\partial f^{(t)}}{\partial \eta}\Big|_{\eta=0} = \frac{t\,L}{m}\sum_{i=1}^m y_i \frac{\langle x_i, x\rangle}{d}, \quad a.s.$$

Proof. We can express the output as

$$f^{(t)}(x) = V^\top \left[\prod_{\ell=1}^L \left(W_\ell^{(0)} - \eta \sum_{s=0}^{t-1} m^{-1} \sum_{i=1}^m \chi_i^{(s)}\, b_{\ell+1}^{(s)} (a_{\ell-1,i}^{(s)})^\top\right)\right] W_0 x.$$

Expanding in η, we have

$$\chi_i^{(s)} = f^{(s)}(x_i) - y_i = f^{(0)}(x_i) - y_i + \eta \times \tilde{\chi}_i^{(s)}(\eta),$$

for some polynomial $\chi_i^{(s)}$. Similarly,

$$b_\ell^{(s)} = b_\ell^0 + \eta \tilde{b}_\ell^{(s)}(\eta),$$

and

$$a_\ell^{(s)} = a_\ell^0 + \eta \tilde{a}_\ell^{(s)}(\eta).$$

Therefore, we can express $f^{(t)}$ as follows

$$f^{(t)}(x) = V^\top \left[\prod_{\ell=1}^L \left(W_\ell^{(0)} - \eta\, t\, m^{-1} \sum_{i=1}^m \chi_i^{(0)}\, b_{\ell+1}^{(0)} (a_{\ell-1,i}^{(0)})^\top + \eta^2 \Psi_\ell(\eta)\right)\right] W_0 x,$$

where Ψ_ℓ is a polynomial in η. It follows that

$$\frac{\partial f^{(t)}}{\partial \eta}\Big|_{\eta=0} = -\frac{t}{m}\sum_{\ell=1}^L V^\top J_{\ell+1}^{(0)} \sum_{i=1}^m \chi_i^{(0)} b_{\ell+1}^{(0)} (a_{\ell,i}^{(0)})^\top W_0 x.$$

Taking the width n to infinity yields the desired result, with almost sure convergence. □

The next result is used in the proof of LR transfer for general step t. It shows the almost sure convergence of the argmin of a polynomial under some conditions.

Theorem 6 (Argmin stability with a.s. coefficient convergence and positive polynomials)**.** *Fix an integer $p \geq 1$. For each $n \geq 1$, let*

$$P_n(x) = \sum_{k=0}^{p} a_{n,k}\, x^k, \qquad x \in [0, \infty),$$

where the coefficients $a_{n,k}$ are real-valued random variables on a common probability space. Assume there exist deterministic reals $(a_k)_{k=0}^{p}$ such that, for every $k = 0, \ldots, p$,

$$a_{n,k} \xrightarrow[n\to\infty]{a.s.} a_k,$$

and set the (deterministic) limit polynomial

$$P_\infty(x) = \sum_{k=0}^{p} a_k x^k.$$

Suppose:

(1) For each n, $P_n(x) \geq 0$ for all $x \geq 0$ almost surely.

(2) P_∞ has a unique minimizer $x_\star \in [0, \infty)$.

Then, for any constant $R > 0$, and for any $x_n \in \operatorname{argmin}_{[0,R]} P_n$ we have

$$x_n \xrightarrow{a.s.} x_\star.$$

Proof. Let Ω_0 be the probability-one event on which $a_{n,k} \to a_k$ for all k and $P_n(x) \geq 0$ for all $x \geq 0$ and all n. Let's fix $\omega \in \Omega_0$ and argue deterministically.

(i) Uniform convergence on compacts: For any $R > 0$, we have

$$\sup_{x\in[0,R]} |P_n(x) - P_\infty(x)| \leq \sum_{k=0}^{p} |a_{n,k} - a_k|\, R^k \xrightarrow[n\to\infty]{} 0,$$

so $P_n \to P$ uniformly on every compact subset of $[0, \infty)$.

(ii) Convergence of minimizers. Let $R > 0$. By uniqueness, for each $\delta > 0$ the compact set $K_\delta = \{x \in [0, R] : |x - x_\star| \geq \delta\}$ satisfies

$$\Delta_\delta \overset{def}{=} \min_{x\in K_\delta} \big(P(x) - P(x_\star)\big) > 0.$$

Uniform convergence on $[0, R]$ yields n_δ with $\sup_{x\in[0,R]} |P_n(x) - P(x)| \leq \Delta_\delta/3$ for all $n \geq n_\delta$. Thus, for $n \geq \max\{N, n_\delta\}$ and $x \in K_\delta$,

$$P_n(x) \geq P(x) - \tfrac{\Delta_\delta}{3} \geq P(x_\star) + \tfrac{2\Delta_\delta}{3} \geq P_n(x_\star) + \tfrac{\Delta_\delta}{3},$$

so no minimizer lies in K_δ, i.e. $|x_n - x_\star| < \delta$. As $\delta > 0$ is arbitrary, $x_n \to x_\star$. Since $\omega \in \Omega_0$ was arbitrary, the convergence holds almost surely. $\square$

C Technical Lemmas

The following lemma is used in the proofs of 1-step convergence results.

Lemma 11. *Let $L \geq 1$ be fixed. For $\ell = 1, \ldots, L$, let $W^{(\ell)} \in \mathbb{R}^{n\times n}$ have i.i.d. entries with $\mathbb{E}W_{ij}^{(\ell)} = 0$ and $\mathbb{E}(W_{ij}^{(\ell)})^2 = n^{-1}$, and assume the entries are uniformly sub-gaussian. Assume the matrices $\{W^{(\ell)}\}_{\ell=1}^{L}$ are independent. Let*

$$J := W^{(L)}W^{(L-1)}\cdots W^{(1)} \in \mathbb{R}^{n\times n}.$$

Let $x, y \in \mathbb{R}^n$ be independent of $\{W^{(\ell)}\}_{\ell=1}^{L}$, with i.i.d. coordinates of zero mean, unit variance, and uniformly sub-gaussian. Set

$$S := x^\top J^\top J J^\top y, \qquad A := J^\top J J^\top.$$

Then for every fixed $p > 0$ there exists a constant $C_{p,L} < \infty$ such that, for all sufficiently large n,

$$\mathbb{E}|S|^p \;\leq\; C_{p,L}\, n^{p/2}, \qquad \textit{equivalently} \qquad \|S\|_{\mathbb{L}^p} \;\leq\; C_{p,L}\sqrt{n}.$$

Proof. Constants may depend on p, L, and the sub-gaussian parameters, but not on n. Let $S = x^\top Ay$ with $A = J^\top JJ^\top$.

Step 1: $\mathbb{L}^p$ bound for $x^\top Ay$. Everything in this step is conditioned on J (so A is deterministic). Fix $p \geq 1$ and condition on y. Writing $v := Ay$, we have $S = \sum_{i=1}^{n} v_i x_i$. By the subgaussian Khintchine inequality (Vershynin, 2026, Prop. 2.7.5) and $\psi_2 \Rightarrow \mathbb{L}^p$ (Vershynin, 2026, Prop. 2.6.6(ii)),

$$\|S\|_{\mathbb{L}^p \,|\, y} \leq C\sqrt{p}\,\|Ay\|_2. \tag{9}$$

Taking $\mathbb{L}^p$ in y gives

$$\|S\|_{\mathbb{L}^p} \leq C\sqrt{p}\,\big\|\|Ay\|_2\big\|_{\mathbb{L}^p}. \tag{10}$$

Next, by anisotropic concentration of the norm (Vershynin, 2026, Ex. 6.13) (applied with $B = A$ and $X = y$),

$$\big\|\|Ay\|_2 - \|A\|_F\big\|_{\psi_2} \leq C\|A\|_{\mathrm{op}}.$$

Using again (Vershynin, 2026, Prop. 2.6.6(ii)), we obtain

$$\big\|\|Ay\|_2\big\|_{\mathbb{L}^p} \leq \|A\|_F + C\sqrt{p}\,\|A\|_{\mathrm{op}}. \tag{11}$$

Combining (10) and (11) yields

$$\|S\|_{\mathbb{L}^p} \leq C\Big(\sqrt{p}\,\|A\|_F + p\,\|A\|_{\mathrm{op}}\Big). \tag{12}$$

For $p \in (0, 1)$, use $\|Z\|_{\mathbb{L}^p} \leq \|Z\|_{\mathbb{L}^1}$.

Step 2: $\mathbb{L}^p$ bound for $\|A\|_F$. We have $\|A\|_F^2 = \mathrm{Tr}((J^\top J)^3)$. Since $\mathrm{Tr}(B) \leq n\|B\|_{\mathrm{op}}$ for $B \succeq 0$,

$$\|A\|_F^2 = \mathrm{Tr}\left((J^\top J)^3\right) \leq n\,\|(J^\top J)^3\|_{\mathrm{op}} = n\,\|J^\top J\|_{\mathrm{op}}^3 = n\,\|J\|_{\mathrm{op}}^6,$$

hence

$$\|A\|_F \leq \sqrt{n}\,\|J\|_{\mathrm{op}}^3. \tag{13}$$

By the subgaussian operator-norm bound for i.i.d. matrices (applied to $\sqrt{n}\,W^{(\ell)}$) and tail-to-moment integration (e.g. (Vershynin, 2026, Thm. 4.4.3 and Lem. 1.6.1)), for every $q \geq 1$ we have $\sup_n \mathbb{E}\|W^{(\ell)}\|_{\mathrm{op}}^q < \infty$. Using submultiplicativity and independence,

$$\|J\|_{\mathrm{op}} \leq \prod_{\ell=1}^{L} \|W^{(\ell)}\|_{\mathrm{op}} \qquad \Rightarrow \qquad \sup_n \mathbb{E}\|J\|_{\mathrm{op}}^q < \infty \quad \text{for every fixed } q \geq 1.$$

Therefore, taking $q = 3p$ in (13) gives

$$\big\|\|A\|_F\big\|_{\mathbb{L}^p} \leq \sqrt{n}\,\big\|\|J\|_{\mathrm{op}}^3\big\|_{\mathbb{L}^p} = \sqrt{n}\,\|J\|_{\mathbb{L}^{3p}}^3 \leq C_{p,L}\sqrt{n}.$$

Step 3: $\mathbb{L}^p$ upper bound for S. From (12) and the trivial bound $\|A\|_{\text{op}} \leq \|A\|_F$, we get

$$\|S\|_{\mathbb{L}^p} \leq C(\sqrt{p}+p)\left\|\|A\|_F\right\|_{\mathbb{L}^p}.$$

By Step 2 and uniform integrability of polynomial spectral statistics, for each fixed $p > 0$,

$$\left\|\|A\|_F\right\|_{\mathbb{L}^p} \leq C'_{p,L}\sqrt{n}.$$

Therefore $\|S\|_{\mathbb{L}^p} \leq C''_{p,L}\sqrt{n}$. □

The next lemma is used in the proof of the 1-step result for SP.

Lemma 12 (Lemma for SP)**.** *Let $P(\eta) = a_0 + a_1\eta + a_2\eta^2 + \cdots + a_L\eta^L$ be a polynomial where the coefficients $a_0, a_1, \ldots, a_L$ are random variables satisfying the following conditions:*

1. *$E[a_0^2] = O(1)$ and a_0 converges weakly to some random variable $\bar{a}_0$ of order 1 in distribution as $n \to \infty$.*
2. *$E[a_i^2] = O(n)$ for $i = 1, \ldots, L$, and $a_1/\sqrt{n}$ converges in $\mathbb{L}_2$ to a deterministic constant $\bar{b}_1 \neq 0$ as $n \to \infty$, with $a_1/\sqrt{n} = \bar{b}_1 + \mathcal{O}_{\mathbb{L}_2}(n^{-1/2})$.*

Let $K > 0$ be a constant and η_n be a minimizer of $P(\eta)^2$ on $[0, K]$, i.e., $\eta_n \in \arg\min_{\eta\in[0,K]} P(\eta)^2$. Then, η_n converges to 0 in probability as $n \to \infty$.

Proof. The proof proceeds by rescaling the domain of the polynomial to analyze its behavior in a neighborhood of 0, similar to the treatment of the μP case.

Consider the change of variables $\eta = \beta/\sqrt{n}$. Let η_n be a minimizer of $P(\eta)^2$. The corresponding minimizer in the β domain is $\beta_n = \eta_n\sqrt{n}$.

We now prove that the sequence of random variables $\{\hat{\beta}_n\}$ is bounded in probability, i.e. $\beta_n = O_p(1)$. This will imply the convergence of η_n.

Let's define a new sequence of random polynomials in the variable β by substituting $\eta = \beta/\sqrt{n}$ into $P(\eta)$

$$R_n(\beta) = P(\beta/\sqrt{n}) = a_0 + a_1\frac{\beta}{\sqrt{n}} + a_2\frac{\beta^2}{(\sqrt{n})^2} + \cdots + a_L\frac{\beta^L}{(\sqrt{n})^L}$$

Define a new set of coefficients $b_i^{(n)} = a_i/\sqrt{n}$ for $i \geq 1$. We can now rewrite the rescaled polynomial as

$$R_n(\beta) = a_0 + b_1^{(n)}\beta + b_2^{(n)}\frac{\beta^2}{\sqrt{n}} + b_3^{(n)}\frac{\beta^3}{n} + \cdots + b_L^{(n)}\frac{\beta^L}{n^{(L-1)/2}}$$

For any fixed $\beta \in \mathbb{R}$, as $n \to \infty$, every term for $i \geq 2$ converges to zero in $\mathbb{L}_2$. For instance, for the term $i = 2$, we have $b_2^{(n)}\beta^2/\sqrt{n} \xrightarrow{L_2} 0$ because $b_2^{(n)}$ is bounded in $\mathbb{L}_2$. This holds for all $\ell \in \{2, \ldots, L\}$.

Therefore, the sequence of random polynomials $R_n(\beta)$ in asymptotically controlled as follows

$$R_n(\beta) - R(\beta) = O_{\mathbb{L}_2}(n^{-1/2}),$$

where $R(\beta) = a_0 + b_1\beta$.

Let $\beta_n^* \in \text{argmin}_{\eta\in[0,K]} R_n(\beta)^2$ for K large enough (so that the global minimizer is covered). The second derivative of $R_n(.)^2$ is given by $2R_n''R_n + 2(R_n')^2$. We know that uniformly on $[0, K]$, $R_n''(\beta) = o_{\mathbb{L}_2}(1)$, and $R_n'(\beta) = b_1^{(n)} + \mathcal{O}_{\mathbb{L}_2}(n^{-1/2})$. Therefore, uniformly over $\beta \in [0, K]$, we have that $(R_n(\beta)^2)'' = 2(b_1^{(n)})^2 + \mathcal{O}_{\mathbb{L}_2}(n^{-1/2})$ $= 2\,\bar{b}_1^2 + \mathcal{O}_{\mathbb{L}_2}(n^{-1/2})$.

As a result, as $n \to \infty$, with high probability, there exists a constant $c > 0$ such that $\inf_{[0,K]}(R_n(\beta)^2)'' \geq c$. Using the Intermediate Value Theorem, we have that

$$|\beta_n^*| = |\beta_n^* - 0| \leq \frac{|(R_n^2)'(0)|}{c} = \frac{|b_1^{(n)}a_0|}{c}.$$

Which shows that $\beta_n^* = \mathcal{O}_{\mathbb{P}}(1)$ and concludes the proof. □