

# WEIGHT DECAY MAY MATTER MORE THAN $\mu$ P FOR LEARNING RATE TRANSFER IN PRACTICE

Atli Kosson,<sup>1,2†</sup> Jeremy Welborn,<sup>1</sup> Yang Liu,<sup>1</sup> Martin Jaggi,<sup>2</sup> Xi Chen<sup>1</sup>

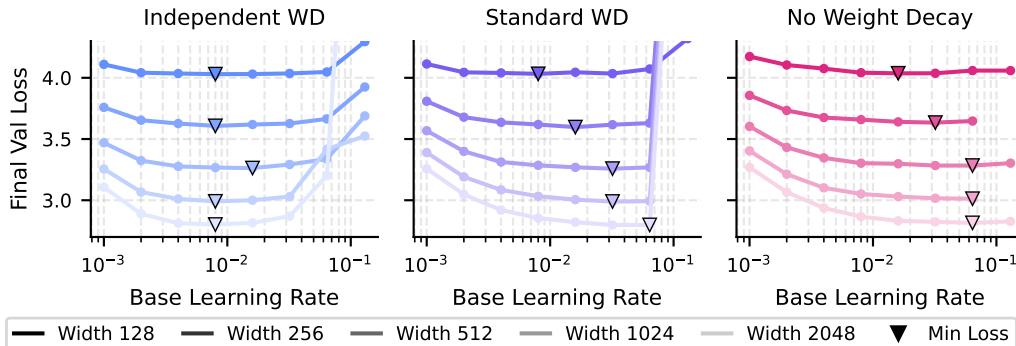
<sup>1</sup>Amazon FAR (Frontier AI & Robotics), <sup>2</sup>EPFL

## ABSTRACT

Transferring the optimal learning rate from small to large neural networks can enable efficient training at scales where hyperparameter tuning is otherwise prohibitively expensive. To this end, the Maximal Update Parameterization ( $\mu$ P) proposes a learning rate scaling designed to keep the update dynamics of internal representations stable across different model widths. However, the scaling rules of  $\mu$ P rely on strong assumptions, particularly about the geometric alignment of a layer’s inputs with both its weights and gradient updates. In this large-scale empirical investigation, we show that these assumptions hold only briefly at the start of training in the practical setups where learning rate transfer is most valuable, such as LLM training. For the remainder of training it is weight decay rather than  $\mu$ P that correctly stabilizes the update dynamics of internal representations across widths, facilitating learning rate transfer. This suggests  $\mu$ P’s scaling primarily acts as a form of implicit learning rate warmup, allowing us to largely replace it with modified warmup schedules. Together these findings fundamentally challenge prevailing beliefs about learning rate transfer and can explain empirical practice such as why  $\mu$ P requires the independent weight decay variant for successful transfer.

## 1 INTRODUCTION

The Maximal Update Parameterization ( $\mu$ P) has become a cornerstone of efficient training at scale, particularly for large language models (LLMs). Many open recipe LLMs have used variants of  $\mu$ P (Dey et al., 2023; Hu et al., 2024; Liu et al., 2023; Falcon-LLM Team, 2025; Cohere et al., 2025), and some commercial models with unknown training details likely use similar learning rate (LR) scaling approaches. For example, Llama4 used an internal technique called MetaP, Grok-2 configuration files suggest  $\mu$ P use, and the GPT-4 report (Achiam et al., 2023) cites a key  $\mu$ P paper coauthored by OpenAI staff (Yang et al., 2021). It has also impacted theory, showing how feature learning can occur in infinitely wide networks, contrasting with the neural tangent kernel regime (Jacot et al., 2018).



**Figure 1:  $\mu$ P specifically requires independent weight decay for good learning rate transfer** when pre-training LLaMA networks with AdamW (20B tokens, width 2048 and 1B params, see Appendix F.3). Independent weight decay achieves this by counteracting  $\mu$ P’s update scaling later in training which is needed because the core assumptions  $\mu$ P’s scaling is based on quickly break down. Standard weight decay does not counteract  $\mu$ P’s scaling and consequently results in a poor learning rate transfer on longer experiments. No weight decay behaves somewhere in-between, see Appx. B.

<sup>†</sup>Work done while interning at Amazon FAR

---

Despite this success, several works (Wortsman et al., 2024; Blake et al., 2025; Wang & Aitchison, 2025; Bergsma et al., 2025) have empirically found that learning rates only transfer well in practice when  $\mu P$  is combined with the independent variant of weight decay (WD), see Figure 1. In the conventional AdamW formulation (Algorithm 1), weight decay multiplies the weights by  $1 - \eta\lambda$  at each step where  $\eta$  is the learning rate and  $\lambda$  the weight decay coefficient. With independent weight decay, the weights are decayed by  $1 - \lambda$  which is not affected by  $\mu P$ 's learning rate scaling. Equivalently,  $\lambda$  can be scaled to keep the product  $\eta\lambda$  constant as  $\eta$  is scaled in the formulation above.

In this work, we demonstrate that independent weight decay is not merely a helpful addition but a central mechanism for enabling learning rate transfer in practice. We begin by developing a unifying framework for weight decay and  $\mu P$  based on relative updates, the size of updates as a fraction of the current weights or features (Section 2). Within this framework, we show that independent weight decay directly counteracts and ultimately overrides  $\mu P$ 's scaling during training (Section 3). We then show why this is crucial: for the majority of training, it is independent weight decay, not  $\mu P$ , that correctly stabilizes feature learning across different model widths (Section 4). This stabilization of the feature learning seems to be what facilitates learning rate transfer and is a key objective of  $\mu P$ .

The need to override  $\mu P$ 's prescribed update scaling arises because the underlying alignment assumptions fail to hold in practice. We show this can happen when the batch size is large relative to the model width, something that commonly happens in practice unlike the infinite-width limit that inspired  $\mu P$  (Section 5). This analysis fundamentally reframes  $\mu P$ 's practical role as an implicit learning rate warmup. We empirically measure this effect (Section 6) and demonstrate that the practical benefit of  $\mu P$  can largely be replicated by using stronger, explicit warmup schedules (Section 7). Together, our findings challenge prevailing beliefs about learning rate transfer and provide guidance for achieving robust transfer in practice.

## 2 A UNIFYING FRAMEWORK FOR $\mu P$ AND WEIGHT DECAY

### 2.1 SINGLE LAYER MODEL AND REPRESENTATION CHANGES

The core  $\mu P$  and weight decay concepts discussed in this work can be understood from the linear transformation of a fully connected layer in isolation. For an input feature dimension  $C$ , output feature dimension  $K$  and batch size  $B$ , we define this as:

$$\mathbf{Y} = \mathbf{W}\mathbf{X}, \quad \mathbf{X} \in \mathbb{R}^{C \times B} \text{ (inputs)}, \quad \mathbf{W} \in \mathbb{R}^{K \times C} \text{ (weights)}, \quad \mathbf{Y} \in \mathbb{R}^{K \times B} \text{ (outputs)} \quad (1)$$

A weight update  $\mathbf{W} \mapsto \mathbf{W} + \Delta\mathbf{W}$  causes the layer's outputs to change as  $\mathbf{Y} \mapsto \mathbf{Y} + \Delta\mathbf{Y}$  for the same input  $\mathbf{X}$ . Note that in the context of neural networks,  $\Delta\mathbf{Y}$  refers to the *local* representation change from this layer alone, the *total* change would include potential changes  $\Delta\mathbf{X}$  in the inputs arising from the updates of earlier layers. Local changes can be related to total changes as shown by Yang et al. (2023), but we will focus on local ones in this work for simplicity.

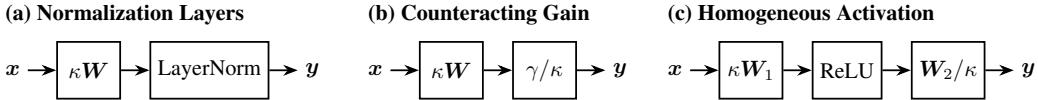
### 2.2 ALIGNMENT RELATES WEIGHT MAGNITUDES TO REPRESENTATION SIZES

Central to this work and  $\mu P$  in general is connecting the size of the *representation changes*  $\|\Delta\mathbf{Y}\|$  to the size of weight updates  $\|\Delta\mathbf{W}\|$ . We measure this in terms of the Frobenius norm  $\|\cdot\| := \|\cdot\|_F$ . It is easy for an optimizer to give a specific weight update size  $\|\Delta\mathbf{W}\|$ , Adam (Kingma & Ba, 2015) does this approximately and sign-based optimizers like Lion (Chen et al., 2023) can do this perfectly. However it is ultimately the size of the representation changes  $\|\Delta\mathbf{Y}\|$  that determine the impact of an update. For example, a weight update that does not change  $\mathbf{Y}$  at all has no immediate impact. However, controlling  $\|\Delta\mathbf{Y}\|$  is much harder than  $\|\Delta\mathbf{W}\|$  and is the core objective of  $\mu P$ 's LR scaling.

We describe the relationship between  $\|\Delta\mathbf{Y}\|$  and  $\|\Delta\mathbf{W}\|$  via the *update alignment*, defined as:

$$\alpha_{\Delta\mathbf{W}} := \frac{\|\Delta\mathbf{Y}\|}{\|\Delta\mathbf{W}\|\|\mathbf{X}\|} = \frac{\|\Delta\mathbf{W}\mathbf{X}\|}{\|\Delta\mathbf{W}\|\|\mathbf{X}\|} = \sqrt{\sum_{k,b} \frac{\|\Delta\mathbf{w}_k\|^2 \|\mathbf{x}_b\|^2}{\|\Delta\mathbf{W}\|^2 \|\mathbf{X}\|^2} \cos^2 \angle(\Delta\mathbf{w}_k, \mathbf{x}_b)} \in [0, 1] \quad (2)$$

This is a weighted root-mean-square average of the cosine similarities between input samples  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_B]$  and rows in the weight update  $\Delta\mathbf{W} = [\Delta\mathbf{w}_1, \dots, \Delta\mathbf{w}_K]^\top$ . Note that a high update alignment  $\alpha_{\Delta\mathbf{W}}$  means that  $\mathbf{w}_k$  and  $\mathbf{x}_b$  tend to point in similar directions and will result in larger representation changes  $\|\Delta\mathbf{Y}\|$  for a given update size  $\|\Delta\mathbf{W}\|$ . We similarly define the *weight alignment* as  $\alpha_{\mathbf{W}} := \frac{\|\mathbf{Y}\|}{\|\mathbf{W}\|\|\mathbf{X}\|}$  which relates the weight magnitude  $\|\mathbf{W}\|$  to representation size  $\|\mathbf{Y}\|$ .



**Figure 2: Neural networks are commonly invariant to certain types of rescaling transformations.** Examples where applying an in-place scaling factor  $\kappa > 0$  does not affect the current block output  $y$ . Across such transformations, an update with a fixed absolute size  $\|\mathbf{W}\|$  has a varying impact on the output while a given relative change  $\|\Delta\mathbf{W}\|/\|\mathbf{W}\|$  always has the same impact. This makes the relative change a more informative measure of the size of updates that better captures their impact.

### 2.3 WEIGHT DECAY MODULATES THE RELATIVE SIZE OF WEIGHT UPDATES

Weight decay is widely thought to mainly function as regularizer despite a long line of work that argues this is not the case in deep learning (Van Laarhoven, 2017; Zhang et al., 2019; Chiley et al., 2019; Li & Arora, 2020; Li et al., 2020; Wan et al., 2021; Li et al., 2022; Kosson et al., 2024b; D’Angelo et al., 2024). Instead it primarily acts as a secondary learning rate hyperparameter that modulates the size of relative weight updates  $\|\Delta\mathbf{W}\|/\|\mathbf{W}\|$ . Kosson et al. (2024b) describe these effects in AdamW in detail. In the standard variant (Algorithm 1), weight decay will cause the weight norms to converge to a fixed *equilibrium* value given by  $\|\mathbf{W}\| \approx \sqrt{KC \cdot \eta/\lambda}$  for learning rate  $\eta$ , weight decay  $\lambda$  and matrix size  $K \times C$ . At this value, the shrinking effect of weight decay and growth effect of gradient updates balance out, keeping the norm constant in expectation. AdamW’s normalization of the weight updates makes  $\|\Delta\mathbf{W}\| \propto \eta\sqrt{KC}$ , which in turn gives relative updates  $\|\Delta\mathbf{W}\|/\|\mathbf{W}\| \propto \sqrt{\eta\lambda}$ . Crucially, note that the learning rate  $\eta$  and weight decay  $\lambda$  affect the relative updates in equilibrium the same way, only their product  $\eta\lambda$  matters. In summary:

$$\|\mathbf{W}\| \propto \sqrt{KC\eta/\lambda}, \quad \frac{\|\Delta\mathbf{W}\|}{\|\mathbf{W}\|} \propto \sqrt{\eta\lambda} \quad (\text{equilibrium effect of weight decay}) \quad (3)$$

Due to the structure of neural networks, relative changes are often more meaningful than absolute changes (Bernstein et al., 2020; Kosson et al., 2024b;a). A given relative update size  $\|\Delta\mathbf{W}\|/\|\mathbf{W}\|$  has the same impact across a number of equivalent parameter states obtained via transformations like those shown in Figure 2, unlike a fixed absolute update size  $\|\Delta\mathbf{W}\|$ . A similar argument extends to the representations, a given change  $\Delta\mathbf{Y}$  typically has a greater impact when  $\|\mathbf{Y}\|$  is small than when it is large. Rescaling transformations do not occur in standard training but weight decay can have a similar overall effect by changing the weight norms over time while normalization layers or learnable gains can undo the effects of this on the output. In this work we focus on relative changes rather than absolute ones based on their informativeness and connection to weight decay, deviating from existing work on  $\mu\mathbf{P}$ . Note that at the start of training these approaches are equivalent since  $\|\mathbf{W}\|$  and  $\|\mathbf{Y}\|$  are roughly determined by the initialization. However, later in training they can differ significantly.

### 2.4 FORMULATING $\mu\mathbf{P}$ IN TERMS OF RELATIVE UPDATES

The primary design goal of  $\mu\mathbf{P}$  is to ensure stable and non-trivial feature learning at any width. Yang et al. (2023) formulate this as  $\|\mathbf{Y}\|_{\text{RMS}} := \|\mathbf{Y}\|_F/\sqrt{KB} = \Theta(1)$  and  $\|\Delta\mathbf{Y}\|_{\text{RMS}} = \Theta(1)$ , i.e., the representations and their changes neither explode nor vanish as the input dimension  $C$  grows.<sup>1</sup> In our relative formulation we want to keep  $\|\Delta\mathbf{Y}\|/\|\mathbf{Y}\|$  roughly constant across widths for a given learning rate. The hope is that a fixed relationship between learning rates and representation changes will result in a transfer of the optimal learning rate, but this is ultimately a heuristic without guarantees. We can express the relative representation change in terms of the relative weight update as follows:<sup>2</sup>

$$\frac{\|\Delta\mathbf{Y}\|}{\|\mathbf{Y}\|} = \frac{\alpha_{\Delta\mathbf{W}}}{\alpha_{\mathbf{W}}} \frac{\|\Delta\mathbf{W}\|}{\|\mathbf{W}\|} \quad (4)$$

where we call  $\alpha_{\Delta\mathbf{W}}/\alpha_{\mathbf{W}}$  the *alignment ratio*. To stabilize the relative representation changes  $\|\Delta\mathbf{Y}\|/\|\mathbf{Y}\|$  across different widths  $C$ ,  $\mu\mathbf{P}$  must perfectly counter changes in the alignment ratio by

<sup>1</sup>Technically it is the total representation change across training that matters, but Yang et al. (2023) show this can follow from bounded local changes with assumptions about the spectral norms of weights and updates.

<sup>2</sup>We assume matrices are non-zero which is typically the case throughout training, barring special initialization.

---

scaling the weight updates appropriately (via the learning rate). This requires knowing how both the update alignment  $\alpha_{\Delta\mathbf{W}}$  and weight alignment  $\alpha_{\mathbf{W}}$  behave as a function of the network width  $C$ .

At the start of training the weights are randomly initialized. If we assume the weight vector  $\mathbf{w} \in \mathbb{R}^C$  has independent and identically distributed (IID) zero-mean elements that are also independent of the inputs  $\mathbf{x} \in \mathbb{R}^C$ , we can show that  $\mathbb{E}[\langle \mathbf{w}, \mathbf{x} \rangle^2] = \mathbb{E}[\|\mathbf{w}\|^2]\mathbb{E}[\|\mathbf{x}\|^2]/C$  resulting in  $\alpha_{\mathbf{W}} \approx 1/\sqrt{C}$ . This is analogous to the computation used to derive variance preserving initialization schemes which call for  $\|\mathbf{W}\| \approx \sqrt{K}$ , which notably does not depend on the input dimension  $C$ .

The update alignment is slightly more complicated since the update  $\Delta\mathbf{W}$  is a function of the inputs  $\mathbf{X}$  rather than independent, causing correlations that tend to increase the update alignment. Specifically the weight gradient is an outer product of the inputs  $\mathbf{X}$  and the gradients with respect to  $\mathbf{Y}$ . For SGD at batch size one, this makes the update a scaled version of the input, giving perfect alignment  $\alpha_{\Delta\mathbf{W}} = 1$  regardless of  $C$ . Based on this we get the core  **$\mu\mathbf{P}$  alignment assumptions**:

$$\alpha_{\Delta\mathbf{W}} = \Theta(1), \quad \alpha_{\mathbf{W}} = \Theta(1/\sqrt{C}), \quad \frac{\alpha_{\Delta\mathbf{W}}}{\alpha_{\mathbf{W}}} = \Theta(\sqrt{C}) \quad (5)$$

With this we can scale the learning rate to counteract changes in the alignment ratio in Equation 4. When the network width is multiplied  $C \mapsto mC$ , we thus get the following  **$\mu\mathbf{P}$ -prescribed scaling**:

$$\frac{\|\Delta\mathbf{W}\|}{\|\mathbf{W}\|} \propto \left(\frac{\alpha_{\Delta\mathbf{W}}}{\alpha_{\mathbf{W}}}\right)^{-1} \propto \frac{1}{\sqrt{m}} \text{ (relative update size)} \implies \eta = \eta_{\text{base}}/m \text{ (Adam LR)} \quad (6)$$

The  $\eta$  scaling follows from the relative one based on  $\|\Delta\mathbf{W}\| \propto \eta\sqrt{KC}$  for Adam and  $\|\mathbf{W}\| \approx \sqrt{K}$ . The optimal *base learning rate*  $\eta_{\text{base}}$  is what we aim to transfer across widths, tuning it at a small scale rather than the target scale we are interested in, saving significant compute in the process.

We emphasize that the prescribed relative update scaling in Equation 6 is fully consistent with prior works on  $\mu\mathbf{P}$  and not an artifact of our formulation, see for example Appendix B.3 of Yang et al. (2021) or Adafactor “Full Align” in Table 1 of Everett et al. (2024). Besides scaling the learning rate of hidden layers,  $\mu\mathbf{P}$  involves special handling of the final output layer and attention normalization. Everett et al. (2024) shows these are not necessary for learning rate transfer, combining standard parameterization with layer-wise scaling of the learning rate works equally well or better in practice. We adopt their simpler version in this work although we refer to the scaling as  $\mu\mathbf{P}$ -prescribed or simply  $\mu\mathbf{P}$  for brevity. See Appendix F.2 for details.

### 3 INDEPENDENT WEIGHT DECAY EVENTUALLY NEUTRALIZES $\mu\mathbf{P}$ ’S SCALING

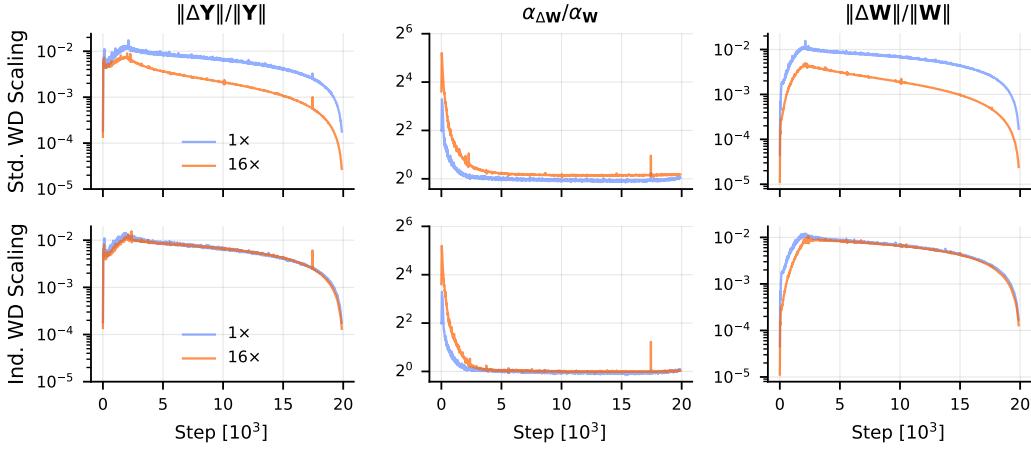
There are two common formulations of AdamW. In the original work by Loshchilov & Hutter (2019), the rate of weight decay was determined by  $\lambda$  alone, i.e., at each step the weights are scaled by  $1 - \lambda$ . However in the default implementation in PyTorch the weights are instead shrunk by  $1 - \eta\lambda$ . These are equivalent since we can scale  $\lambda$  to get one from the other, but will behave differently when  $\eta$  is modified by  $\mu\mathbf{P}$ . In this work we base all discussion on the PyTorch variant but consider the following two approaches for weight decay as  $\mu\mathbf{P}$  scales the learning rate  $\eta \mapsto \eta/m$ :

$$(\eta, \lambda) \mapsto (\eta/m, \lambda) \quad (\text{standard scaling}) \quad (7)$$

$$(\eta, \lambda) \mapsto (\eta/m, m\lambda) \quad (\text{independent scaling}) \quad (8)$$

Note that independent scaling keeps the product  $\eta\lambda$  constant, meaning the relative updates in equilibrium are not scaled at all (see Equation 3). This fundamentally differs from  $\mu\mathbf{P}$ ’s prescribed scaling in Equation 6, meaning that **independent weight decay scaling contradicts  $\mu\mathbf{P}$  by eventually overriding its update scaling**. On the other hand standard weight decay scaling results in equilibrium behavior that follows the relative update scaling prescribed by  $\mu\mathbf{P}$ . This is surprising because multiple works have empirically found that using independent weight decay gives better learning rate transfer (Wortsman et al., 2024; Blake et al., 2025; Wang & Aitchison, 2025; Bergsma et al., 2025). We replicate this finding for LLaMA (Touvron et al., 2023) training on DCLM (Li et al., 2024) next token prediction in Figure 1. Together these results suggest that **applying the  $\mu\mathbf{P}$ -theoretical scaling of the relative update size throughout training does not work well in practice**.

To the best of our knowledge, this key mismatch between theory and what empirically works has not been pointed out before. In the following sections we use our unified relative framework to clarify the exact roles of both  $\mu\mathbf{P}$ ’s scaling and weight decay in achieving learning rate transfer in practice.



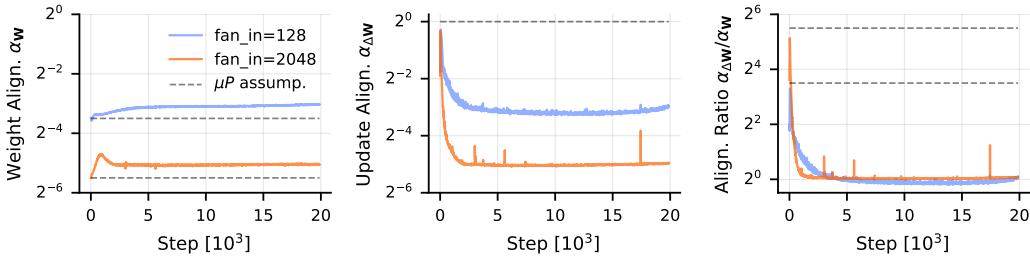
**Figure 3: Standard  $\mu$ P scaling fails to maintain relative representation changes across widths.** Comparing the relative representation change (RRC) for  $\mu$ P with standard weight decay scaling (first row) and independent weight decay scaling (second row) for LLaMA training with a 10% linear warmup followed by linear decay. Plots shown for a single representative layer with  $C = 128$  at  $1\times$  width, see F.3.2 for details. **Left:** Standard scaling does not preserve the RRC for the same learning rate unlike independent scaling. Recall that the RRC is the product of the alignment ratio and the relative weight update (Equation 4). **Middle:** The alignment ratio is similar for both widths and approximately 1. **Right:** To maintain the RRC across widths the relative weight update size should then be kept constant as achieved by independent weight decay rather than  $\mu$ P’s prescribed scaling.

#### 4 WEIGHT DECAY COUNTERACTS $\mu$ P’S BROKEN ALIGNMENT ASSUMPTIONS

Recall that in our formulation  $\mu$ P’s core objective becomes mapping a given base learning rate to the same rate of relative representation changes across different network widths. In the left column of Figure 3 we experimentally measure how well this is achieved under both standard and independent weight decay scaling during LLaMA pre-training. We find that independent weight decay scaling successfully maintains the relative representation changes across widths as indicated by the similarity of the traces on the lower left panel. In contrast, standard weight decay scaling fails to do this, the same base learning rate results in vastly different relative feature updates for the narrow and wide networks later in training. Since learning rates transfer well with independent weight decay (Figure 1), this suggest that the our goal of **maintaining relative representation changes across widths can be an effective objective for learning rate transfer, but  $\mu$ P relies on independent weight decay to achieve this later in training**.

The reason independent weight decay helps maintain relative representation changes can be understood from the middle and right columns of Figure 3. Recall from Equation 4 that the relative representation change is the product of the alignment ratio shown in the middle column and the relative weight change in the right column. We see that the alignment ratio quickly falls from the width-dependent initial value to approximately one, violating  $\mu$ P’s assumptions. Once this happens, **relative weight changes should be the same across both network widths in order to produce the same relative representation changes, which is exactly what independent weight decay achieves** in practice, matching theoretical models. Standard weight decay scaling results in smaller relative updates for the larger network, as prescribed by  $\mu$ P. This fails to maintain the relative representation change with the lower alignment ratios that dominate training, preventing learning rate transfer.

In Appendix B we explore the effects of not using weight decay at all. Somewhat counterintuitively, this also results in similarly sized relative weight updates across widths like with independent weight decay. The reason is that scale-invariant weights grow perpetually when updated without weight decay. Once the cumulated updates outweigh the initialization, the weight norm becomes proportional to the (peak) learning rate used, making both the weights and their updates proportional to it. This means their ratios, **the relative weight updates, eventually lose their dependence on the peak learning rate without weight decay**. We believe this is a key reason learning rates somewhat transfer without weight decay despite  $\mu$ P’s broken assumptions. The ever decreasing relative changes, and the lack of control over them, likely also contributes to the loss of performance without weight decay.



**Figure 4: The alignment assumptions of  $\mu P$  do not hold in practice.** Measurements of the weight alignment, the update alignment and the alignment ratio for LLaMA training at different widths. The update alignment varies significantly over time, becoming width-dependent. As a result the alignment ratio loses its width dependence. This violates  $\mu P$ 's core scaling assumptions which only hold very early in training. Plots shown for one representative MLP layer, see F.3.3 for details.

## 5 WHY DO $\mu P$ 'S ALIGNMENT ASSUMPTIONS BREAK DOWN IN PRACTICE?

Figure 4 shows how  $\mu P$ 's alignment assumptions only hold very briefly at the start of training. Both the update and weight alignment can deviate significantly, for the weight alignment this is more prominent in our ResNet experiments (see Figure 13, Appendix E). In this section we investigate why and when  $\mu P$ 's assumptions about each alignment component break down in practice.

### 5.1 UPDATE ALIGNMENT

In Appendix C we show how **the update alignment can become width-dependent when the batch size  $B$  is large compared to the input dimension  $C$ , violating  $\mu P$ 's assumptions**. This is frequently the case in practice unlike in the infinite width setting  $\mu P$  was originally developed for.

To understand the conceptual reason for the width-dependence, let us look at the simplified case of SGD for a single neuron i.e.,  $K = 1$  in Equation 1. Assume the input samples  $\mathbf{x}_b \in \mathbb{R}^C$  and the output gradients  $\mathbf{y}'_b \in \mathbb{R}$  are all zero-mean IID random and mutually independent. The gradient for each sample is then given by  $\mathbf{g}_b = \mathbf{x}_b \mathbf{y}'_b$  and is also IID and zero-mean. The weight update is given by  $\Delta \mathbf{w} = -\eta \frac{1}{B} \sum_b \mathbf{g}_b$ , allowing us to write the  $i$ -th element of the output change  $\Delta \mathbf{y}$  as:

$$\Delta y_i = \langle \Delta \mathbf{w}, \mathbf{x}_i \rangle = \langle -\eta \frac{1}{B} \sum_b \mathbf{x}_b \mathbf{y}'_b, \mathbf{x}_i \rangle = -\eta \frac{1}{B} y'_i \langle \mathbf{x}_i, \mathbf{x}_i \rangle - \eta \frac{1}{B} \sum_{b \neq i} y'_b \langle \mathbf{x}_b, \mathbf{x}_i \rangle \quad (9)$$

The output change for each input involves one *self-contribution* term and  $B - 1$  *interference* terms from other samples. Although random alignment weakens each interference term by a factor of  $1/\sqrt{C}$  relative to the fully aligned self-contribution, the random sum across all  $B - 1$  terms amplifies their impact by roughly  $\sqrt{B}$ . As a result, interference dominates when  $B \gg C$ , giving an overall dependence on  $C$ . Conversely, for  $B \ll C$ , the interference becomes negligible making the update alignment invariant to  $C$ . In models like Transformers and CNNs, the spatial or sequence dimensions act similarly to the batch size. For our LLaMA experiments, the *effective batch size* becomes the total number of tokens per batch (1,048,576), which far exceeds the network width ( $C \leq 3 \times 2048$ ).

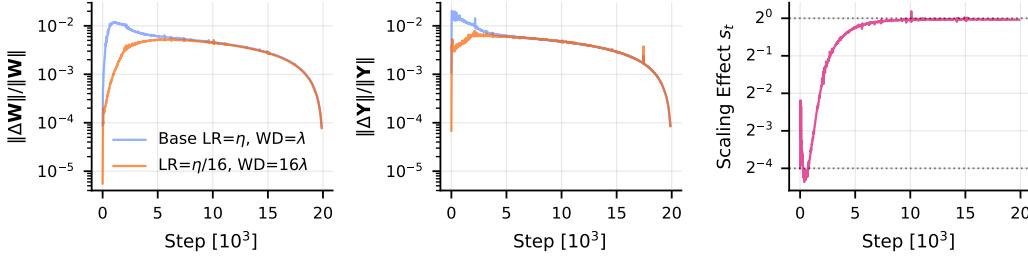
In practice the gradients of different samples are not uncorrelated. High initial correlation results in a large update alignment that resembles the small batch-to-width setting that  $\mu P$  targets. The correlation decreases over time, resulting in width-dependence that violates  $\mu P$ 's assumptions, see Appendix C.

### 5.2 WEIGHT ALIGNMENT

The initial weight alignment matches  $\mu P$ 's assumptions, but analyzing it later in training is hard. Intuitively, we note that we can decompose the weights  $\mathbf{W}_t$  at time  $t$  into weight updates  $\Delta \mathbf{W}_\tau$  from prior time steps  $\tau$ , where we assume  $\Delta \mathbf{W}$  does not include the weight decay term (if used):

$$\mathbf{W}_t = (1 - \eta \lambda)^t \cdot \mathbf{W}_0 + \sum_{\tau=0}^{t-1} (1 - \eta \lambda)^{t-\tau} \cdot \Delta \mathbf{W}_\tau \quad (\text{update composition of weights}) \quad (10)$$

If the decay rate  $\eta \lambda$  is large or  $\|\mathbf{W}_0\|$  is small compared to the updates, the  $\mathbf{W}_0$  term quickly becomes insignificant. Afterwards, the weight alignment at time  $t$  is determined primarily by the alignment of  $\mathbf{X}_t$  and recent update terms  $\Delta \mathbf{W}_\tau$ . **If the current update and recent updates are similarly aligned with the inputs, the overall weight alignment may approximate the update alignment.** This would explain the alignment ratios  $\alpha_{\Delta \mathbf{W}} / \alpha_{\mathbf{W}} \approx 1$  that we often observe in practice (see Figure 16).



**Figure 5:  $\mu$ P with independent WD scaling has a warmup-like effect on relative updates.** Measurements of the relative weight update (left) and relative representation change (middle) for different learning rate (LR) and weight decay (WD) combinations in a LLaMA layer with 2048 input features. Independent weight decay scaling achieves a given LR-WD product using a lower LR and higher WD (orange) compared to not performing  $\mu$ P scaling (blue). This results in smaller relative updates early in training, similar to an additional learning rate warmup factor. The right panel shows this effect through the ratio of the relative weight updates for the two runs, which goes from  $\frac{1}{m} = \frac{1}{16}$  to asymptotically approaching 1 in a roughly exponential manner. See F.3.4 for experimental details.

## 6 THE WARMUP EFFECT OF $\mu$ P WITH INDEPENDENT WEIGHT DECAY

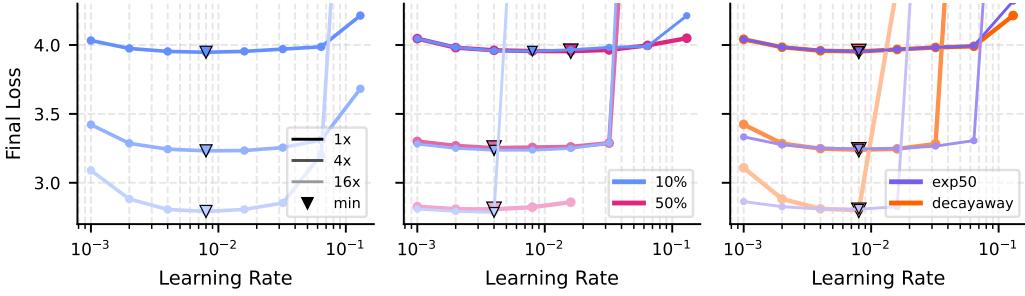
Based on the order-one alignment ratios that we find to dominate practical training, the relative weight updates should be kept roughly constant across widths. Kosson et al. (2024b) described how the magnitude of the relative updates in the steady-state only depends on the product of the learning rate  $\eta$  and weight decay  $\lambda$  rather than their individual values. However,  $(\eta, \lambda)$  pairs with the same product generally affect the initial phases of training differently. Specifically, high weight decay pairs like those obtained from independent  $\mu$ P scaling via  $(\eta, \lambda) \mapsto (\eta/m, m\lambda)$  have a warmup-like effect, where the relative updates at the start of training are comparatively smaller. Therefore,  **$\mu$ P acts as a special form of additional learning rate warmup when used with independent WD scaling**. We measure this effect empirically for LLaMA training in Figure 5, confirming the general effect is still present despite the lack of the perfect scale-invariance typically assumed by works on weight decay. Note that since  $\mu$ P does not scale the updates appropriately for the later phases of training, **this warmup effect is the only practical benefit of  $\mu$ P's learning rate scaling for longer runs**.

The shape of the induced warmup effect is generally quite complex. The initial downscaling of the relative update is always  $1/m$  where  $m$  is width ratio used for  $\mu$ P's scaling. The scaling approaches one over time as the weight norms close in on their equilibrium values. The effective length of the warmup depends on the initialization value, learning rate and weight decay. In Appendix D we analyze a simplified setting with a constant learning rate where the scaling ratio becomes:

$$s_t = \sqrt{\frac{1 + (\rho_0^2/\rho_\infty^2 - 1)a^{2t}}{1 + (m^2\rho_0^2/\rho_\infty^2 - 1)a^{2t}}} \quad \text{if } \rho_0 = \rho_\infty \quad \frac{1}{\sqrt{1 + (m^2 - 1)a^{2t}}} \quad (\text{simple warmup effect}) \quad (11)$$

where  $a := 1 - \eta\lambda$  is the decay multiplier at each step,  $\rho_t$  is the weight RMS,  $\rho_0$  is the initialization RMS value, and  $\rho_\infty = \frac{\eta}{2\lambda}$  is the predicted weight RMS in equilibrium. In practice this model does not accurately predict the shape in Figure 5, overestimating the time required to reach equilibrium. The reason is likely the interaction of momentum with temporal gradient correlations that do not appear in the simplified fully random and uncorrelated setting the model is derived for.

Finally we note that  $\mu$ P's learning rate scaling in AdamW is used to both counteract assumed changes in the alignment ratio and the size of the relative updates. Recall that the learning rate is scaled  $\eta \propto 1/m$  whereas the relative updates should only scale with  $1/\sqrt{m}$ . Even without any assumptions about the alignment ratio, it would make sense to scale the learning rate and weight decay  $(\eta, \lambda) \mapsto (\eta/\sqrt{m}, \sqrt{m} \cdot \lambda)$ . This scales  $\rho_\infty \propto 1/\sqrt{m}$ , matching how variance preserving initialization scales  $\rho_0 \propto 1/\sqrt{m}$ . Such scaling preserves both  $a$  and  $\rho_0/\rho_\infty$  in Equation 11, and is thus more likely to preserve the relative weight update behavior across training. In comparison independent  $\mu$ P scaling has an additional warmup effect of  $1/\sqrt{m}$  in the relative updates (which is desirable), but also decreases  $\rho_\infty \propto 1/m$ . These smaller weight norms could potentially cause issues (signal propagation etc.) if there are not learnable gains or normalization layers to counteract them.



**Figure 6: Stronger warmup can replicate the benefit of  $\mu$ P with independent WD scaling.** Learning rate transfer plots for LLaMA training using different warmup approaches. **Left:**  $\mu$ P with independent scaling transfers well. **Middle:** No  $\mu$ P scaling with either the base 10% linear warmup or a longer 50% warmup. These give a poor learning rate transfer in comparison. **Right:** No  $\mu$ P scaling with additional warmup incorporating the width scaling factor  $m$  can give good transfer, but can require tuning and is less stable at higher learning rates. Additional 50% exponential warmup (Equation 12) compared to a decay-away warmup (Equation 13). See F.3.5 for experimental details.

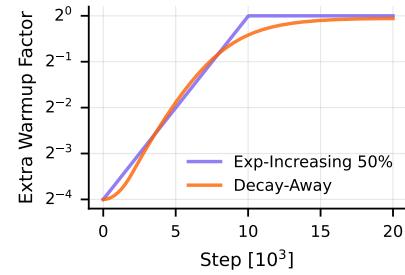
## 7 CAN STRONGER LEARNING RATE WARMUP REPLACE $\mu$ P'S LR SCALING?

Learning rate warmup is already commonly used in practice, making it unclear whether the additional warmup effect from  $\mu$ P is really needed. The first two panels of Figure 6 explore this for LLaMA training, comparing the quality of learning rate transfer with and without  $\mu$ P's scaling. Somewhat surprisingly, we find that **even long linear warmup does not necessarily give the same benefits for learning rate transfer or stability as  $\mu$ P with independent weight decay**. In Appendix E we conduct a similar experiment finding that **the additional warmup effect is not needed for ResNet training**. This could be a part of the reason  $\mu$ P like learning rate scaling did not appear until transformers became popular, well-normalized convolutional networks may not need it at reasonable scales.

We conjecture that linear warmup schedules may increase too fast early on, not allowing enough time for the alignment ratio to fall. Taking inspiration from the more exponential warmup shape from Equation 11, we experiment with two types of additional multiplicative learning rate warmup factors:

$$\hat{s}_t = m^{\min(0, t/T_W - 1)} \quad (\text{exp-increasing warmup}) \quad (12)$$

$$\hat{s}_t = (1 + (m^2 - 1) \prod_{\tau=0}^{t-1} (1 - \eta_\tau \lambda)^2)^{-1/2} \quad (\text{decay-away warmup}) \quad (13)$$



**Figure 7:** Additional warmup factors applied on top of a linear decay schedule with 10% linear warmup for  $m = 16$ ,  $\eta = 0.004$ , and  $\lambda = 0.1$ , matching the minimum in the right panel of Figure 6.

Both scale the initial learning rate by  $1/m$ , similar to  $\mu$ P, and eventually become one. The exponentially increasing one explicitly defines the length of the warmup through a hyperparameter  $T_W$  which can differ from the length of the linear warmup. The decay-away warmup comes directly from the simplified case of Equation 11. In this case the length of the warmup effect varies depending on  $\eta\lambda$ .

The final panel of Figure 6 shows the effect of applying these additional warmup factors on top of the existing 10% linear warmup. This works well, offering better stability and learning rate transfer than with the linear schedules suggesting: **Additional exponential warmup can sometimes replace  $\mu$ P's learning rate scaling in practice**. Although the effect from independent scaling still seems to stabilize higher learning rates slightly better, these results further point to the warmup effect being  $\mu$ P's primary benefit. We note that in our experiments we applied the warmup factors to all parameters. This differs from the warmup effect of the independent scaling which only affects parameters whose learning rate is scaled by  $\mu$ P, i.e., the weight matrices of the hidden and final output layers.

---

## 8 RELATED WORK

We discuss related work in more detail in Appendix A, only briefly mentioning the strongest influences and most closely related works here. Our work adapts the layer-wise LR scaling approach of Everett et al. (2024) and uses a similar alignment definition, although they only consider weight alignment. We also take inspiration from the simpler more accessible approach to  $\mu$ P introduced by Yang et al. (2023) to isolate the key alignment assumptions of  $\mu$ P and work with local changes. Finally, we build closely upon the framework for weight decay introduced by Kosson et al. (2024b) and the additional observations about relative updates and learning rate warmup from Kosson et al. (2024a).

The question of why  $\mu$ P requires independent weight decay was explored before by Wang & Aitchison (2025) who also empirically demonstrated its necessity for learning rate transfer. Their analysis revolves around approximating AdamW as an exponential moving average (EMA) over past updates, which has a form similar to the geometric sum in Equation 10. From this perspective the decay rate defines a time scale for the EMA (analogous to a half-life) which is an alternative characterization of the relative update size in equilibrium. They argue the time scale should intuitively remain constant across networks sizes in order for the final network parameters to form a similar average over different data points, but do not justify this formally. We note that although the timescale interpretation can be a useful mental model and approximation, it does not account for the dependency of later updates upon earlier ones. This dependency is exactly what prevents optimization from being performed in a single step by simply taking an appropriate weighted average over all the data. We take a different approach by relating weight decay directly to the goals of  $\mu$ P through the rate of feature changes.

## 9 DISCUSSION & CONCLUSION

Originally  $\mu$ P was proposed as a theoretical framework that aims to control feature learning in the infinite width limit at initialization. This work expands upon this by studying feature learning throughout training in practical finite-width networks through a more empirical approach. Our relative formulation provides a unified view of  $\mu$ P and weight decay, explaining how both play a role in stabilizing feature learning at different times during training. Crucially, this reveals that  $\mu$ P’s downscaling of the relative weight updates for wider networks is only desirable early in training. As training goes on,  $\mu$ P’s scaling must be counteracted by independent weight decay or similar mechanisms in order to stabilize feature learning and thereby facilitate learning rate transfer.

In this work we have focused exclusively on AdamW for two key reasons. First, AdamW is the still the most common optimizer for LLM training, a key application area for  $\mu$ P and learning rate transfer. Second, the original  $\mu$ P formulation for SGD does not scale the learning rate for hidden layers, using special multipliers instead. This can make standard weight decay behave like independent weight decay under  $\mu$ P’s scaling, making it harder to show the discrepancy. However, we believe the alignment dynamics we have described and  $\mu$ P’s broken alignment assumptions occur in SGD and similar optimizers as well. Matrix-level optimizers on the other hand could significantly change these dynamics. For example, Muon (Jordan et al., 2024) can result in a constant low update alignment that does not vary over time, allowing it to bypass much of the complexities we have discussed in this work. This could potentially explain its reduced need for learning rate warmup and makes it a promising alternative to  $\mu$ P for effectively controlling feature learning in practice.

In Appendix E we show that our findings mostly generalize to ResNet training on ImageNet with AdamW. The main difference there seems to be a potentially confounding regularization benefit from higher learning rates. As the network widths increase, smaller relative representation changes minimize the training loss while larger changes minimize the validation loss. Despite this, independent weight decay still gives much better transfer than standard weight decay.

The complex effects of weight decay are unfortunately often overlooked in optimization research for deep learning. The use of weight decay allows specifying the size of relative updates twice, once at the start of training through the learning rate and again in equilibrium through the product of the learning rate and weight decay. Interestingly,  $\mu$ P with independent weight decay scaling actually makes use of this effect to create the beneficial implicit learning rate warmup. Although we showed a similar effect can be achieved via modified learning rate schedules,  $\mu$ P with independent weight decay scaling is a perfectly valid and practical way of achieving this. We hope our work can inspire confidence in this approach in practice, provide useful conceptual insights into learning rate transfer and weight decay, and finally highlight a promising way forward via matrix-level optimizers.

---

## REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Sanjeev Arora, Zhiyuan Li, and Kaifeng Lyu. Theoretical analysis of auto rate-tuning by batch normalization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=rkxQ-nA9FX>. arXiv:1812.03981.

Shane Bergsma, Nolan Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Power lines: Scaling laws for weight decay and batch size in lilm pre-training. *arXiv preprint arXiv:2505.13738*, 2025.

Jeremy Bernstein, Arash Vahdat, Yisong Yue, and Ming-Yu Liu. On the distance between two neural networks and the stability of learning. *Advances in Neural Information Processing Systems*, 33: 21370–21381, 2020. arXiv:2002.03432.

Charlie Blake, Constantin Eichenberg, Josef Dean, Lukas Balles, Luke Yuri Prince, Björn Deisereth, Andres Felipe Cruz-Salinas, Carlo Luschi, Samuel Weinbach, and Douglas Orr.  $u-\mu$ : The unit-scaled maximal update parametrization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=P7KRIiLM8T>. arXiv:2407.17465.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=ne6zeqLFCZ>. arXiv:2302.06675.

Vitaliy Chiley, Ilya Sharapov, Atli Kosson, Urs Koster, Ryan Reece, Sofia Samaniego de la Fuente, Vishal Subbiah, and Michael James. Online normalization for training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. arXiv:1905.05894.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, et al. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*, 2025.

Francesco D’Angelo, Maksym Andriushchenko, Aditya Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=YrAxxscKM2>. arXiv:2310.04415.

Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, Joel Hestness, et al. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023.

Nolan Dey, Quentin Anthony, and Joel Hestness. The practitioner’s guide to the maximal update parameterization, 2024. URL <https://www.cerebras.ai/blog/the-practitioners-guide-to-the-maximal-update-parameterization>.

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. *arXiv preprint arXiv:2505.01618*, 2025.

Katie E 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. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=0ksNeD1SJT>. arXiv:2407.05872.

Falcon-LLM Team. Falcon-h1: A family of hybrid-head language models redefining efficiency and performance, May 2025. URL <https://falcon-lm.github.io/blog/falcon-h1>.

---

Moritz Haas, Sebastian Bordt, Ulrike von Luxburg, and Leena Chennuru Vankadara. On the surprising effectiveness of large learning rates under standard width scaling. *arXiv preprint arXiv:2505.22491*, 2025.

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 (CVPR)*, June 2016. arXiv:1512.03385.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. arXiv:1806.07572.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, 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 Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015. arXiv:1412.6980.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Analyzing & reducing the need for learning rate warmup in GPT training. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a. URL <https://openreview.net/forum?id=ZgDNrpS46k>. arXiv:2410.23922.

Atli Kosson, Bettina Messmer, and Martin Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. In *Forty-first International Conference on Machine Learning*, 2024b. URL <https://openreview.net/forum?id=MQirrNNU2pC>. arXiv:2305.17212.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Kamal Mohamed Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Joshua P Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander T Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alex Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-LM: In search of the next generation of training sets for language models. In *The Thirty-eighth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL <https://openreview.net/forum?id=CNWdWn47IE>. arXiv:2406.11794.

Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=rJg8TeSFDH>. arXiv:1910.07454.

Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33:14544–14555, 2020. arXiv:2010.02916.

Zhiyuan Li, Srinadh Bhojanapalli, Manzil Zaheer, Sashank Reddi, and Sanjiv Kumar. Robust training of neural networks using scale invariant architectures. In *International Conference on Machine Learning*, pp. 12656–12684. PMLR, 2022. arXiv:2202.00980.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023.

---

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>. arXiv:1711.05101.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.

Christian H.X. Ali Mehmeti-Göpel and Michael Wand. On the weight dynamics of deep normalized networks. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://openreview.net/forum?id=AzUCfhJ9Bs>. arXiv:2306.00700.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. arXiv:1912.01703.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y. arXiv:1409.0575.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4596–4604. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/shazeer18a.html>. arXiv:1804.04235.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>.

Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

Mathurin Videau, Badr Youbi Idrissi, Daniel Haziza, Luca Wehrstedt, Jade Copet, Olivier Teytaud, and David Lopez-Paz. Meta Lingua: A minimal PyTorch LLM training library, 2024. URL <https://github.com/facebookresearch/lingua>.

Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using sgd and weight decay. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6380–6391. Curran Associates, Inc., 2021. URL <https://proceedings.neurips.cc/paper/2021/file/326a8c055c0d04f5b06544665d8bb3ea-Paper.pdf>. arXiv:2006.08419.

Xi Wang and Laurence Aitchison. How to set adamw’s weight decay as you scale model and dataset size. In *Forty-second International Conference on Machine Learning*, 2025. URL <https://openreview.net/forum?id=IszVnczhfz>. arXiv:2405.13698.

Ross Wightman. Pytorch image models. <https://github.com/rwightman/pytorch-image-models>, 2019.

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>. arXiv:2309.14322.

Ge Yang, Edward 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 M. Ranzato, A. Beygelzimer,

---

Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17084–17097. Curran Associates, Inc., 2021. URL [https://proceedings.neurips.cc/paper\\_files/paper/2021/file/8df7c2e3c3c3be098ef7b382bd2c37ba-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/8df7c2e3c3c3be098ef7b382bd2c37ba-Paper.pdf). arXiv:2203.03466.

Greg Yang and Edward J. Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11727–11737. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/yang21c.html>. arXiv:2011.14522.

Greg Yang, James B. Simon, and Jeremy Bernstein. A spectral condition for feature learning. *ArXiv*, abs/2310.17813, 2023. URL <https://api.semanticscholar.org/CorpusID:264555180>. arXiv:2310.17813.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs VI: Feature learning in infinite depth neural networks. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=17pVDnpwwl>. arXiv:2310.02244.

Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=B11z-3Rct7>. arXiv:1810.12281.

---

## A EXTENDED RELATED WORK

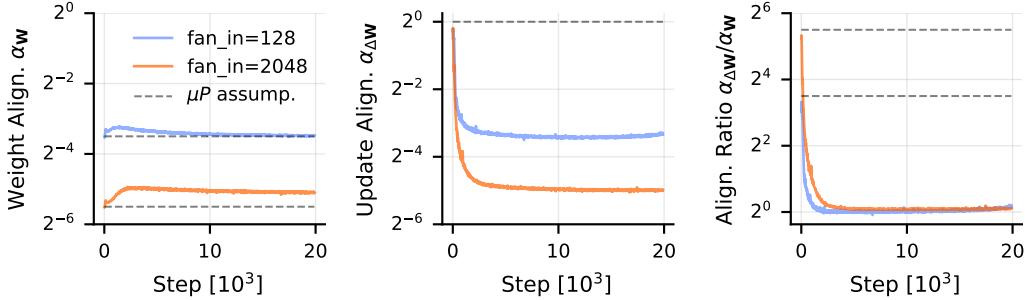
Adam (Kingma & Ba, 2015) is a widely used optimizer in deep learning. It was originally used with  $\ell_2$ -regularization but Loshchilov & Hutter (2019) empirically found that decoupled weight decay works better, resulting in AdamW. The version of AdamW we use is shown in Algorithm 1 and corresponds to the most popular variant, the default in PyTorch (Paszke et al., 2019).

Weight decay is unfortunately still widely believed to act primarily as an explicit regularizer despite a long line of work that argues this is not the case in deep learning (Van Laarhoven, 2017; Zhang et al., 2019; Chiley et al., 2019; Li & Arora, 2020; Li et al., 2020; Wan et al., 2021; Li et al., 2022; Kosson et al., 2024b; D’Angelo et al., 2024). Instead weight decay modulates some notion of an “effective” learning rate, which can be seen as measures of the relative update size. Our work builds closely upon Kosson et al. (2024b), who described the equilibrium dynamics of weight decay in detail for AdamW. This included originally noting the scheduling effects, how weight decay can be replaced by controlling the relative weight update, and how the balanced equilibrium behavior of AdamW explains its effectiveness over Adam with  $\ell_2$ -regularization.

Yang & Hu (2021) proposed the Maximal Update Parameterization  $\mu P$  as a way to ensure stable feature learning in infinitely wide networks. Yang et al. (2021) demonstrated how  $\mu P$  allows transferring the learning rates across different network widths and how this is beneficial for hyperparameter tuning. Later works have expanded  $\mu P$  to network depth rather than only width (Yang et al., 2024; Dey et al., 2025). The original formulation of  $\mu P$  is quite theoretical but Yang et al. (2023) show a simplified and more accessible derivation based on spectral norms which we base our discussion on. Dey et al. (2024) provide a useful overview of  $\mu P$ .

Everett et al. (2024) show hyperparameter transfer is possible for other parameterizations than  $\mu P$ , including standard initialization schemes without the special multipliers used in  $\mu P$ , using analogous learning rate scaling. We adopt this approach in our experiments as it is simpler and results in better performance according to Everett et al. (2024). They also describe how the learning rate scaling depends on alignment metrics similar to those we define in Equation 4, but use the total weight change from the start of training rather than a single optimization step. Our findings fit particularly well with one case they consider, the no-alignment setting. Their analysis of AdaFactor (Shazeer & Stern, 2018) with weight-proportional updates in this case predicts that learning rates should transfer without any scaling. This is similar to what we observe in practice once weight decay makes the updates proportional to the weights. We expand upon their work with measurements of the update-alignment, describing the role of weight decay, and how the batch size affects alignment.

Haas et al. (2025) is very recent work that investigates why higher learning rates than  $\mu P$  prescribes can be used. They investigate alignment changes as a potential cause but dismiss them in favor of the role of the loss function, showing that typical cross entropy losses help stabilize the final layer compared to  $\ell_2$  based losses. Our alignment measure is based on the current update rather than the total update across training, which might explain the different conclusions.



**Figure 8: The alignment assumptions of  $\mu$ P break down even without weight decay.** Measurements of the weight alignment, the update alignment and the alignment ratio for LLaMA training at different widths. The update alignment varies significantly over time, becoming width-dependent. This violates  $\mu$ P’s assumption which only holds very early in training. A base learning rate of  $\eta = 1.6 \cdot 10^{-2}$  is used for the narrower network and scaled down  $16 \times$  for the wider network via  $\mu$ P’s learning rate scaling. Plots shown for one representative MLP layer, `layers.13.feed_forward.w1`.

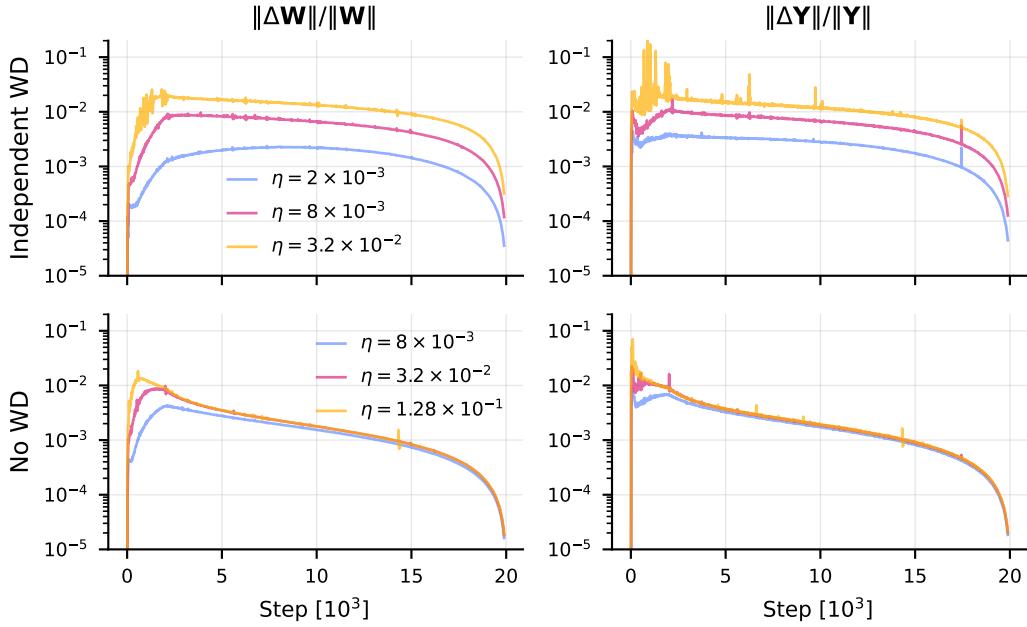
## B $\mu$ P’S ALIGNMENT ASSUMPTIONS STILL FAIL WITHOUT WEIGHT DECAY

In Figure 1 we observed that no weight decay seems to provide better learning rate transfer than the use of standard weight decay, although it is still not as good as with independent weight decay. The final loss is also slightly worse than the best configuration with either weight decay approach. In this section we explore whether the assumptions of  $\mu$ P hold without weight decay.

In Figure 8 we see that this is not the case. The alignment ratio still becomes roughly one, independent of the width, meaning that the alignment assumptions of  $\mu$ P break down exactly like in the weight decay case. **Weight decay is therefore not the cause of  $\mu$ P’s violated assumptions.** This matches our analysis of the update alignment (Section 5) where the width dependence at large batch sizes does not depend on weight decay directly. With an alignment ratio that does not vary across width, the relative updates should be preserved in order to keep the relative representation changes comparable.

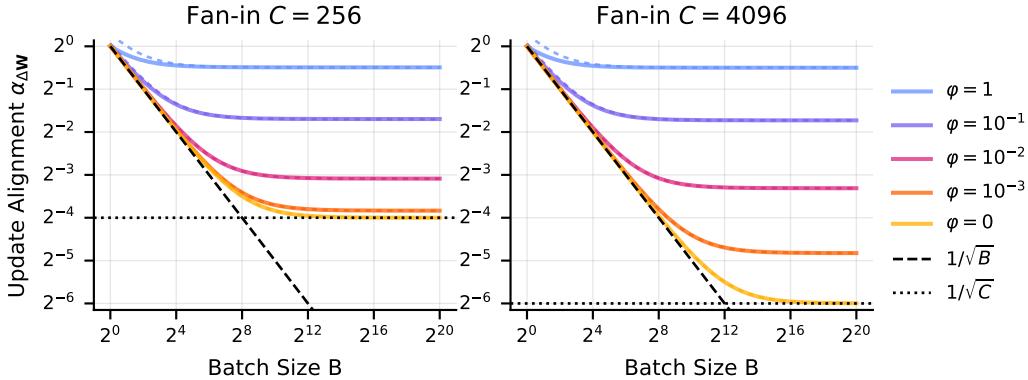
In Figure 9 we measure how the relative updates behave across different learning rates. Notably, **without weight decay the peak learning rate has little effect on the size of relative updates later in training.** This could explain why the no weight decay setting seems to be less sensitive to the specific value of the learning rate in Figure 1. Without weight decay the size of the relative updates falls over time compared to when weight decay is used, similar to an extra decay factor in the learning rate schedule. The way that this decay occurs seems to cause the relative weight updates to lose their dependence on the peak learning rate. In Appendix D.2 we show this can happen in a simplified setting for constant learning rates. The conceptual reason is that if the total weight growth dominates the initial value, both the update and the total weight norm have the same proportional dependency on the learning rate later in training. These factors cancel out resulting in roughly identical relative weight updates across peak learning rates. This assumes the network behaves similar to a scale-invariant network, i.e., that there is no strong gradient signal that affects the weight norms. A similar effect of the peak learning rate losing its relevance over time was observed for SGD and normalization layers before (Arora et al., 2019; Mehmeti-Göpel & Wand, 2024).

The optimal learning rate likely becomes a balance between two factors, meaningful learning and stability. Higher learning rates decrease the contribution of the initial weights to the final model, which may allow the model to better learn the distribution. However, high learning rates can also cause large relative updates early in training that can have detrimental effects as explored by Kosson et al. (2024a), perhaps by saturating non-linearities or causing instability. If these detrimental effects occur sufficiently early, the alignment assumptions of  $\mu$ P may still roughly hold. In this case the optimal learning rate may be decided primarily by early behavior. This may aid optimal learning rate transfer under  $\mu$ P’s scaling, despite its core assumptions only holding briefly at the start of training. We suspect the length of the learning rate warmup may also impact whether the key limiting stability period occurs when  $\mu$ P’s alignment assumptions still hold or not.



**Figure 9: Without weight decay relative updates become similar across learning rates over time.** Measurements of the relative weight update size and relative representation change for LLaMA training. Plots shown for one representative MLP layer, `layers.13.feed_forward.w1`.

Overall the reason for learning rate transfer without weight decay, to the extent it occurs, would then be similar as in the independent weight decay case. Early in training  $\mu\mathbf{P}$  correctly scales down the relative updates which helps transfer relative representation changes under the high, width-dependent, initial alignment ratios. Later in training the relative updates become approximately the same across widths which helps stabilize the rate of feature learning when the alignment ratios are identical across widths. Only standard weight decay follows  $\mu\mathbf{P}$ 's prescribed relative weight update scaling from Equation 6 throughout training, which is exactly why learning rates fail to transfer with it. With either no weight decay or independent weight decay, the relative updates deviate from the prescribed scaling which helps counteract the violated alignment assumptions. However, independent weight decay seems to control the relative updates later in training more tightly, resulting in better transfer overall.



**Figure 10: Predicted update alignment  $\alpha_{\Delta W}$  in the simple noise model based on Equation 25.** The overall behavior is well approximated as  $\alpha_{\Delta W} \approx \sqrt{C^{-1} + B^{-1} + \varphi/(1 + \varphi)}$  shown by the dashed color lines. Note how the alignment for a low signal-to-noise ratio  $\varphi \approx 0$  is roughly  $1/\sqrt{B}$  (no width-dependence as  $\mu P$  assumes) until  $B > C$  when it becomes  $1/\sqrt{C}$  ( $\mu P$ 's assumption fails). Higher values of  $\varphi$  always increase the alignment regardless of  $B$  and  $C$ . This is most prominent when  $\varphi \gg 1/C$  in which case we get  $\alpha_{\Delta W} \approx 1/\sqrt{\varphi^{-1} + 1}$  once  $B \gg \varphi^{-1}$ . This can be viewed as the batch size being effectively capped at  $\varphi^{-1} + 1$ , an estimate of the critical batch size (McCandlish et al., 2018). Early in training this can give a high approximately width-independent alignment, allowing  $\mu P$ 's assumptions to hold initially even when  $B \gg C$  but then fail as the SNR decreases.

### C ANALYSIS OF THE UPDATE ALIGNMENT FOR A SIMPLE NOISE MODEL

In this section we analyze the update alignment for a simple probabilistic model of a single neuron with a fan-in of  $C$  at batch size  $B$ . On the forward pass we have:

$$\mathbf{y}^\top = [y_1, \dots, y_B] = \mathbf{w}^\top \mathbf{X} = \mathbf{w}^\top [\mathbf{x}_1, \dots, \mathbf{x}_B] \quad (14)$$

where  $\mathbf{w} \in \mathbb{R}^C$  are the weights of the neuron,  $\mathbf{X} \in \mathbb{R}^{C \times B}$  is a batch of  $B$  input vectors with dimension  $C$  each, and  $\mathbf{y} \in \mathbb{R}^B$  is the output for each input. The gradient of the final average sample loss  $\mathcal{L} = \frac{1}{B} \sum_{b=1}^B \mathcal{L}_b$  with respect to the weights  $\mathbf{w}$  is given by:

$$\mathbf{g} := \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{1}{B} \sum_{b=1}^B \frac{\partial \mathcal{L}_b}{\partial y_b} \mathbf{x}_b = \frac{1}{B} \sum_{b=1}^B y'_b \mathbf{x}_b \quad (15)$$

where we have defined  $y'_b := \partial \mathcal{L}_b / \partial y_b$ . Note how the gradient for each sample  $\mathbf{g}_b := y'_b \mathbf{x}_b$  is simply a scaled version of the input vector  $\mathbf{x}_b$  and is thus fully aligned with it.

We want to be able to introduce analytically tractable correlations between the random variables in our probabilistic model. Taking inspiration from Kosson et al. (2024a), we use the Signal-to-Noise ratio  $\varphi$  to capture the correlation. It is defined as the square ratio of the norms of the expected gradient (signal) to the expected deviation (noise):

$$\varphi := \frac{\|\bar{\mathbf{g}}\|^2}{\mathbb{E}_b[\|\mathbf{g}_b - \bar{\mathbf{g}}\|^2]}, \quad \bar{\mathbf{g}} := \mathbb{E}_b[\mathbf{g}_b] \quad (16)$$

We can now define a distributions for  $\mathbf{x}_b$  and  $y'_b$  that result in a specific  $\varphi$  while respecting the mathematical form of the gradient and being simple enough to analyze. We chose the form of the input samples as:

$$\mathbf{x}_b = (\tilde{\mathbf{x}}_b + \sqrt{\varphi} \cdot \mathbf{s} \cdot \frac{1}{y'_b}) / \sqrt{1 + \varphi} \quad (17)$$

where the noise component  $\tilde{\mathbf{x}}_b$  and signal component  $\mathbf{s}$  are independent random vectors drawn from a standard multivariate normal distribution, i.e.,  $\tilde{\mathbf{x}}_b, \mathbf{s} \sim \mathcal{N}(\mathbf{0}, I_C)$ . We choose the distribution of the output gradient to be  $y'_b = \pm 1$  with equal probability to avoid complexities in modeling the inverse.

We could scale the variance of the distributions but this would not affect the alignment ratio. In this model the gradient becomes:

$$\mathbf{g}_b = \mathbf{y}'_b \mathbf{x}_b = (\tilde{\mathbf{x}}_b \mathbf{y}'_b + \sqrt{\varphi} \cdot \mathbf{s}) / \sqrt{1 + \varphi} \quad (18)$$

Assuming  $\|\mathbf{s}\|^2 = C$  (or alternatively in expectation over  $\mathbf{s}$ ), this results in the targeted signal-to-noise ratio  $\varphi$ :

$$\frac{\|\bar{\mathbf{g}}\|^2}{\mathbb{E}_b[\|\mathbf{g}_b - \bar{\mathbf{g}}\|^2]} = \frac{\varphi \|\mathbf{s}\|^2 / (1 + \varphi)}{\mathbb{E}_b[\|\mathbf{y}'_b \tilde{\mathbf{x}}_b / \sqrt{1 + \varphi}\|^2]} = \varphi \frac{C / (1 + \varphi)}{C / (1 + \varphi)} = \varphi \quad (19)$$

Note that there are other forms for  $\mathbf{x}_b$  that result in the desired signal-to-noise ratio including some that give  $\bar{\mathbf{g}} = \mathbf{s}$ , but our choice results in bounded input and gradient norms in expectation for both  $\varphi \rightarrow 0$  and  $\varphi \rightarrow \infty$ . We will also assume that the update is obtained via simple gradient descent:

$$\Delta \mathbf{w} = -\eta \mathbf{g} \quad (20)$$

The update alignment involves a ratio that is hard to analyze under expectation. Instead we will approximate this as the ratio of the square expectations:

$$\mathbb{E}[\alpha_{\Delta \mathbf{w}}] = \mathbb{E} \left[ \frac{\|\Delta \mathbf{w}^\top \mathbf{X}\|}{\|\Delta \mathbf{w}\| \|\mathbf{X}\|} \right] \approx \sqrt{\frac{\mathbb{E}[\|\Delta \mathbf{w}^\top \mathbf{X}\|^2]}{\mathbb{E}[\|\Delta \mathbf{w}\|^2] \mathbb{E}[\|\mathbf{X}\|^2]}} \quad (21)$$

we thus need to compute  $\mathbb{E}[\|\Delta \mathbf{w}^\top \mathbf{X}\|^2]$ ,  $\mathbb{E}[\|\Delta \mathbf{w}\|^2]$ , and  $\mathbb{E}[\|\mathbf{X}\|^2]$ .

Starting with the input size we get:

$$\mathbb{E}[\|\mathbf{X}\|^2] = B \cdot \mathbb{E}_b[\|\mathbf{x}_b\|^2] \quad (22a)$$

$$= B \cdot \mathbb{E}_b \left[ \frac{1}{1 + \varphi} \left\langle \tilde{\mathbf{x}}_b + \sqrt{\varphi} \cdot \mathbf{s} \cdot \frac{1}{\mathbf{y}'_b}, \tilde{\mathbf{x}}_b + \sqrt{\varphi} \cdot \mathbf{s} \cdot \frac{1}{\mathbf{y}'_b} \right\rangle \right] \quad (22b)$$

$$= \frac{B}{1 + \varphi} \left( \mathbb{E}_b[\|\tilde{\mathbf{x}}_b\|^2] + 2\sqrt{\varphi} \mathbb{E}_b \left[ \frac{1}{\mathbf{y}'_b} \langle \tilde{\mathbf{x}}_b, \mathbf{s} \rangle \right] + \varphi \mathbb{E}_b \left[ \frac{1}{(\mathbf{y}'_b)^2} \|\mathbf{s}\|^2 \right] \right) \quad (22c)$$

$$= \frac{B}{1 + \varphi} (C + 0 + \varphi \cdot C) \quad (22d)$$

$$= BC \quad (22e)$$

For the update size we get:

$$\mathbb{E}[\|\Delta \mathbf{w}\|^2] = \eta^2 \mathbb{E}[\|\mathbf{g}\|^2] \quad (23a)$$

$$= \eta^2 \mathbb{E} \left[ \left\| \frac{1}{B} \sum_{b=1}^B \frac{\mathbf{y}'_b \tilde{\mathbf{x}}_b + \sqrt{\varphi} \mathbf{s}}{\sqrt{1 + \varphi}} \right\|^2 \right] \quad (23b)$$

$$= \frac{\eta^2}{B^2(1 + \varphi)} \mathbb{E} \left[ \left\| \left( \sum_{b=1}^B \mathbf{y}'_b \tilde{\mathbf{x}}_b \right) + B \sqrt{\varphi} \mathbf{s} \right\|^2 \right] \quad (23c)$$

$$= \frac{\eta^2}{B^2(1 + \varphi)} \left( \mathbb{E} \left[ \left\| \sum_{b=1}^B \mathbf{y}'_b \tilde{\mathbf{x}}_b \right\|^2 \right] + 2\mathbb{E} \left[ \left\langle \sum_{b=1}^B \mathbf{y}'_b \tilde{\mathbf{x}}_b, B \sqrt{\varphi} \mathbf{s} \right\rangle \right] + \mathbb{E} [\|B \sqrt{\varphi} \mathbf{s}\|^2] \right) \quad (23d)$$

$$= \frac{\eta^2}{B^2(1 + \varphi)} \left( \sum_{b=1}^B \mathbb{E}[\|\tilde{\mathbf{x}}_b\|^2] + 0 + B^2 \varphi \mathbb{E}[\|\mathbf{s}\|^2] \right) \quad (23e)$$

$$= \frac{\eta^2}{B^2(1 + \varphi)} (BC + B^2 \varphi C) \quad (23f)$$

$$= \frac{\eta^2 C}{B} \frac{1 + B \varphi}{1 + \varphi} \quad (23g)$$

Finally we can compute the expected output change as:

$$\mathbb{E}[\|\Delta \mathbf{w} \mathbf{X}\|^2] = B\mathbb{E}[\langle \mathbf{x}_b, -\eta \mathbf{g} \rangle^2] \quad (24a)$$

$$= \frac{B\eta^2}{(1+\varphi)^2} \mathbb{E}\left[\left\langle \tilde{\mathbf{x}}_b + \sqrt{\varphi} \mathbf{s} / \mathbf{y}'_b, \frac{1}{B} \sum_{i=1}^B (\tilde{\mathbf{x}}_i \mathbf{y}'_i + \sqrt{\varphi} \mathbf{s}) \right\rangle^2\right] \quad (24b)$$

$$= \frac{\eta^2}{(1+\varphi)^2 B} \mathbb{E}\left[\left(\mathbf{y}'_b \|\tilde{\mathbf{x}}_b\|^2 + \sum_{i \neq b} \mathbf{y}'_i \langle \tilde{\mathbf{x}}_b, \tilde{\mathbf{x}}_i \rangle + (1+B)\sqrt{\varphi} \langle \tilde{\mathbf{x}}_b, \mathbf{s} \rangle + \frac{\sqrt{\varphi}}{\mathbf{y}'_b} \sum_{i \neq b} \langle \mathbf{s}, \tilde{\mathbf{x}}_i \rangle \mathbf{y}'_i + \frac{B\varphi}{\mathbf{y}'_b} \|\mathbf{s}\|^2\right)^2\right] \quad (24c)$$

$$= \frac{\eta^2}{(1+\varphi)^2 B} \left[ \mathbb{E}[\|\tilde{\mathbf{x}}_b\|^4] + \mathbb{E}\left[\left(\sum_{i \neq b} \mathbf{y}'_i \langle \tilde{\mathbf{x}}_b, \tilde{\mathbf{x}}_i \rangle\right)^2\right] + (1+B)^2 \varphi \mathbb{E}[\langle \tilde{\mathbf{x}}_b, \mathbf{s} \rangle^2] + \varphi \mathbb{E}\left[\left(\sum_{i \neq b} \langle \mathbf{s}, \tilde{\mathbf{x}}_i \rangle \mathbf{y}'_i\right)^2\right] + B^2 \varphi^2 \mathbb{E}[\|\mathbf{s}\|^4] + 2B\varphi \mathbb{E}[\|\tilde{\mathbf{x}}_b\|^2 \|\mathbf{s}\|^2] \right] \quad (24d)$$

$$= \frac{\eta^2}{(1+\varphi)^2 B} \left[ \mathbb{E}[\|\tilde{\mathbf{x}}_b\|^4] + \sum_{i \neq b} \mathbb{E}[\langle \tilde{\mathbf{x}}_b, \tilde{\mathbf{x}}_i \rangle^2] + (1+B)^2 \varphi \mathbb{E}[\langle \tilde{\mathbf{x}}_b, \mathbf{s} \rangle^2] + \varphi \sum_{i \neq b} \mathbb{E}[\langle \mathbf{s}, \tilde{\mathbf{x}}_i \rangle^2] + B^2 \varphi^2 \mathbb{E}[\|\mathbf{s}\|^4] + 2B\varphi \mathbb{E}[\|\tilde{\mathbf{x}}_b\|^2 \|\mathbf{s}\|^2] \right] \quad (24e)$$

$$= \frac{\eta^2}{(1+\varphi)^2 B} \left[ C(C+2) + (B-1)C + (1+B)^2 \varphi C + \varphi(B-1)C + B^2 \varphi^2 C(C+2) + 2B\varphi C^2 \right] \quad (24f)$$

Now we have all the factors needed to approximate the update alignment via Equation 21 giving:

$$\mathbb{E}[\alpha_{\Delta \mathbf{w}}] \approx \sqrt{\frac{(C+2) + (B-1) + (1+B)^2 \varphi + \varphi(B-1) + B^2 \varphi^2(C+2) + 2B\varphi C}{(1+B\varphi)(1+\varphi)CB}} \quad (25)$$

Figure 10 plots this prediction showing  $C$ -dependent behavior for large batch-to-width ratios as expected. We find the overall behavior in Equation 25 well approximated by:

$$\mathbb{E}[\alpha_{\Delta \mathbf{w}}] \approx \sqrt{C^{-1} + B^{-1} + \varphi/(1+\varphi)} \quad (26)$$

The last term is the inverse of  $\varphi^{-1} + 1$  which is an estimate of the critical batch size (McCandlish et al., 2018) where the overall gradient starts to become dominated by the signal component and the benefit of increasing the batch size further diminishes quickly. In the fully uncorrelated case  $\varphi = 0$  we roughly get  $\alpha_{\Delta \mathbf{w}} \approx 1/\sqrt{B}$  when  $B \ll C$  which is the regime targeted by  $\mu P$ , infinitely wide networks trained with finite batch sizes. In the practical training settings we often have  $B \gg C$  which results in  $\alpha_{\Delta \mathbf{w}} \approx 1/\sqrt{C}$ . This  $C$  dependence breaks the scaling behavior of  $\mu P$ . When the critical batch size is small compared to  $C$  this can instead become  $\alpha_{\Delta \mathbf{w}} \approx \sqrt{\frac{\varphi}{1+\varphi}}$  which eliminates the  $C$  dependence again. The overall alignment behavior is thus highly dependent on the specific settings of  $B$  and  $C$  as well as the signal-to-noise ratio  $\varphi$  which dynamically changes throughout training.

---

## D ANALYZING THE SCHEDULING EFFECTS OF WEIGHT DECAY

### D.1 THE EVOLUTION OF THE WEIGHT NORM OVER TIME

Here we construct a simple model of how the weight norms evolve over time, similar to the one used in prior work (Kosson et al., 2024b). In this section we use  $\Delta \mathbf{W}$  to refer to this “gradient component” of the update, treating the weight decay component separately. We will assume that the weight update (excluding weight decay) at each timestep is orthogonal to the current weights. This results in a simple recurrence relation for the weight norms:

$$\|\mathbf{W}_{t+1}\|^2 = (1 - \eta_t \lambda)^2 \|\mathbf{W}_t\|^2 + \|\Delta \mathbf{W}_t\|^2 \quad (27)$$

We will assume that the elementwise RMS size of the weight update is roughly one, for example due to the second moment normalization in Adam. Defining  $\rho_t := \|\mathbf{W}_t\|/\sqrt{KC}$  as the RMS value of the elements in  $\|\mathbf{W}_t\|$  and  $a_t := 1 - \eta_t \lambda$  as the decay multiplier we get a recurrence relation:

$$\rho_{t+1}^2 = a_t^2 \rho_t^2 + \eta_t^2 \quad (28)$$

This model involves several significant simplifications:

- The assumption that the update is orthogonal to the current weights can either come from randomness or due to scale-invariance from normalization layers. The gradient of a scale-invariant weight is always orthogonal to the weight and zero-mean uncorrelated random gradient is orthogonal to the weights on average. We note that with momentum the update is not necessarily orthogonal on average, even in these special cases.
- This model only holds for momentum when the gradients are uncorrelated over time. In this case we can replace the update term by the total contribution of the current gradient over time (the sum of the impulse response) as done by Kosson et al. (2024b). This also assumes  $\beta_1$  is not too large, so that the impact of a single gradient happens over a short timeframe compared to the  $t$  values we are interested in. If there is correlation the behavior of the system changes in a way we can not predict without additional information. See Wan et al. (2021) for an analysis that partially accounts for this in SGDM.
- The size of the update assumes that  $\beta_2$  can accurately track the expected magnitude of the gradient over time. This either requires small  $\beta_2$  values in comparison to how fast the gradient magnitude changes. We note that at the start of training the gradient magnitude often changes very rapidly.

This means that in practice Equation 28 does not accurately predict the behavior of the weight norm in neural network training. However, we believe it is the best we can do based on the hyperparameters alone without measurements of the temporal gradient correlation and gradient magnitude over time (which we expect strongly depend on the other hyperparameters too). We still find the predictions of this model to be informative for real training, particularly those based on the equilibrium value that satisfies  $\rho_{t+1} = \rho_t$ . The RMS values seen in real training often stabilize near this value (Kosson et al., 2024b), but how fast this happens depends strongly on the complex temporal effects described above.

With this disclaimer, we proceed with the analysis of Equation 28. Rolling out the recurrence relation:

$$\rho_{t+1}^2 = \rho_0^2 \prod_{\tau=0}^t a_\tau^2 + \sum_{\tau=0}^t \eta_\tau^2 \prod_{i=\tau+1}^t a_i^2 \quad (29)$$

In the case where the learning rate is constant  $\eta_t = \eta$ , making  $a_t = a$  a constant as well, we get:

$$\rho_t^2 = \rho_0^2 \cdot a^{2t} + \eta^2 \sum_{\tau=0}^{t-1} a^{2(t-\tau)} = \rho_0^2 \cdot a^{2t} + \eta^2 \frac{1 - a^{2t}}{1 - a^2} \quad (30)$$

assuming  $0 < a < 1$ . If  $\eta_t$  is not constant, we can still easily compute the behavior of the weight norm over time using the recurrence relation or the rolled-out form directly.

In the constant learning rate case the RMS value approaches the equilibrium value  $\rho_\infty$  exponentially:

$$\rho_t^2 = \rho_\infty^2 + (\rho_0^2 - \rho_\infty^2) a^{2t} \quad \rho_\infty^2 = \frac{\eta^2}{1 - a^2} \approx \frac{\eta}{2\lambda} \quad (31)$$

---

We can use this to estimate how long it takes for the weights to approximately reach equilibrium. If we define this by the timestep  $T_{EQ}$  when the RMS value falls within a factor  $\sqrt{2}$  of the equilibrium value we get:

$$T_{EQ} \approx \begin{cases} \frac{1}{2\eta\lambda} \ln \left( \frac{1-\rho_0^2/\rho_\infty^2}{1-1/\sqrt{2}} \right) & \text{if } \rho_0 < \frac{1}{\sqrt{2}}\rho_\infty \\ \frac{1}{2\eta\lambda} \ln \left( \frac{\rho_0^2/\rho_\infty^2-1}{\sqrt{2}-1} \right) & \text{if } \rho_0 > \sqrt{2}\rho_\infty \\ 0 & \text{otherwise} \end{cases} \quad (32)$$

## D.2 THE EVOLUTION OF THE RELATIVE WEIGHT UPDATE OVER TIME

In this simplified model the relative weight update becomes:

$$\frac{\|\Delta \mathbf{W}_t\|}{\|\mathbf{W}\|} = \frac{b\eta_t}{\rho_t} \quad (33)$$

which we can compute exactly given the expressions for  $\rho_t$  from the previous subsection.

In the constant learning rate case we get:

$$\frac{\|\Delta \mathbf{W}_t\|}{\|\mathbf{W}\|} = \sqrt{\frac{b^2\eta^2}{\rho_t^2}} = \sqrt{\frac{b^2\eta^2}{\rho_0^2 \cdot a^{2t} + b^2\eta^2 \frac{1-a^{2t}}{1-a^2}}} \xrightarrow{t \rightarrow \infty} \sqrt{2\eta\lambda - \eta^2\lambda^2} \approx \sqrt{2\eta\lambda} \quad (34)$$

showing that the product  $\eta\lambda$  determines the size of relative weight updates in the steady-state. However, at the start of training, e.g. the first step, the relative weight update (excluding the weight decay component) is determined by just the learning rate and the initialization norm.

It is also interesting to see what happens without weight decay,  $\lambda = 0$ . For a constant learning rate, this gives:

$$\rho_t^2 = \rho_0^2 + t \cdot \eta^2 \quad \frac{\eta_t}{\rho_t} = \frac{\eta}{\sqrt{\rho_0^2 + t \cdot \eta^2}} = \frac{1}{\sqrt{(\rho_0^2/\eta^2) + t}} \quad (35)$$

which means that the relative updates decay towards zero over time roughly as  $t^{-0.5}$ , rather than stabilizing at a fixed value. For this reason it is important to use learning rate schedules with weight decay, otherwise the effective learning rate measured in terms of relative updates never decays which typically leads to suboptimal outcomes. This decay in the relative update size can also be detrimental, if the relative update size decreases too quickly the network stops learning (sometimes referred to as a loss of plasticity). An example of this was seen in the AdamW paper (Loshchilov & Hutter, 2019) where the optimal weight decay value in an experiment was zero for a constant learning rate but non-zero when a cosine schedule was used. It is better to vary the effective learning rate over time explicitly according to our desired schedule, ideally exactly through optimizers like LionAR (Kosson et al., 2024a) or approximately via weight decay.

## D.3 THE WARMUP EFFECT OF HIGH WEIGHT DECAY CONFIGURATIONS

Let us now consider two configurations for the learning rate and weight decay,  $(\eta, \lambda)$  vs  $(\eta/m, m\lambda)$ . These correspond to the configurations from independent weight decay scaling compared to no scaling (or in standard scaling at a higher base learning rate). We first note that their product is identical, so the relative weight update will converge to the same equilibrium value. However, the very first update will also be  $m$  times smaller for the high weight decay configuration. A learning rate warmup starting at  $\eta/m$  and reaching  $\eta$  over time has an identical effect on the initial update and the steady-state updates. However, the exact shape of this warmup effect is not trivial, i.e. how it affects the relative update size over time.

We use the previous symbols to denote quantities for the base hyperparameters  $(\eta, \lambda)$  and  $\hat{\cdot}$  over the symbol to denote the corresponding quantities for the scaled configuration. We are interested in the ratio of the relative weight updates between the two configurations:

$$s_t := \frac{\hat{b}\hat{\eta}_t/\hat{\rho}_t}{b\eta_t/\rho_t} \quad (36)$$

---

which captures the size of the warmup effect. For the simple weight norm model we can always compute this using Equation 28, even when the learning rate varies over time. For constant learning rates we would get:

$$s_t = \frac{1}{m} \frac{\rho_t}{\hat{\rho}_t} \quad (37a)$$

$$= \frac{1}{m} \sqrt{\frac{\rho_0^2 \cdot a^{2t} + b^2 \eta^2 \frac{1-a^{2t}}{1-a^2}}{\rho_0^2 \cdot a^{2t} + b^2 \hat{\eta}^2 \frac{1-a^{2t}}{1-a^2}}} \quad (37b)$$

$$= \sqrt{\frac{\rho_\infty^2 + (\rho_0^2 - \rho_\infty^2)a^{2t}}{\rho_\infty^2 + (m^2 \rho_0^2 - \rho_\infty^2)a^{2t}}} \quad (37c)$$

$$= \sqrt{\frac{1 + (\rho_0^2/\rho_\infty^2 - 1)a^{2t}}{1 + (m^2 \rho_0^2/\rho_\infty^2 - 1)a^{2t}}} \quad (37d)$$

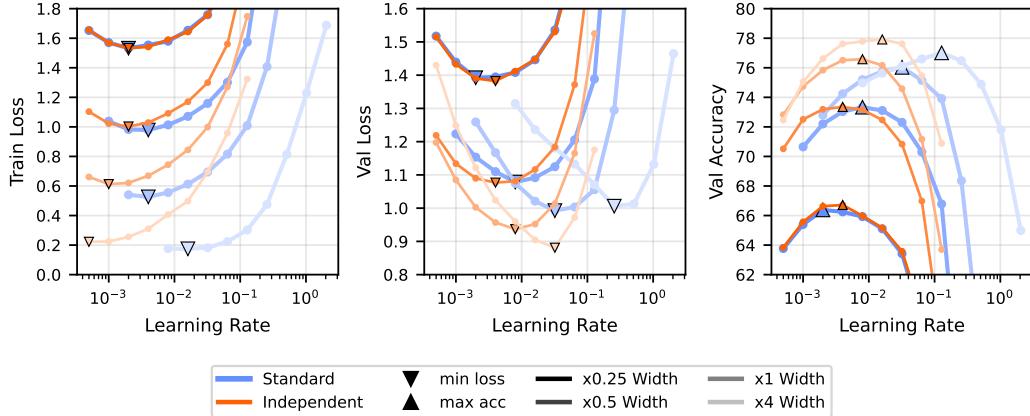
This is already a relatively complicated expression with behavior that depends on the initialization magnitude  $\rho_0$  as well as the exact values of  $\eta$  and  $\lambda$  used. The effective length of the additional warmup can vary from a single step, e.g. when  $\rho_0 \rightarrow 0$ , to the whole training, for example when  $\rho_0$  is large or decays slowly. We note it will also depend significantly on  $\eta\lambda$ , so when sweeping learning rates the length of the warmup effect will differ between points.

Finally we note that simply scaling the learning schedule in for the base configuration by  $s_t$  will generally not give relative updates matching the high weight decay configuration. This is because modifying the learning rate schedule also changes how the norms evolve over time. A scaling factor that would make the profiles match exactly could still be computed using the recurrence relations by keeping track of how the modifications affect the norms over time. However, exactly matching the warmup effect from higher weight decay values may not be that important since it is largely arbitrary as far as we know in the sense that it does not depend on the alignment ratio. Applying the scaling factor from Equation 36 to a relative or rotational optimizer like LionAR (Kosson et al., 2024a) would result in the desired warmup effect.

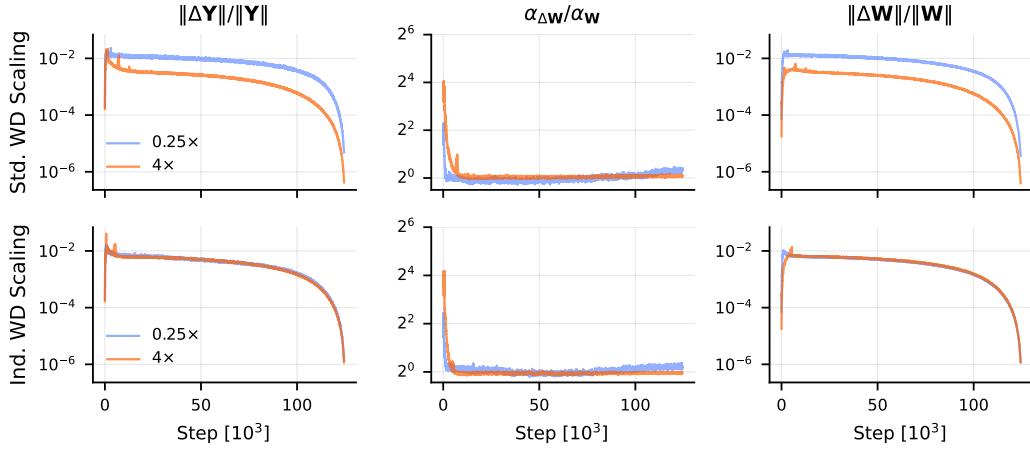
## E RESNET EXPERIMENTS

In this section we repeat some of our experiments for ResNet training to see if they generalize to other architectures and settings. The training setup is described in Appendix F.4 Overall our main conclusions from the LLaMA experiments hold well. The main differences are that the learning rate transfer is worse, likely due to regularization effects and that the warmup effects seem to matter less. The results are shown in the following figures.

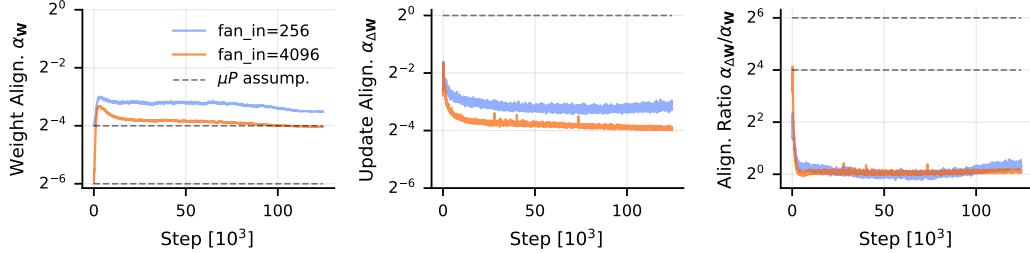
- Figure 11 compares learning rate transfer between standard and independent weight decay scaling. Independent weight decay gives better transfer but there is a shift in the optimal learning rates for both scaling approaches.
- Figure 12 shows that independent weight decay scaling helps preserve relative representation changes. The behavior is essentially identical to the LLaMA setting.
- Figure 13 measures the weight alignment, update alignment and alignment ratio. Just like in the LLaMA setting  $\mu$ P’s assumptions only hold for a very brief period at the start of training. The weight alignment changes more than we observed with LLaMA.
- Figure 14 empirically measures the warmup effect from the use of high weight decay. The conclusions are identical to the LLaMA setting.
- Figure 15 compares learning rate transfer between  $\mu$ P with independent learning rate scaling and not applying  $\mu$ P scaling. We find that both learning rate transfer quality and stability are similar. The  $\mu$ P configuration results in a slightly higher accuracy which may at least in part be due to the warmup effect.



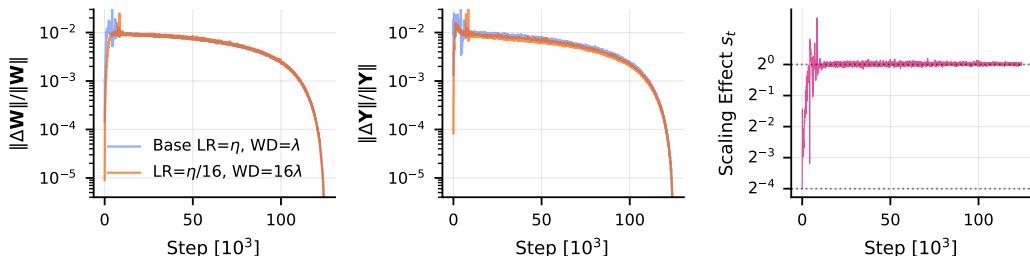
**Figure 11: Comparison of standard vs independent  $\mu$ P scaling for ResNet training on ImageNet.** Learning rates generally transfer much worse than for LLaMA. Independent scaling still gives better results than standard scaling overall. Note how optimal learning rates in terms of training loss shift downwards with independent scaling while they shift upwards for the validation loss and accuracy, a strong indication of a confounding regularization effect. Widths are specified relative to a standard ResNet-50 but the base width for scaling is the  $0.25\times$  variant. Learning rates for parameters that do not scale with width are frozen before the sweep at the optimal value for the  $0.25\times$  network.



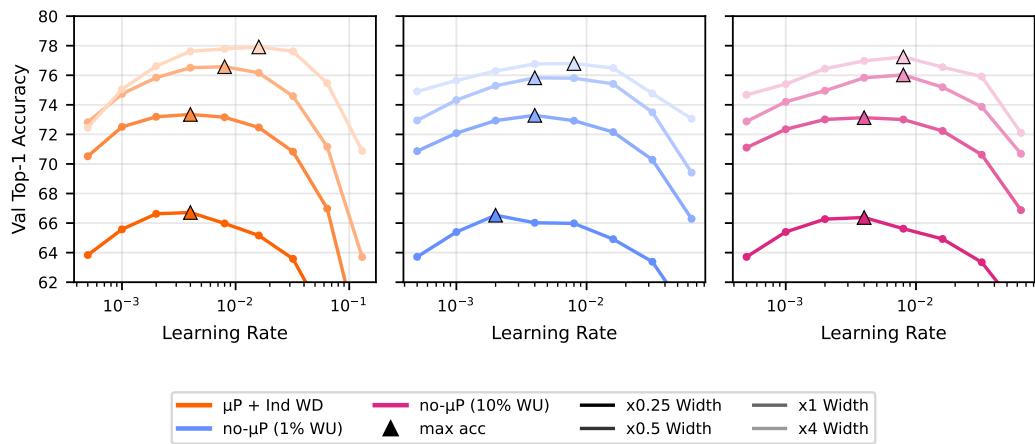
**Figure 12:** ResNet variant of Figure 3. Similar to the LLaMA case, standard weight decay scaling fails to preserve the relative representation change across widths. The alignment ratio is similar across widths contrary to  $\mu$ P’s assumptions, requiring maintaining relative weight updates as achieved by independent weight decay scaling. Plots show for `layer3.2.conv1` which is a  $1 \times 1$  convolutional layer with `fan_in=4096` for the  $4 \times$  variant. The learning rate used here is  $\eta = 4 \times 10^{-3}$  with a 1 epoch warmup (1%) followed by cosine decay to zero.



**Figure 13:** ResNet variant of Figure 4, showing how the alignment assumptions of  $\mu$ P do not hold. Plots show for `layer3.2.conv1` which is a  $1 \times 1$  convolutional layer. Both the weight alignment and update alignment deviate significantly from  $\mu$ P’s assumptions. Note the larger changes in the weight alignment over time compared to the LLaMA experiment. Learning rate  $\eta = 4 \times 10^{-3}$ , no  $\mu$ P scaling applied.



**Figure 14:** ResNet version of Figure 5, showing the warmup effect of higher weight decay configurations. Plots show for `layer3.2.conv1` which is a  $1 \times 1$  convolutional layer. Hyperparameters used are  $\eta = 8 \times 10^{-3}$ ,  $\lambda = 0.1$  corresponding to the best standard WD configuration in Figure 11 (after  $\mu$ P scaling). The additional warmup effect here is comparable in absolute length to the LLaMA setting but proportionally shorter.



**Figure 15:** Learning rate transfer comparison between  $\mu$ P with independent weight decay scaling (left) and no- $\mu$ P scaling for ResNet training (middle/right). The left and middle panels use a 1% linear warmup, the right panel uses a 10% linear warmup (WU). The additional warmup effect of  $\mu$ P with independent weight decay scaling is not needed for stability or transfer, but may contribute to the slightly higher accuracy achieved on the left. Other differences such as the resulting weight norms could also matter, particularly for the final FC layer. For normalized layers the weight norms also inversely scale the gradient norms which may affect the result through interactions with the epsilon value of Adam. A similar gap also appears in Figure 11 between standard and independent weight decay scaling.

---

**Algorithm 1** AdamW (PyTorch variant, [differs from the original](#) by Loshchilov & Hutter (2019))

**Require:** Learning rate  $\eta_t$ , weight decay  $\lambda$ , momentum  $\beta_1$ , magnitude smoothing  $\beta_2$ ,  $\epsilon$  for numerical stability

1: **Initialize:** Time step  $t \leftarrow 0$ , parameter vector  $\theta_0$ , momentum vector  $\mathbf{m}_0 \leftarrow 0$ , magnitude vector  $\mathbf{v}_0 \leftarrow 0$

2: **while** stopping criteria not met :

3:    $t \leftarrow t + 1$

4:    $\mathbf{g}_t \leftarrow$  Mini-batch gradient w.r.t.  $\theta_{t-1}$

5:    $\mathbf{m}_t \leftarrow \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t$

6:    $\mathbf{v}_t \leftarrow \beta_2 \mathbf{v}_{t-1} + (1 - \beta_2) \mathbf{g}_t^2$

7:    $\hat{\mathbf{m}}_t \leftarrow \mathbf{m}_t / (1 - \beta_1^t)$

8:    $\hat{\mathbf{v}}_t \leftarrow \mathbf{v}_t / (1 - \beta_2^t)$

9:    $\theta_t \leftarrow (1 - \eta_t \lambda) \theta_{t-1} - \eta_t \hat{\mathbf{m}}_t / (\sqrt{\hat{\mathbf{v}}_t} + \epsilon)$

---

## F EXPERIMENTAL DETAILS

### F.1 ALGORITHMS

The AdamW variant we use is shown in Algorithm 1. Note how the weight decay hyperparameter in line 9 is scaled by the learning rate. This differs from the original variant described in Loshchilov & Hutter (2019), which is often referred to as “independent” or “fully-decoupled”. This variant only has  $\lambda$  instead of  $\eta\lambda$  in line 9, but the weight decay hyperparameter is still varied according to the same schedule as the learning rate.

### F.2 LEARNING RATE TRANSFER EXPERIMENTS

Our learning rate transfer experiments consist of two phases. We start from an existing hyperparameter configuration and first sweep the global learning rate used across all parameter types, i.e., the input layer, hidden layers, output layer, biases and gains. This is done for the  $1 \times$  scale model. We then freeze the learning rate of the input layer, biases and gains, changing only the learning rate for the layers that scale with the width, i.e., the hidden layer and output layer. This second base learning rate is scaled via  $\mu$ P’s rules before applying it in the optimizer, but the plots show the base rate before scaling (which is what transfers across scale). This second sweep gives the final performance shown in the learning rate transfer plots.

We adopt this approach because it better captures how well the learning rate scaling works by leaving the parameters it is not applied to unaffected. Compared to sweeping a single global learning rate across scales, this typically gives better performance with standard scaling but highlights the shift more. The reason is that without freezing the learning rate for the other layers, they can become unstable at the shifted (higher) base learning rates needed for the hidden layers to learn well. This gives worse overall results since no single global learning rate works well for all parameters anymore, but the optimal learning rates may shift less as a result.

Instead of using the full Maximal Update Parameterization  $\mu$ P for training we combine standard parameterization with the learning rate scaling of  $\mu$ P as proposed by Everett et al. (2024). They find this simpler approach empirically works just as well or better. We note there are many variants of  $\mu$ P that achieve the same effects slightly differently, see for example Tables 3, 8 and 9 in Tensor Programs V (Yang et al., 2021). However, these are generally not equivalent with standard weight decay, which would complicate the use of full  $\mu$ P for our experiments.

For transformer training the main differences between full  $\mu$ P and our experiments are the handling of the final output layer and the normalization factor applied in attention. The output layer is initialized using a smaller standard deviation that scales with  $1/C$  rather than  $1/\sqrt{C}$  like in standard parameterization. This results in an initial relative update size that is not scaled with the width compared to the  $1/\sqrt{m}$  scaling that occurs in the approach from Everett et al. (2024). The other difference is that the scaling factor in the attention heads is  $1/d_{\text{head}}$  rather than  $1/\sqrt{d_{\text{head}}}$ . However we keep  $d_{\text{head}}$  constant in our scaling experiments, opting to scale the number of heads instead. This difference could thus be absorbed into a tunable factor that does not change across scales.

---

### F.3 LLaMA EXPERIMENT CONFIGURATION

Our LLaMA Touvron et al. (2023) experiments are performed using a modified version of Lingua (Videau et al., 2024). The networks have 20 transformer blocks, each consisting of one attention and one MLP sub-block. We vary the embedding dimension  $d_{\text{emb}}$  between 128 and 2048, which we sometimes refer to as  $1\times$  and  $16\times$ . The MLP block has an expansion factor of exactly  $3\times$  giving exact scaling ratios for all layers. We used a fixed attention head size of  $d_{\text{head}} = 128$ , varying the number of heads across scale. All linear layers are initialized using a standard deviation of  $1/\sqrt{C}$ . We do not use embedding tying but the embeddings are still initialized with a standard deviation of  $1/\sqrt{d_{\text{emb}}}$  like the final FC layer (default behavior in Lingua).

The task is standard next token prediction pre-training using a cross-entropy loss. The dataset used is DLM (Li et al., 2024), specifically a 10% subset obtained by using the first global shard from the Hugging Face dataset. We use the llama3 tokenizer<sup>3</sup> with a vocabulary size of 128,256.

Unless otherwise stated we use the PyTorch variant of the AdamW optimizer shown in Algorithm 1. We vary the learning rate between runs but other hyperparameters are fixed at  $\beta_1 = 0.9$ ,  $\beta_2 = 0.95$ ,  $\epsilon = 10^{-8}$ , and  $\lambda = 0.1$  (except when scaled through independent weight decay scaling). Weight decay is applied to all parameters except the gains. We train for 20,000 steps at a global batch size of 256 and sequence length of 4096 giving 1,048,576 ( $2^{20}$ ) tokens in each batch and 20.97 billion tokens in total. This results in 19.23 tokens per non-embedding parameter in the largest configuration we use (embedding dim 2048). Unless otherwise noted, the learning rate schedule consists of a 10% linear warmup from 0 to the specified peak learning rate, followed by a linear decay down to 0. The losses reported are the average of the last 100 training steps unless otherwise specified. This is due to initial issues we had with the evaluation part of the code base.

For an embedding dim of 2048 it takes roughly 14 hours to complete a single run using 8 H200 GPUs. This was not optimized and is slowed down by heavy logging of a wide range of alignment related metrics. Training was performed in bfloat16 with mixed precision.

#### F.3.1 DETAILS FOR FIGURE 1

This experiment follows our base LLaMA training F.3 and learning rate transfer setup F.2. However these are more recent runs with working validation losses. For the weight decay runs, the learning rate for the input layer and gains is frozen at  $\eta = 1.6 \times 10^{-2}$  based on an initial sweep of a global learning rate at the  $1\times$  scale. For the no weight decay runs, this secondary learning rate is set to  $\eta = 6.4 \times 10^{-2}$ . The best loss without weight decay is 2.815 at  $\eta_{\text{base}} = 6.4 \times 10^{-2}$ , for independent weight decay it is 2.801 at  $\eta_{\text{base}} = 8 \times 10^{-3}$ , and for standard weight decay it is 2.798 at  $\eta_{\text{base}} = 6.4 \times 10^{-2}$ .

#### F.3.2 DETAILS FOR FIGURE 3

This experiment shows `layers.13.feed_forward.w1` for a base learning rate  $\eta = 8 \times 10^{-3}$ , the optimal for value for the  $1\times$  width in Figure 1. The learning rate for gains and the input layer is  $\eta = 6.4 \times 10^{-2}$ . The metrics are logged and plotted every 10 steps, the final 0.5% of training is excluded for numerical reasons (the changes go to zero with the learning rate). The representation based metrics are computed for 1/64th of a batch (corresponding to one micro-batch on the master rank). Other details follow the base setup described in F.3.

#### F.3.3 DETAILS FOR FIGURE 4

This experiment shows `layers.13.feed_forward.w1` for a learning rate  $\eta = 4 \times 10^{-3}$  and weight decay  $\lambda = 0.1$ . Note that this is the LR-WD product that gives the best results for standard scaling of the  $16\times$  configuration in Figure 1, after applying the  $\mu\text{P}$  scaling to the LR. No  $\mu\text{P}$  scaling is performed between the two configurations, the hyperparameters are exactly the same and the specified learning rate is applied to all parameters. The metrics are logged and plotted every 10 steps, the final 0.5% of training is excluded for numerical reasons. The representation based metrics are computed for 1/64th of a batch (corresponding to one micro-batch on the master rank). See Figure 16 for the alignment ratio of other layers. Other details follow the base setup described in F.3.

---

<sup>3</sup><https://huggingface.co/meta-llama/Meta-Llama-3-8B>

---

### F.3.4 DETAILS FOR FIGURE 5

This experiment shows `layers.13.feed_forward.w1` for a learning rate  $\eta = 4 \times 10^{-3}$  (after  $\mu$ P scaling) and weight decay  $\lambda = 0.1$ . This corresponds to the best configuration for standard weight decay scaling in Figure 1 and the independent scaling run with the same LR-WD product. Metrics are logged and displayed every 10 steps, the final 0.5% are excluded for numerical reasons, representation changes are computed on 1/64th of a batch. Other details follow the base setup described in F.3.

### F.3.5 DETAILS FOR FIGURE 6

This experiment follows our base setup for LLaMA training F.3 and learning rate transfer experiments F.2. The learning rate for the gains and input layer is fixed at  $\eta = 1.6 \times 10^{-2}$ . The modified learning rate schedules are applied to all parameters, not only the width dependent ones. The scaling factor for `exp50` and `decayaway` varies between widths as in  $\mu$ P, they have no effect at the  $1 \times$  scale. The additional warmup factors are applied on top of the existing 10% warmup and linear decay. For the longer 50% linear warmup, the run corresponding to  $\eta = 4 \times 10^{-3}$  was repeated due to sudden divergence half way through, which did not occur in the repeated run with exactly the same seed and configuration.

## F.4 RESNET EXPERIMENT CONFIGURATION

Our ResNet He et al. (2016) experiments are performed using a modified version of PyTorch Image Models (Wightman, 2019). All networks are 50 layers deep with a bottleneck block configuration, specifically the original post-norm variant but with the downsampling on the residual stream performed on the 3x3 convolutions (“ResNet v1.5”). When scaling the width we scale all hidden dimensions of all layers evenly relative to the original network, going down to  $0.25 \times$  width up to  $4 \times$ . The  $4 \times$  variant is fairly large for a convolutional network with roughly 384M parameters.

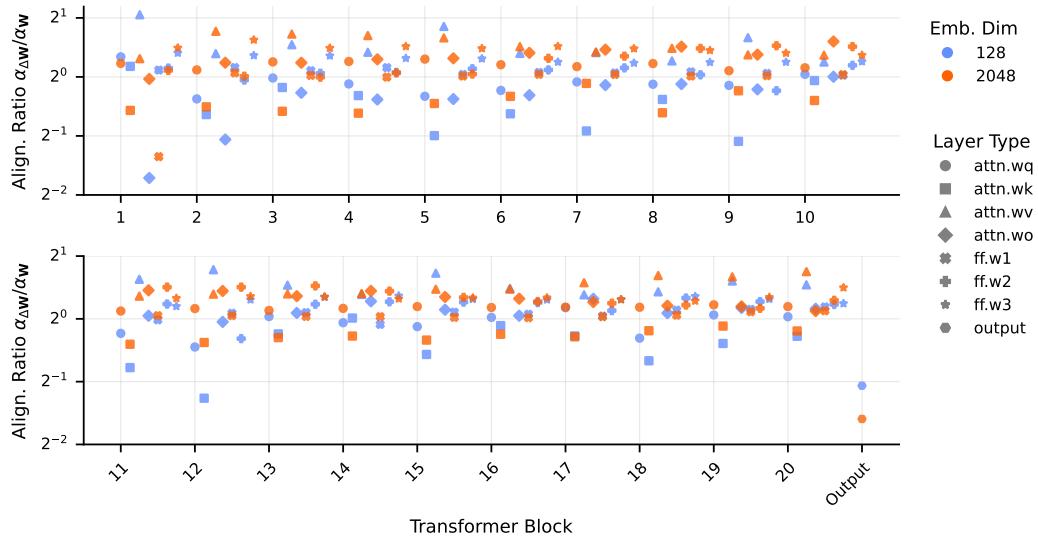
The training task is image classification using cross-entropy loss. We train on ImageNet-1k (Russakovsky et al., 2015) for 100 epochs. Data augmentation consists of random cropping and random horizontal flips.

Optimization is performed using the PyTorch variant of AdamW. The learning rate varies between experiments. Other hyperparameters are  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ ,  $\epsilon = 10^{-8}$ ,  $\lambda = 0.1$  (before independent weight decay scaling). Unless otherwise stated we use a 1 epoch linear warmup from 0 followed by a cosine decay to 0. The batch size is 1024 spread out over 8 GPUs without the use of synchronized batch normalization during training.

The training of the  $4 \times$  variant takes roughly 10 hours to complete using 8 H200 GPUs. This was not optimized and is slowed down by heavy logging of a wide range of alignment related metrics. Training was performed in bfloat16 with mixed precision.

---

## G SUPPLEMENTARY FIGURES



**Figure 16: Halfway through training the alignment ratio of all layers is approximately one.** The alignment ratio for each layer in the LLaMA experiments shown in Figure 4 measured at step 10,000. The ratio is similar across widths, contrary to  $\mu$ P's assumptions.