
Sparse maximal update parameterization: A holistic approach to sparse training dynamics

Nolan Dey Shane Bergsma Joel Hestness
Cerebras Systems
{nolan,joel}@cerebras.net

Abstract

Several challenges make it difficult for sparse neural networks to compete with dense models. First, setting a large fraction of weights to zero impairs forward and gradient signal propagation. Second, sparse studies often need to test multiple sparsity levels, while also introducing new hyperparameters (HPs), leading to prohibitive tuning costs. Indeed, the standard practice is to re-use the learning HPs originally crafted for dense models. Unfortunately, we show sparse and dense networks do not share the same optimal HPs. Without stable dynamics and effective training recipes, it is costly to test sparsity at scale, which is key to surpassing dense networks and making the business case for sparsity acceleration in hardware. A holistic approach is needed to tackle these challenges and we propose sparse maximal update parameterization (S μ Par) as one such approach. For random unstructured static sparsity, S μ Par ensures activations, gradients, and weight updates all scale independently of sparsity level. Further, by reparameterizing the HPs, S μ Par enables the same HP values to be optimal as we vary both sparsity level and model width. HPs can be tuned on small dense networks and transferred to large sparse models, greatly reducing tuning costs. On large-scale language modeling, S μ Par shows increasing improvements over standard parameterization as sparsity increases, leading up to 11.9% relative loss improvement at 99.2% sparsity. A minimal implementation of S μ Par is available at <https://github.com/ElleutherAI/nanoGPT-mup/tree/supar>.

1 Intro

Sparsity has emerged as a key technique to mitigate the increasing computational costs of training and inference in deep neural networks. This work focuses on *weight sparsity*, whereby a significant fraction of model weights are kept at zero. It has long been known that dense neural networks can be heavily pruned *after* training [30]. With the goal of reducing costs *during* training, recent work has explored static weight sparsity from initialization. In this work we focus on random unstructured static sparsity, which has re-emerged as a surprisingly effective strategy [33, 58]. This type of sparsity can be accelerated by CPUs, Cerebras, Graphcore, and SambaNova. Furthermore, GPUs and TPUs support 2:4 block structured sparsity which is quite similar to 50% unstructured sparsity.

Unfortunately, several challenges have hindered progress in weight-sparse neural networks. First, sparsity impairs signal propagation during training [31, 11, 1]. Second, with today’s techniques, sparse training is costly. Sparse techniques typically introduce extra hyperparameters (HPs), e.g., number of pruning iterations at initialization [60, 7, 56], and it is common to train models across different sparsity levels. Since tuning should be performed at each level and the search space grows exponentially with the number of HPs, the tuning costs essentially “defeat the purpose” of sparsity, i.e., to *reduce* computation [60]. Finally, today there is only a nascent ecosystem of hardware acceleration for unstructured sparsity, so most researchers get little sparsity benefit when tuning.

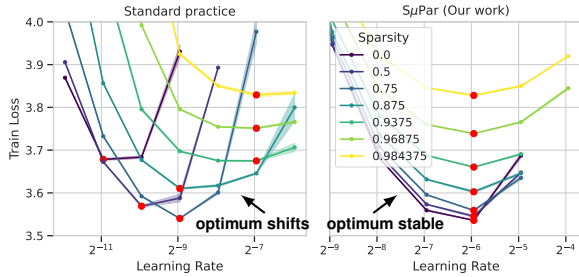


Figure 1: SμPar (Our work) allows stable optimum HPs for any sparsity level, unlike standard practice.

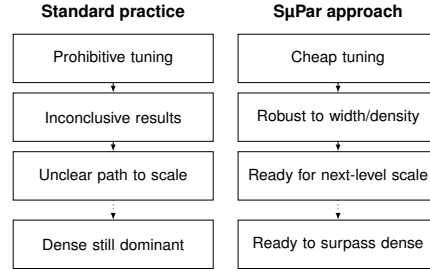


Figure 2: SμPar enables sparse training at scale, helping to surpass dense and motivate sparsity in hardware.

These costs have led to the standard practice of *simply re-using HPs that were previously optimized for the baseline dense models* (Section 2). One might hope that sparse models thrive with the same learning rates and other HPs as their dense counterparts. Unfortunately, they do not: optimal HPs *systematically* vary with sparsity level (Figure 1, left). With impaired training dynamics, prohibitive tuning cost, and lacking the established training recipes enjoyed by dense models, it is often inefficient to train sparse networks at scale (Figure 2).

To remedy this situation, we propose sparse maximal update parameterization (SμPar, pronounced “soo-pahr”), a novel, holistic approach to stabilize sparse training dynamics. SμPar fulfills the Feature Learning Desiderata (Section 3) by parameterizing weight initialization and learning rates with respect to change in width *and* sparsity level. As a generalization of maximal update parameterization (μP) [64, 63], SμPar enjoys well-controlled activation, gradient, and weight update scales in expectation, avoiding exploding or vanishing signal when changing both sparsity and model width.

By reparameterizing HPs in this way, SμPar enables the same HP values to be optimal as sparsity varies (Figure 1, right). We therefore enjoy μTransfer: we can tune small proxy models and transfer optimal HPs directly to models at scale. In fact, we discovered our μP HPs, tuned for dense models in prior work (and equivalent to SμPar with sparsity=0%), correspond to the optimal learning rate and initial weight variance for *all* sparse models tuned in this paper! As sparsity increases, our formulation shows the standard parameterization (SP) and μP suffer from vanishing signal, further clarifying prior observations of gradient flow issues in sparse networks. The improvements enabled by SμPar set the Pareto-frontier best loss across sparsity levels. Figure 3 previews this improvement for large language models trained from compute-optimal configurations [23]. Here, SμPar benefits grow with increasing sparsity, to 11.9% better loss than SP and 1.9% better loss than μP at 99.2% random unstructured sparsity. See Section 4.3 for details on this experiment.

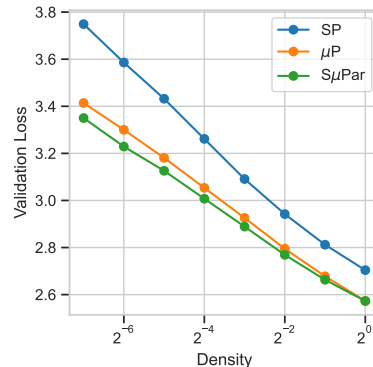


Figure 3: For LLMs, SμPar forms the Pareto frontier loss across sparsity levels, with no HP tuning required.

2 Related work

Sparse training landscape Sparse training can be divided into static sparsity, where the connectivity is fixed (our focus) and dynamic sparsity, where the sparsity mask can evolve [22]. We use *unstructured* sparsity, though our approach generalizes to structured approaches where a particular sparsity pattern increases efficiency on specific hardware [67, 26, 38, 14, 29, 1]. Unstructured connectivity may be based on both random pruning [40, 18, 57, 33, 58] and various pruning-at-initialization criteria [32, 60, 61, 56, 7]. Liu et al. [33] found that as models scale, the relative performance of randomly pruned networks grow. Furthermore, Frantar et al. [15] found the optimal level of sparsity increases with the amount of training data. Together, these findings suggest that as neural networks

continue to get wider and deeper, and trained on more and more data, very sparse randomly-pruned networks may emerge as an attractive option.

Improving sparse training dynamics Many prior works identify various sparse training dynamics issues. In particular, prior works note sparsity impacts weight initialization [35, 31, 49, 11], activation variance [29], gradient flow [61, 37, 57, 11, 1], and step sizes during weight updates [15]. These prior works each only address a subset of these issues in targeted ways, often showing benefits to sparse model training loss. We advocate for a holistic approach, and discuss the relationship between these prior works and our approach in Section 5 after describing and evaluating SμPar.

Sparse sensitivity to HPs Due to the costs of training with fixed weight sparsity, re-using dense HPs is standard practice. Such re-use is typically indicated in appendices or supplemental materials, e.g., [40, 32, 35, 31, 16, 60, 61, 56, 13, 7, 18, 57, 33, 58]. Also, dynamic sparsity approaches often compare to fixed sparsity; these baselines are likewise reported to re-use the dense HPs [2, 41, 10, 34, 11, 59]. However, some prior work has suggested such training is sensitive to HPs, e.g., learning rates [35, 57], learning rate schedules [16], or training length [28], although systematic tuning was not performed. For dynamic sparse training (DST), it is also conventional to re-use dense HPs, whether in dense-to-sparse [37, 15] or sparse-to-sparse (evolving mask) training [2, 8, 34, 11, 59]. As with fixed sparsity, work here has also suggested sensitivity to HPs, e.g., to dropout and label smoothing [16]. DST may also benefit from extra training steps [10] or smaller batch sizes [34], although in DST this may mainly be due to a greater number of opportunities for connectivity exploration [34].

3 Sparse maximal update parameterization (SμPar)

We now provide background, motivation, and derivation for SμPar, first introducing notation (Section 3.1) and then defining Feature Learning Desiderata (Section 3.2) with a brief overview of μP (Section 3.3). Finally we motivate SμPar and provide an overview of the parameterization (Section 3.4).

3.1 Notation

The operations for a single sparse training step are illustrated in Figure 4. The definition and dimensions are: batch size B , learning rate η , loss function \mathcal{L} , forward pass function \mathcal{F} , input dimension d_{in} , input activations $\mathbf{X} \in \mathbb{R}^{B \times d_{\text{in}}}$, input activation gradient $\frac{\partial \mathcal{L}}{\partial \mathbf{X}} = \nabla_{\mathbf{X}} \mathcal{L} \in \mathbb{R}^{B \times d_{\text{in}}}$, output dimension d_{out} , output activations $\mathbf{Y} \in \mathbb{R}^{B \times d_{\text{out}}}$, output activation gradient $\frac{\partial \mathcal{L}}{\partial \mathbf{Y}} = \nabla_{\mathbf{Y}} \mathcal{L} \in \mathbb{R}^{B \times d_{\text{out}}}$, weights $\mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, initialization variance σ_W for weights \mathbf{W} , weight update $\Delta \mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}}$, and $\Delta \mathbf{Y} \in \mathbb{R}^{B \times d_{\text{out}}}$ is the effect of the weight update on output activations: $\Delta \mathbf{Y} = \mathbf{X}(\Delta \mathbf{W} \odot \mathbf{M})$. Unless otherwise specified, $\mathbf{M} \in \{0, 1\}^{d_{\text{in}} \times d_{\text{out}}}$ is an unstructured random static mask with sparsity s and density $\rho = 1 - s$. When changing model scale or sparsity, we refer to a width multiplier $m_d = \frac{d_{\text{in}}}{d_{\text{in, base}}} = \frac{d_{\text{out}}}{d_{\text{out, base}}}$ and density multiplier $m_\rho = \frac{\rho}{\rho_{\text{base}}}$.

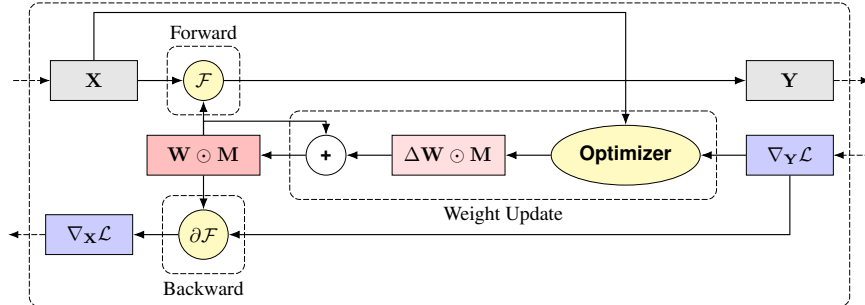


Figure 4: The three operations associated with training a layer with weights that perform the function \mathcal{F} : Forward activation calculation, backward gradient propagation, and the weight update.

If we apply sparsity to a linear layer (i.e., \mathcal{F} is a fully-connected layer), our aim is to control:

1. **Forward pass:** $\mathbf{Y} = \mathcal{F}(\mathbf{X}, \mathbf{W} \odot \mathbf{M}) = \mathbf{X}(\mathbf{W} \odot \mathbf{M})$.

2. **Backward pass:** $\nabla_{\mathbf{X}}\mathcal{L} = (\nabla_{\mathbf{Y}}\mathcal{L}) \cdot (\mathbf{W} \odot \mathbf{M})^\top$.
3. **Effect of weight update $\Delta\mathbf{W}$ on \mathbf{Y} :** $\Delta\mathbf{Y} = \mathbf{X}(\Delta\mathbf{W} \odot \mathbf{M})^1$.

3.2 Feature learning: Defining the goal of μP and SuPar

Prior works [64, 63, 65] introduce the Feature Learning Desiderata (FLD) to ensure stable training dynamics as width is varied. Building on prior works, we include gradients $\nabla_{\mathbf{X}}\mathcal{L}$ in the desiderata.

Feature Learning Desiderata (FLD): For layer l and token i , we desire that $\|\mathbf{Y}_i^l\|_2 = \Theta(\sqrt{d_{\text{out}}})$, $\|\nabla_{\mathbf{X}}\mathcal{L}_i^l\|_2 = \Theta(\sqrt{d_{\text{in}}})$, $\|\Delta\mathbf{Y}_i^l\|_2 = \Theta(\sqrt{d_{\text{out}}})$, $\forall i, \forall l$.

Recall that if all the entries of some vector $\mathbf{v} \in \mathbb{R}^n$ are some constant c , then $\|\mathbf{v}\|_2 = \Theta(\sqrt{n})$ with respect to width n . Therefore we can satisfy the FLD by ensuring the *typical element size* of \mathbf{Y} , $\nabla_{\mathbf{X}}\mathcal{L}$, and $\Delta\mathbf{Y}$ is $\Theta(1)$ with respect to **some variable(s)** we would like to scale. Variables to scale include width [64, 63, 65], depth [66, 4], and sparsity (this work). The FLD prescribes a **holistic** signal propagation approach of controlling each of the three operations in a training step, not a subset².

3.3 Maximal update parameterization (μP)

Here we provide a brief overview of maximal update parameterization (μP) [64, 63, 65]. With the standard parameterization (SP), Yang and Hu [64] show the scale of activations throughout training increases as model width increases, motivating the development of μP . μP [64, 63] is defined as the unique parameterization that satisfies the FLD by ensuring the *typical element size* of \mathbf{Y} , $\nabla_{\mathbf{X}}\mathcal{L}$, and $\Delta\mathbf{Y}$ is $\Theta(1)$ **with respect to change in width** m_d . The FLD can also be satisfied by controlling the spectral norm of weights [65]. μP enables $\mu\text{Transfer}$: the optimum learning rate, initialization weight variance, scalar multipliers, and learning rate schedule all remain consistent as width is increased for μP models [63]. $\mu\text{Transfer}$ can be leveraged to take a *tune small, train large* approach where hyperparameters are extensively tuned for a small model then transferred, enabling reduced tuning budgets and superior tuning for large models compared to standard practice.

3.4 Sparse maximal update parameterization (SuPar)

Yang et al. [63] show activation magnitudes explode with increasing model width. In Figure 5 we show sparsity has the opposite effect: increasing sparsity causes shrinking activation magnitudes.

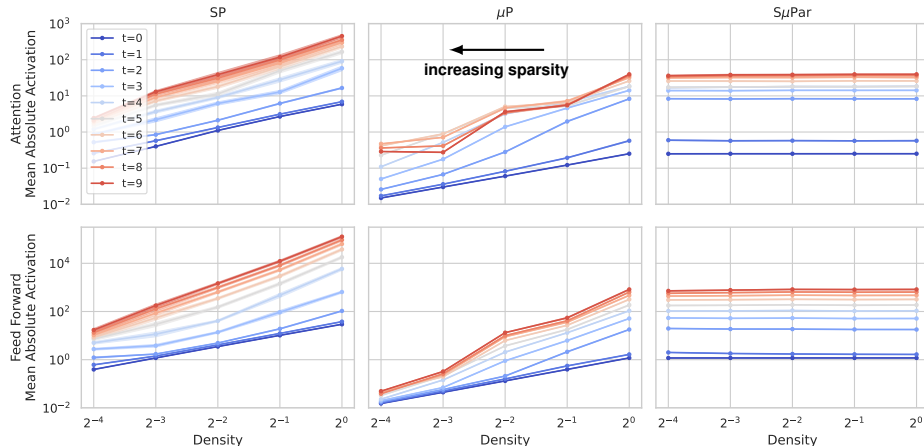


Figure 5: Mean absolute value activations for attention and feed forward blocks after training step t (10 seeds). In SP and μP models, decreasing density causes activations to vanish (note axes on log-scale). In SuPar models, density has little effect on activation scales and there is no vanishing.

¹After a weight update $\Delta\mathbf{W}$ is applied, new output activations can be written as $\mathbf{Y} + \Delta\mathbf{Y} = \mathbf{X}(\mathbf{W} \odot \mathbf{M}) + \mathbf{X}(\Delta\mathbf{W} \odot \mathbf{M})$. Our goal is to control $\Delta\mathbf{Y}$.

²For example, initialization methods alone can only control $\|\mathbf{Y}\|_F$ and $\|\nabla_{\mathbf{X}}\mathcal{L}\|_F$ at the first time step.

Finding 1: Increasing sparsity causes vanishing activations and gradients with both SP and μP .

SuPar is defined as the unique parameterization that satisfies the FLD by ensuring the *typical element size* of \mathbf{Y} , $\nabla_{\mathbf{X}}\mathcal{L}$, and $\Delta\mathbf{Y}$ is $\Theta(1)$ **with respect to change in width m_d and change in density m_ρ** . SuPar enables stable activation scales across sparsity levels (Figure 5, right). In this section, we walk through the changes required to control each of the three operations in a sparse training step, providing an overview of the SuPar derivation. We focus on the AdamW [36] optimizer used in our experiments. For a more detailed derivation, including both SGD and Adam, see Appendix D.

Forward pass at initialization To ensure the *typical element size* of \mathbf{Y} is $\Theta(1)$ with respect to change in width $m_{d_{\text{in}}}$ and change in density m_ρ , we can control the mean and variance of \mathbf{Y}_{ij} . Since at initialization $\mathbb{E}[\mathbf{W}] = 0$, $\mathbb{E}[\mathbf{Y}] = 0$, and $\mathbf{W} \perp \mathbf{Y}$, the mean is controlled. The variance of \mathbf{Y}_{ij} can be written as:

$$\text{Var}(\mathbf{Y}_{ij}) = m_{d_{\text{in}}} d_{\text{in,base}} m_\rho \rho_{\text{base}} \sigma_{\mathbf{W}}^2 (\text{Var}(\mathbf{X}) + \mathbb{E}[\mathbf{X}]^2) \quad (1)$$

To ensure $\text{Var}(\mathbf{Y}_{ij})$ scales independent of $m_{d_{\text{in}}}$ and m_ρ , we choose $\sigma_{\mathbf{W}}^2 = \frac{\sigma_{\mathbf{W},\text{base}}^2}{m_{d_{\text{in}}} m_\rho}$.

Backward gradient pass at initialization To ensure the *typical element size* of $\nabla_{\mathbf{X}}\mathcal{L}$ is $\Theta(1)$ with respect to change in width $m_{d_{\text{out}}}$ and change in density m_ρ , we can control the mean and variance of $\nabla_{\mathbf{X}}\mathcal{L}$. Since at initialization $\mathbb{E}[\mathbf{W}] = 0$, $\mathbb{E}[\nabla_{\mathbf{X}}\mathcal{L}] = 0$ and the mean is controlled³. The variance of $\nabla_{\mathbf{X}}\mathcal{L}_{ij}$ can be written as:

$$\text{Var}(\nabla_{\mathbf{X}}\mathcal{L}_{ij}) = m_{d_{\text{out}}} d_{\text{out,base}} m_\rho \rho_{\text{base}} \sigma_{\mathbf{W}}^2 \text{Var}(\nabla_{\mathbf{Y}}\mathcal{L}) \quad (2)$$

To ensure $\text{Var}(\nabla_{\mathbf{X}}\mathcal{L}_{ij})$ scales independent of $m_{d_{\text{out}}}$ and m_ρ , we choose $\sigma_{\mathbf{W}}^2 = \frac{\sigma_{\mathbf{W},\text{base}}^2}{m_{d_{\text{out}}} m_\rho}$. Typically $m_{d_{\text{out}}} = m_{d_{\text{in}}}$, allowing the same $\sigma_{\mathbf{W}}^2$ to control both forward and backward scales.

Effect of Adam weight update $\Delta\mathbf{W}$ on \mathbf{Y} To ensure the *typical element size* of $\Delta\mathbf{Y}$ is $\Theta(1)$ with respect to change in width $m_{d_{\text{out}}}$ and change in density m_ρ . By the law of large numbers, the expected size of each element can be written as:

$$\mathbb{E}[\Delta\mathbf{Y}_{ij}] \rightarrow \eta m_{d_{\text{in}}} d_{\text{in,base}} m_\rho \rho_{\text{base}} \mathbb{E} \left[\mathbf{X}_{ik} \left(\frac{\sum_t \gamma_t \sum_b \mathbf{X}_{bk}^t \nabla_{\mathbf{Y}} \mathcal{L}_{bj}^t}{\sqrt{\sum_t \omega_t \sum_b (\mathbf{X}_{bk}^t \nabla_{\mathbf{Y}} \mathcal{L}_{bj}^t)^2}} \right) \right], \text{ as } (d_{\text{in}}\rho) \rightarrow \infty \quad (3)$$

To ensure $\Delta\mathbf{Y}_{ij}$ and $\|\Delta\mathbf{Y}\|_F$ are scale invariant to $m_{d_{\text{in}}}$, m_ρ , we choose $\eta = \frac{\eta_{\text{base}}}{m_{d_{\text{in}}} m_\rho}$.

Implementation summary Table 1 summarizes the differences between SP, μP , and SuPar. Since we only sparsify hidden weights, SuPar matches μP for input, output, bias, layer-norm, and attention logits. Also note width multipliers m_d and density multipliers m_ρ are usually the same for all layers, allowing simplified notation. This correction is equivalent to μP [63] when $\rho = 1$ and $m_\rho = 1$. The correction to hidden weight initialization we derive is similar to the sparsity-aware initialization in prior work [35, 49, 11]. SuPar should also easily extend to 2:4 sparsity patterns because, in expectation, the rows and columns of M^l should have equal density. A minimal implementation of SuPar is available at <https://github.com/EleutherAI/nanoGPT-mup/tree/supar>.

4 SuPar Training Results

Here, we present empirical results showing the effectiveness of SuPar over SP and μP when training sparse models. When using SP or μP , optimal HPs drift as we change the sparsity level, possibly leading to inconclusive or even reversed findings. SuPar has stable optimal HPs across both model width and sparsity level, and we show it improves over SP and μP across different scaling approaches. Taken together, we see that SuPar sets the Pareto frontier best loss across all sparsities and widths,

³Although the gradients $\nabla_{\mathbf{Y}}\mathcal{L}$ will have some correlation with weights \mathbf{W} even at initialization, we assume for simplicity that they are fully independent. Future work could investigate this assumption more deeply.

Table 1: Summary of SP, μP , and $\text{S}\mu\text{Par}$ implementations.

Parameterization	SP	μP	$\text{S}\mu\text{Par}$
Embedding Init. Var.	σ_{base}^2	σ_{base}^2	σ_{base}^2
Embedding LR	η_{base}	η_{base}	η_{base}
Embedding Fwd.	$\mathbf{X}^0 \mathbf{W}_{\text{emb}}$	$\alpha_{\text{input}} \cdot \mathbf{X}^0 \mathbf{W}_{\text{emb}}$	$\alpha_{\text{input}} \cdot \mathbf{X}^0 \mathbf{W}_{\text{emb}}$
Hidden Init. Var.	σ_{base}^2	$\sigma_{\text{base}}^2 / m_d$	$\sigma_{\text{base}}^2 / (m_d m_\rho)$
Hidden LR (Adam)	η_{base}	η_{base} / m_d	$\eta_{\text{base}} / (m_d m_\rho)$
Unembedding Fwd.	$\mathbf{X}^L \mathbf{W}_{\text{emb}}^\top$	$\alpha_{\text{output}} \mathbf{X}^L \mathbf{W}_{\text{emb}}^\top / m_d$	$\alpha_{\text{output}} \mathbf{X}^L \mathbf{W}_{\text{emb}}^\top / m_d$
Attention logits	$\mathbf{Q}^\top \mathbf{K} / \sqrt{d_{\text{head}}}$	$\mathbf{Q}^\top \mathbf{K} / d_{\text{head}}$	$\mathbf{Q}^\top \mathbf{K} / d_{\text{head}}$

including when we scale to a large dense model with width equal to GPT-3 XL [5]. Optimal *dense* μP HPs—when adjusted using $\text{S}\mu\text{Par}$ —are also optimal HPs for all sparse models that we test here.

All tests in this section use GPT-like transformer language models [48, 9], trained on the SlimPajama dataset [54] with a 2048 token context length. We apply random unstructured static sparsity to all projection weights in attention and feedforward blocks while keeping embedding, layer normalization, and bias parameters dense. We refer the reader to Appendix E for full methodology of all experiments.

4.1 Sparse hyperparameter transfer

We first show sparsifying a dense model using either SP or μP leads to significant drift in optimal HPs as the sparsity level changes. Figure 6 shows train loss for SP, μP , and $\text{S}\mu\text{Par}$ models when trained with varying sparsity levels and sweeping across different peak learning rates. For the SP configuration, as sparsity increases, the optimal learning rate increases in a somewhat unpredictable way. μP experiences similar shift in optimal learning rate, though shifts are even more abrupt. For $\text{S}\mu\text{Par}$, the optimal learning rate is consistently near 2^{-6} across all sparsity levels.

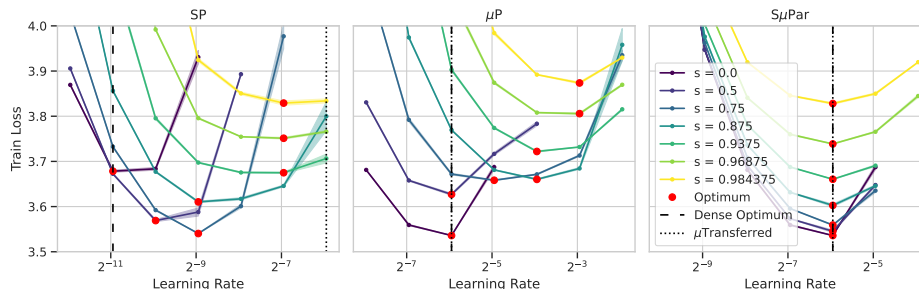


Figure 6: $\text{S}\mu\text{Par}$ ensures stable optimal learning rate for any sparsity s , unlike SP and μP (3 seeds).

We also sweep base weight initialization values and find even more chaotic behaviors for SP and μP with different sparsity levels (Figure 7, left and center, respectively)⁴. μP even shows discontinuities in optimal initialization values at different sparsity levels. We attribute this discontinuity to widely varying expected activation scales between embedding and transformer decoder layers, where embedding activation scales will tend to dominate for high sparsity levels. $\text{S}\mu\text{Par}$ shows consistent optimal initialization (right plot). Figures 6 and 7 demonstrate our second finding.

Finding 2: *With SP and μP , dense and sparse networks do not share the same optimal HPs.*

Figure 8 summarizes our HP transfer tests, showing loss for each parameterization across all sparsities. Even when selecting the best learning rate at each sparsity level for SP and μP , $\text{S}\mu\text{Par}$ (largely) forms the Pareto frontier with an average gap of 0.8% better than SP and 2.1% better than μP .

⁴These results are taken from a point early in training as models with widely varying initialization tend to become unstable later in training.

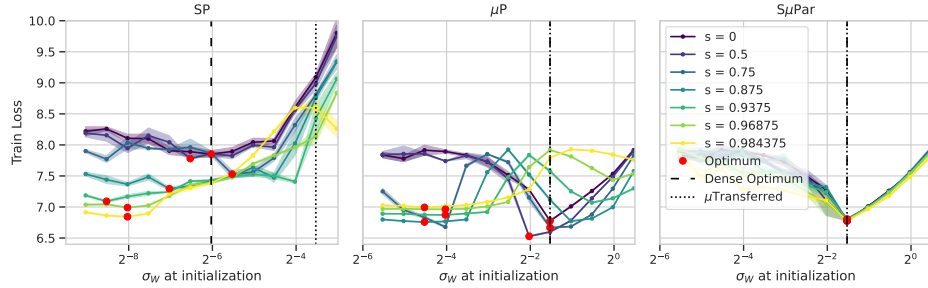


Figure 7: Across sparsity s , SP and μP show unstable optimal initialization. $\text{S}\mu\text{Par}$ is stable (3 seeds).

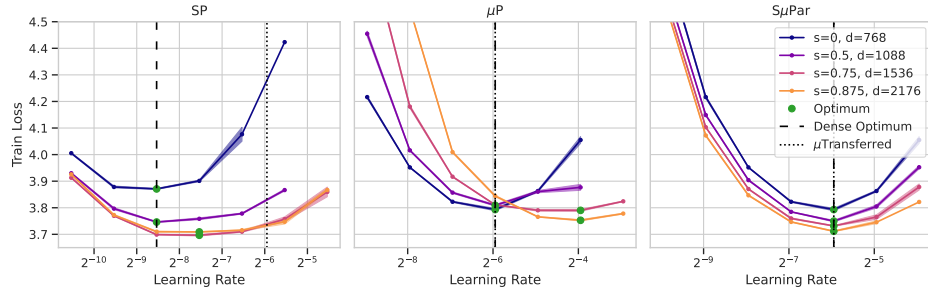


Figure 9: $\text{S}\mu\text{Par}$ ensures stable optimal learning rate in Iso-Parameter sparse + wide scaling (3 seeds).

4.2 Studying $\text{S}\mu\text{Par}$ indicates how some sparse scaling techniques appear to work

So far, we see $\text{S}\mu\text{Par}$ can transfer optimal HPs across sparsity levels, but we have also designed it to transfer HPs across different model widths (hidden sizes), similar to μP . Here, we further demonstrate that $\text{S}\mu\text{Par}$ transfers optimal HPs across width. More generally, sparse scaling that keeps a fixed number of non-zero weights per neuron allows SP and μP to also transfer HPs.

Figure 9 shows learning rate transfer tests when changing both the model’s hidden size, d_{model} , and sparsity level in a common scaling approach called *Iso-Parameter scaling* [18, 10, 59]. Iso-Parameter scaling keeps the model’s number of non-zero parameters approximately the same, as width and sparsity are varied⁵. Here, we see the common result that SP models starting from dense HPs *do* tend to significantly improve as we increase width and sparsity. Note, though, the optimal learning rate for each sparsity level still shifts. When we correct dense HPs using μP or $\text{S}\mu\text{Par}$, the dense baseline significantly improves, but only $\text{S}\mu\text{Par}$ shows consistent loss improvement and stable HPs.

Based on the $\text{S}\mu\text{Par}$ formulation: When the number of non-zero weights per neuron (WPN) in the network is the same, μP and $\text{S}\mu\text{Par}$ become synonymous, because initialization and learning rate adjustment factors will be constant (i.e., $d_{\text{model}} \cdot \rho = \text{WPN} = O(1)$). Optimized SP HPs will also tend to work well. We define this new scaling setting, which we call *Iso-WPN*, to verify this hypothesis. In Figure 11, we test SP HPs with Iso-WPN scaling and see the optimal learning rate stays consistently between 2^{-7} and 2^{-6} with roughly aligned curves (we omit similar μP and $\text{S}\mu\text{Par}$ plots for space, because their corrections are the same). The conclusion is that when scaling SP models in an Iso-WPN sparse setting, HPs should maintain similar training dynamics. More generally, as WPN decreases

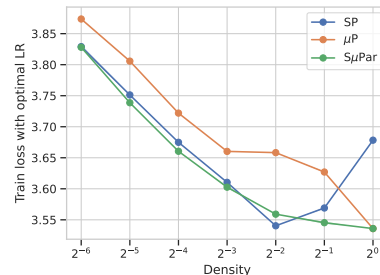


Figure 8: Summarizing loss results from Figure 6 with the optimal learning rate for each parameterization and sparsity.

⁵Not perfectly Iso-Parameter due to unsparsified layers (embedding, bias, layer-norm, etc.)

(e.g., by increasing sparsity), the optimal learning rate will tend to increase proportionally, and vice versa⁶.

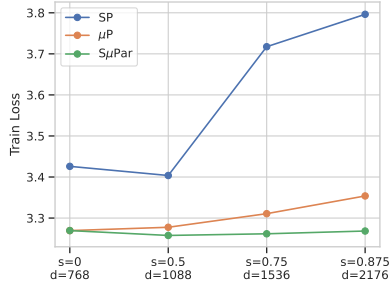


Figure 10: Losses at the end of training when Iso-Parameter scaling.

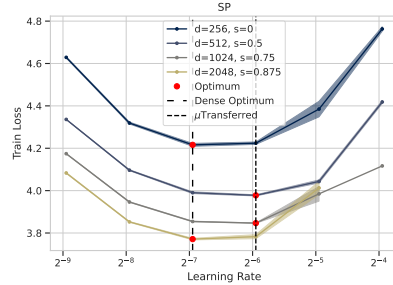


Figure 11: The SP optimized LR is relatively stable with iso-WPN scaling (3 seeds).

Figures 5, 6, 7, and 9 show SμPar is the only parameterization that ensures stable activation scales and stable optimal HPs across model widths and sparsities, satisfying the FLD.

Finding 3: SμPar enables stable activation and stable optimal HPs for any width and sparsity.

4.3 SμPar scaling to large language model pretraining

We conclude this section reflecting on the demonstration of SμPar improvements in a large-scale language model. We train 610M parameter models starting from a Chinchilla [23] compute-optimal training configuration with 20 tokens per parameter from the SlimPajama dataset. This larger model—with hidden size 2048, 10 layers, and attention head size 64—permits sweeping over a larger range of sparsity levels, so we test up to 99.2% sparsity (density 2^{-7}).

Figure 3 shows validation loss for each parameterization as we sweep sparsity levels. Additionally, in Table 2, we evaluate the models from Figure 3 on five downstream tasks: ARC-easy, lambda, RACE, PIQA, and BoolQ, which collectively test for common sense reasoning, world knowledge, and reading comprehension. As sparsity increases, results across pretraining loss and average downstream task accuracy consistently show SP and μP fall farther behind SμPar. Since these models are trained with a large number of tokens, we attribute the widening loss gap mostly to increasingly under-tuned learning rates for SP and μP as sparsity increases—the weight updates lose gradient information throughout training. Figure 8 shows retuning SP and μP could recover some of the gap to SμPar, but that could be costly: These runs take 3-6 hours on a Cerebras CS-3 system (or > 9 days on an NVIDIA A100 GPU).

Finally, returning to the Iso-Parameter scaling setting, Figure 10 shows losses for 111M parameter models trained on 1B tokens and scaled up while using dense optimal HPs. The SP and μP models experience detuning as sparsity increases, allowing SμPar to achieve superior losses⁷.

Finding 4: Sparse networks trained with SμPar improve over SP and μP due to improved tuning.

4.4 Dynamic sparsity hyperparameter transfer

In Figure 12 we test the transfer of optimal learning rate across sparsity levels for two popular dynamic sparse training methods: Rigging the Lottery (RigL) [10]⁸ and Gradual Magnitude Pruning (GMP) [68]⁹. We show that none of SP, μP, or SμPar achieve transfer of optimal learning rate across sparsity levels. For SP and μP we see that higher sparsity levels have higher optimal learning rates.

⁶Our results generalize the Yang et al. finding that optimal LR decreases as width increases [63, Figure 1].

⁷Note this is not an Iso-FLOP comparison because increasing d_{model} also increases attention dot product and embedding FLOPs, which aren't be sparsified. This is so significant that our 87.5% sparse model from Figure 10 has double the training FLOPs of the dense baseline, with virtually unchanged loss.

⁸RigL: Uniform sparsity distribution, drop fraction of 0.3, and mask updates every 100 steps.

⁹GMP: Uniform sparsity distribution, cubic sparsity schedule, and mask updates every 100 steps.

This is because sparsity reduces activation and gradient scales such that a larger learning rate is needed to counteract this. $S\mu$ Par sees the opposite trend where higher sparsity levels have lower optimal learning rates, indicating that $S\mu$ Par is “overcorrecting”.

Dynamic sparse methods can make updates to the weight mask such that the distribution of unmasked/non-zero weights changes to something non-Gaussian, which prevents $S\mu$ Par from being correct in expectation. Compared to random pruning, a mask obtained from magnitude pruning will better preserve the size of activations and gradients seen in the dense network. Since $S\mu$ Par assumes weights are drawn from a Gaussian distribution, $S\mu$ Par ends up “overcorrecting” the initialization and learning rate. In future work it would be impactful to develop a parameterization which generalizes $S\mu$ Par to work for an arbitrary sparse training algorithm.

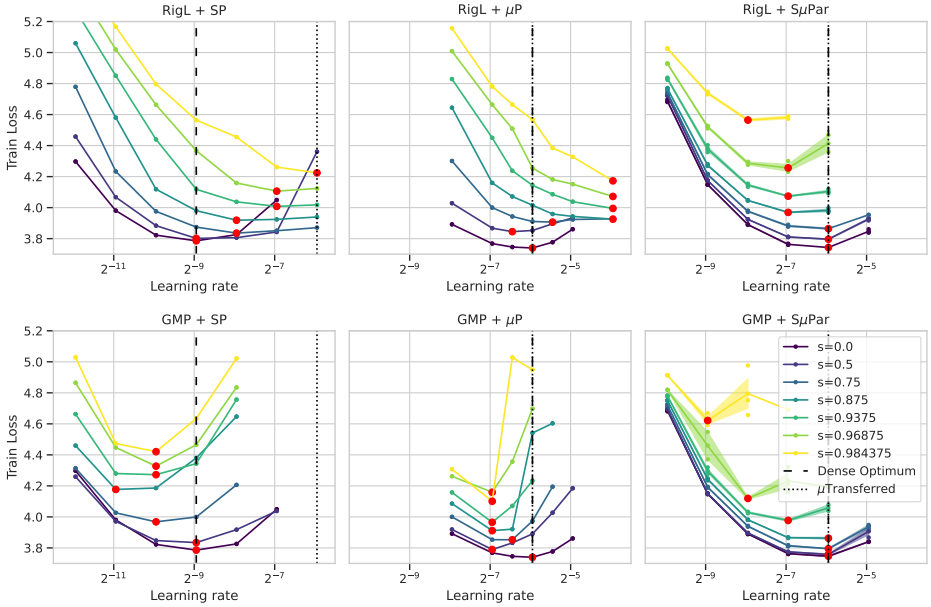


Figure 12: For dynamic sparse training methods RigL and GMP, none of SP, μ P, or $S\mu$ Par achieve stable optimal learning rate across sparsity (3 seeds). Missing points indicate diverged training runs.

5 Discussion

To improve sparse training, prior works make targeted corrections which arise from observations that sparsity can cause degraded activation, gradient, and/or weight update signal propagation. We review these observations and corrections to advocate for holistic control of sparse training dynamics.

Sparsifying Can Cause Vanishing Activations Evci et al. [11] note that by initializing weights using dense methods (e.g., [17, 21]), the “vast majority” of sparse networks have vanishing activations. Lasby et al. [29, App. A] analyze activation variance as a guide for selecting structured sparsity. The FLD suggest activation norms be measured and controlled with respect to sparsity, so activation variance can be considered a proxy to whether sparsity might negatively impact training dynamics. Evci et al. [11] ultimately initialize variances via neuron-specific sparse connectivity, while Liu et al. [35] and Ramanujan et al. [49] propose scaling weight variances proportional to layer sparsity. These corrections, however, only target controlling activations but not weight updates.

Gradient Flow Partially Measures the Weight Update μ Desideratum Sparsity also impairs *gradient flow*—the magnitude of the gradient to the weights—during training [11, 1]. Since gradient flow is measured using the norm of the weight gradients, it measures a piece of the weight update. Unfortunately, gradient flow does not directly measure the effect of the weight update step, which can also involve adjustments for things like optimizer state (e.g., momentum and velocity), the learning rate, and weight decay. Prior works propose techniques to improve gradient flow during sparse

training and pruning by adjusting individual hyperparameters or adding normalization [61, 37, 11, 1]. However, these techniques might overlook the effects of the optimizer and learning rates in weight updates. Notably, Tessera et al. [57] *do* consider some of these effects, but their proposed techniques maintain gradient flow only in the Iso-Parameter scaling setting rather than arbitrary sparsification.

Frantar et al. [15, App. A.1] also endeavor to control weight updates, where they observe diminished step sizes when optimizing sparse networks with Adafactor [52]. They correct this by computing Adafactor’s root-mean-square scaling adjustments over *unpruned* weights and updates. However, such normalization does not prevent activations from scaling with model width [63, 65]. In this sense, sparsity-aware fixes to Adafactor can improve dynamics, but will not address instability holistically. In Figure 14 we show the S μ Par LR correction alone is not even sufficient to achieve stable optimal η .

Weight Initialization Only Controls Dynamics at Initialization We noted works above that adjust sparse weight initializations [11, 35, 49]. Additionally, Lee et al. [31] explore orthogonal weight initialization [46], both before pruning (to ensure SNIP [32] pruning scores are on a similar scale across layers) and after (to improve trainability of the sparse network). While adjusting weights can improve sparse training dynamics at initialization, such adjustments are insufficient to stabilize signals *after multiple steps of training*, in the same way that standard weight initializations fail to stabilize training of dense networks. In Figure 13 we show the S μ Par init. alone is not even sufficient to achieve stable optimal σ_W .

6 Limitations

As Section 4.4 shows, S μ Par requires further extension for dynamic sparse training due to unpredictable changes in weight distributions. The same applies to methods which prune at initialization or after pretraining in a non-random fashion. Iterative magnitude pruning (IMP) is an interesting case since it involves rewinding weights back to their initial values while maintaining the same mask [12]. If the IMP mask at initialization still allows the non-zero weights to have a Gaussian distribution, then S μ Par would apply to this case. Therefore, it’s possible S μ Par *could* prevent “HP detuning” in later IMP iterations, and *potentially* improve IMP losses, though we do not explore this. S μ Par would also work for random structured pruning of entire neurons at initialization because this case simply reduces to training with a narrower dense model.

For weight sparsity more generally, the most pressing limitation is the lack of hardware acceleration [38]. While new software [50, 29, 43] continues to better leverage existing hardware, the growth of software and hardware co-design is also encouraging [59, 6], as effective sparsity techniques can be specifically optimized in deep learning hardware. But to effectively plan hardware, we need to train and test sparse prototypes at next-level sizes, at scales where the optimum sparsity level may be higher than in current networks [15]. Performing such *scaling law*-style studies requires incredible resources even for dense models with well-established training recipes [27, 23]. As S μ Par reduces training and tuning costs, it can help unlock these studies and guide future hardware design.

Finally, the scaling factors for the weight update need to be derived for each optimizer, which might limit the usability of S μ Par in practice. For a discussion of the broader impacts of S μ Par, see Appendix A.

7 Conclusion

Nobody said training with sparsity was easy. We showed that with the standard parameterization and μ P, increasing sparsity level directly correlates with vanishing activations. Impaired training dynamics prevent sparse models from sharing the same optimal hyperparameters, suggesting prior results based on re-use of dense HPs merit re-examination. In contrast, by holistically controlling the training process, S μ Par prevents vanishing activations and enables HP transfer (across both width and sparsity). LLMs trained with S μ Par improve over μ P and the standard parameterization. As such, we hope S μ Par makes things a little easier for sparsity research going forward.

Acknowledgements

We would like to thank Gavia Gray, who provided helpful feedback on the manuscript, and Gurpreet Gosal, who tuned the μ Transferred hyperparameters seen throughout the document.

References

- [1] Abhimanyu Rajeshkumar Bambhaniya, Amir Yazdanbakhsh, Suvinay Subramanian, Sheng-Chun Kao, Shivani Agrawal, Utku Evci, and Tushar Krishna. 2024. Progressive Gradient Flow for Robust N:M Sparsity Training in Transformers. *arXiv preprint arXiv:2402.04744* (2024).
- [2] Guillaume Bellec, David Kappel, Wolfgang Maass, and Robert Legenstein. 2017. Deep rewiring: Training very sparse deep networks. *arXiv preprint arXiv:1711.05136* (2017).
- [3] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.
- [4] Blake Bordelon, Lorenzo Noci, Mufan Bill Li, Boris Hanin, and Cengiz Pehlevan. 2024. Depthwise Hyperparameter Transfer in Residual Networks: Dynamics and Scaling Limit. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=KZJehvRKGD>
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
- [6] Cerebras Systems. 2024. Train a Model with Weight Sparsity. Cerebras Systems Documentation. https://docs.cerebras.net/en/2.1.1/wsc/how_to_guides/sparsity.html Version 2.1.1.
- [7] Pau de Jorge, Amartya Sanyal, Harkirat S Behl, Philip HS Torr, Gregory Rogez, and Puneet K Dokania. 2020. Progressive skeletonization: Trimming more fat from a network at initialization. *arXiv preprint arXiv:2006.09081* (2020).
- [8] Tim Dettmers and Luke Zettlemoyer. 2019. Sparse networks from scratch: Faster training without losing performance. *arXiv preprint arXiv:1907.04840* (2019).
- [9] Nolan Dey, Daria Soboleva, Faisal Al-Khateeb, Bowen Yang, Ribhu Pathria, Hemant Khachane, Shaheer Muhammad, Zhiming Chen, Robert Myers, Jacob Robert Steeves, Natalia Vassilieva, Marvin Tom, and Joel Hestness. 2023. BTLM-3B-8K: 7B Parameter Performance in a 3B Parameter Model. *arXiv:2309.11568* [cs.AI]
- [10] Utku Evci, Trevor Gale, Jacob Menick, Pablo Samuel Castro, and Erich Elsen. 2020. Rigging the lottery: Making all tickets winners. In *International conference on machine learning*. PMLR, 2943–2952.
- [11] Utku Evci, Yani Ioannou, Cem Keskin, and Yann Dauphin. 2022. Gradient flow in sparse neural networks and how lottery tickets win. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 6577–6586.
- [12] Jonathan Frankle and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635* (2018).
- [13] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. 2020. Pruning neural networks at initialization: Why are we missing the mark? *arXiv preprint arXiv:2009.08576* (2020).
- [14] Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*. 10323–10337.

- [15] Elias Frantar, Carlos Riquelme, Neil Houlsby, Dan Alistarh, and Utku Evci. 2023. Scaling laws for sparsely-connected foundation models. *arXiv preprint arXiv:2309.08520* (2023).
- [16] Trevor Gale, Erich Elsen, and Sara Hooker. 2019. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574* (2019).
- [17] Xavier Glorot and Yoshua Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (PMLR)*.
- [18] Anna Golubeva, Behnam Neyshabur, and Guy Gur-Ari. 2020. Are wider nets better given the same number of parameters? *arXiv preprint arXiv:2010.14495* (2020).
- [19] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862* (2019).
- [20] Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. 2018. Sparse dnns with improved adversarial robustness. *Advances in neural information processing systems* 31 (2018).
- [21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.
- [22] Torsten Hoefer, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. 2021. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research* 22, 241 (2021), 1–124.
- [23] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An Empirical Analysis of Compute-optimal Large Language Model Training. In *The Conference on Neural Information Processing Systems (NeurIPS)*.
- [24] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2019. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248* (2019).
- [25] Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. 2020. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058* (2020).
- [26] Hyeong-Ju Kang. 2019. Accelerator-aware pruning for convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 7 (2019), 2093–2103.
- [27] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs.LG]* <https://arxiv.org/abs/2001.08361>
- [28] Denis Kuznedelev, Eldar Kurtic, Eugenia Iofinova, Elias Frantar, Alexandra Peste, and Dan Alistarh. 2023. Accurate neural network pruning requires rethinking sparse optimization. *arXiv preprint arXiv:2308.02060* (2023).
- [29] Mike Lasby, Anna Golubeva, Utku Evci, Mihai Nica, and Yani Ioannou. 2023. Dynamic Sparse Training with Structured Sparsity. *arXiv preprint arXiv:2305.02299* (2023).
- [30] Yann LeCun, John Denker, and Sara Solla. 1989. Optimal brain damage. *Advances in Neural Information Processing Systems* 2 (1989).
- [31] Namhoon Lee, Thalaiyasingam Ajanthan, Stephen Gould, and Philip HS Torr. 2019. A signal propagation perspective for pruning neural networks at initialization. *arXiv preprint arXiv:1906.06307* (2019).
- [32] Namhoon Lee, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018. SNIP: Single-shot network pruning based on connection sensitivity. *arXiv preprint arXiv:1810.02340* (2018).

- [33] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Li Shen, Decebal Constantin Mocanu, and Mykola Wang, Zhangyang and Pechenizkiy. 2022. The unreasonable effectiveness of random pruning: Return of the most naive baseline for sparse training. *arXiv preprint arXiv:2202.02643* (2022).
- [34] Shiwei Liu, Lu Yin, Decebal Constantin Mocanu, and Mykola Pechenizkiy. 2021. Do we actually need dense over-parameterization? In-time over-parameterization in sparse training. In *International Conference on Machine Learning*. PMLR, 6989–7000.
- [35] Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. 2018. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270* (2018).
- [36] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
- [37] Ekdeep Singh Lubana and Robert P Dick. 2020. A gradient flow framework for analyzing network pruning. *arXiv preprint arXiv:2009.11839* (2020).
- [38] Asit Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. 2021. Accelerating sparse deep neural networks. *arXiv preprint arXiv:2104.08378* (2021).
- [39] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.
- [40] Decebal Constantin Mocanu, Elena Mocanu, Phuong H Nguyen, Madeleine Gibescu, and Antonio Liotta. 2016. A topological insight into restricted Boltzmann machines. *Machine Learning* 104 (2016), 243–270.
- [41] Decebal Constantin Mocanu, Elena Mocanu, Peter Stone, Phuong H Nguyen, Madeleine Gibescu, and Antonio Liotta. 2018. Scalable training of artificial neural networks with adaptive sparse connectivity inspired by network science. *Nature communications* 9, 1 (2018), 2383.
- [42] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).
- [43] Neural Magic. 2024. DeepSparse. GitHub repository. <https://github.com/neuralmagic/deepsparse>
- [44] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training Language Models to Follow Instructions With Human Feedback. *arXiv:2203.02155 [cs.CL]* <https://arxiv.org/abs/2203.02155>
- [45] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350* (2021).
- [46] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. 2017. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems* 30 (2017).
- [47] Ofir Press, Noah Smith, and Mike Lewis. 2022. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation. In *International Conference on Learning Representations*.
- [48] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. <https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf>
- [49] Vivek Ramanujan, Mitchell Wortsman, Aniruddha Kembhavi, Ali Farhadi, and Mohammad Rastegari. 2020. What’s hidden in a randomly weighted neural network?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11893–11902.

- [50] Erik Schultheis and Rohit Babbar. 2023. Towards Memory-Efficient Training for Extremely Large Output Spaces – Learning with 670k Labels on a Single Commodity GPU. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 689–704.
- [51] Noam Shazeer. 2020. GLU Variants Improve Transformer. arXiv:2002.05202 [cs.LG] <https://arxiv.org/abs/2002.05202>
- [52] Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*. PMLR, 4596–4604.
- [53] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326* (2019).
- [54] Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. <https://www.cerebras.net/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama>. <https://huggingface.co/datasets/cerebras/SlimPajama-627B>
- [55] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).
- [56] Hidenori Tanaka, Daniel Kunin, Daniel L Yamins, and Surya Ganguli. 2020. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in neural information processing systems* 33 (2020), 6377–6389.
- [57] Kale-ab Tessera, Sara Hooker, and Benjamin Rosman. 2021. Keep the gradients flowing: Using gradient flow to study sparse network optimization. *arXiv preprint arXiv:2102.01670* (2021).
- [58] Vithursan Thangarasa, Abhay Gupta, William Marshall, Tianda Li, Kevin Leong, Dennis DeCoste, Sean Lie, and Shreyas Saxena. 2023. SPDF: Sparse pre-training and dense fine-tuning for large language models. In *Uncertainty in Artificial Intelligence*. 2134–2146.
- [59] Vithursan Thangarasa, Shreyas Saxena, Abhay Gupta, and Sean Lie. 2023. Sparse-IFT: Sparse Iso-FLOP transformations for maximizing training efficiency. *arXiv preprint arXiv:2303.11525* (2023).
- [60] Stijn Verdenius, Maarten Stol, and Patrick Forré. 2020. Pruning via iterative ranking of sensitivity statistics. *arXiv preprint arXiv:2006.00896* (2020).
- [61] Chaoqi Wang, Guodong Zhang, and Roger Grosse. 2020. Picking winning tickets before training by preserving gradient flow. *arXiv preprint arXiv:2002.07376* (2020).
- [62] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems* 36 (2023).
- [63] Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. 2021. Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. In *Advances in Neural Information Processing Systems*.
- [64] Greg Yang and Edward J Hu. 2020. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522* (2020).
- [65] Greg Yang, James B Simon, and Jeremy Bernstein. 2023. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813* (2023).
- [66] Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. 2023. Feature Learning in Infinite Depth Neural Networks. In *International Conference on Learning Representations*.
- [67] Zhuliang Yao, Shijie Cao, Wencong Xiao, Chen Zhang, and Lanshun Nie. 2019. Balanced sparsity for efficient DNN inference on GPU. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5676–5683.
- [68] Michael Zhu and Suyog Gupta. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878* (2017).

A Broader impacts

Sparsity is recognized to reduce carbon emissions [45] and offset well-known environmental and financial costs of large model training [3]. For example, unstructured sparsity can be accelerated by the Cerebras Wafer-Scale Engine¹⁰ and 2:4 block sparsity can be accelerated by NVIDIA Ampere GPUs¹¹. There is growing recognition that HP tuning is a key contributor to these costs. HP tuning is costly, possibly undermining equity in AI research due to financial resources [55]. During model retraining, *sensitivity* to HPs also leads to downstream costs [55]. S μ Par can reduce these costs and sensitivities and thus improve equity.

Sparsity also has potential drawbacks. Hooker et al. [24] showed that even when top-line performance metrics are comparable, pruned networks may perform worse on specific subsets of the data (including on underrepresented groups [25]), may amplify sensitivity to adversarial examples, and may be more sensitive to distribution shift. These sensitivities may depend on the degree of sparsity [20]. It remains an open question whether such drawbacks occur only with pruning or when training with sparsity from scratch (as in S μ Par) [22], and how such sensitivity may impact susceptibility to misuse [62]. We require sparsity-specific methods to detect [53, 42] and mitigate [19, 44] harm. Moreover, since many large models are later pruned for deployment, we recommend testing and documenting in the model card [39] any adverse affects of sparsification at the time of model release.

B Downstream task comparison of parameterizations

In Table 2, we evaluate the models from Figure 3 on five downstream tasks: ARC-easy, lambda, RACE, PIQA, and BoolQ, which collectively test for common sense reasoning, world knowledge, and reading comprehension. We also specifically chose tasks that are easy enough for even extremely sparse models to significantly outperform random chance.

Sparsity	-	0			0.5			0.75			0.875		
	Rand.	SP	μ P	S μ Par	SP	μ P	S μ Par	SP	μ P	S μ Par	SP	μ P	S μ Par
ARC-easy	0.25	0.49	0.51	0.51	0.45	0.49	0.48	0.44	0.45	0.45	0.42	0.43	0.44
LAMBADA	0.00	0.32	0.36	0.36	0.27	0.31	0.32	0.22	0.27	0.28	0.20	0.23	0.25
RACE	0.25	0.30	0.30	0.30	0.29	0.31	0.30	0.28	0.30	0.29	0.27	0.28	0.28
PIQA	0.50	0.67	0.67	0.67	0.63	0.65	0.67	0.63	0.64	0.64	0.62	0.63	0.63
BoolQ	0.50	0.53	0.57	0.57	0.58	0.55	0.51	0.57	0.62	0.52	0.61	0.62	0.62
Avg.	0.30	0.46	0.48	0.48	0.44	0.46	0.46	0.43	0.46	0.44	0.42	0.44	0.44

Sparsity	-	0.9375			0.96875			0.984375			0.992188		
	Rand.	SP	μ P	S μ Par	SP	μ P	S μ Par	SP	μ P	S μ Par	SP	μ P	S μ Par
ARC-easy	0.25	0.41	0.40	0.43	0.39	0.42	0.41	0.38	0.38	0.41	0.37	0.38	0.38
LAMBADA	0.00	0.19	0.20	0.21	0.16	0.18	0.19	0.13	0.15	0.17	0.12	0.13	0.14
RACE	0.25	0.25	0.27	0.28	0.24	0.25	0.27	0.24	0.24	0.25	0.25	0.24	0.26
PIQA	0.50	0.61	0.61	0.61	0.60	0.61	0.60	0.59	0.60	0.59	0.58	0.60	0.60
BoolQ	0.50	0.62	0.62	0.61	0.57	0.62	0.62	0.45	0.62	0.61	0.41	0.61	0.61
Avg.	0.30	0.42	0.42	0.43	0.39	0.42	0.42	0.36	0.40	0.41	0.34	0.39	0.40

Table 2: Downstream evaluation accuracy; higher is better. S μ Par performs best or within 0.01 of best across all sparsity levels and tasks, except *boolq* at 50% and 75% sparsity. Even at 99% sparsity, S μ Par models maintain 40%+ average accuracy, whereas the SP model drops to 34%, close to the 30% accuracy of the random baseline.

C Individual ablations of S μ Par initialization and learning rate corrections

In Figures 13 and 14, we individually ablate the effect of the S μ Par initialization and the S μ Par learning rate. We show that using only the S μ Par initialization in conjunction with μ P (μ P + S μ Par initialization only) does not allow for transfer of optimal initialization standard deviation or optimal learning rate across sparsity levels. We also show that using only the S μ Par learning rate in conjunction with μ P does not achieve transfer either. Therefore, **both** the S μ Par initialization and learning rate corrections are required to achieve optimal hyperparameter transfer across sparsity levels.

¹⁰<https://www.cerebras.net/blog/harnessing-the-power-of-sparsity-for-large-gpt-ai-models>

¹¹<https://www.nvidia.com/en-us/data-center/ampere-architecture/>

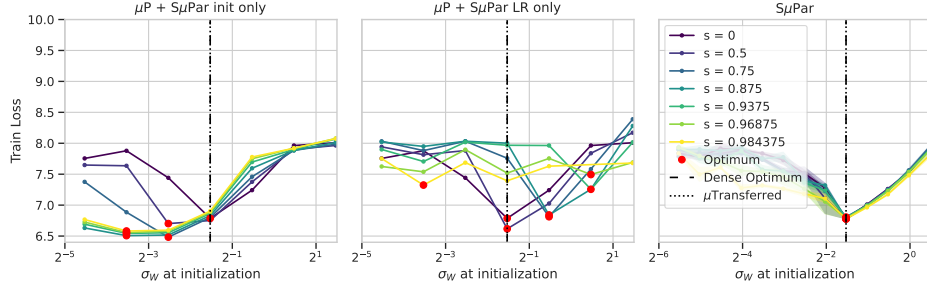


Figure 13: $S_{\mu}Par$ ensures stable optimal weight initialization standard deviation, unlike SP, μP , $\mu P + S_{\mu}Par$ init only, and $\mu P + S_{\mu}Par$ LR only.

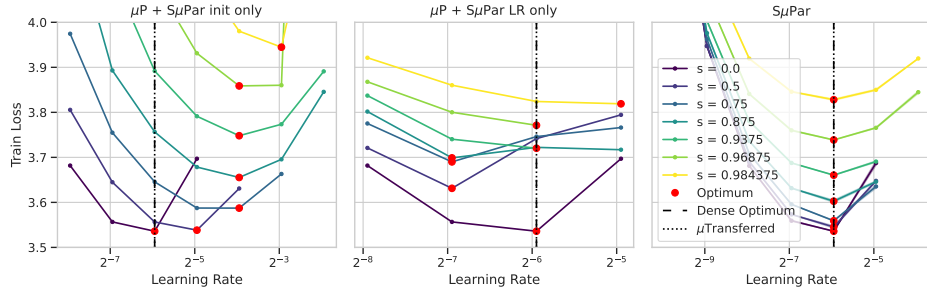


Figure 14: $S_{\mu}Par$ ensures stable optimal learning rate (**Bottom**), unlike SP, μP , $\mu P + S_{\mu}Par$ init only, and $\mu P + S_{\mu}Par$ LR only.

D $S_{\mu}Par$ detailed derivation

D.1 Forward pass at initialization

The first stage where we would like to control training dynamics is in the layer’s forward function. For a random unstructured sparsity mask \mathbf{M} , since each *column* of \mathbf{M} has $d_{in}\rho$ non-zero elements in expectation, we can rewrite the forward pass as:

$$\mathbf{Y}_{ij} = [\mathbf{X}(\mathbf{W} \odot \mathbf{M})]_{ij} = \sum_{q=1}^{d_{in}} \mathbf{X}_{iq} (\mathbf{W}_{qj} \cdot \mathbf{M}_{qj}) = \sum_{k:\mathbf{M}_{kj}=1}^{d_{in}\rho} \mathbf{X}_{ik} \mathbf{W}_{kj} \quad (4)$$

To satisfy the FLD, we desire the *typical element size* of \mathbf{Y} is $\Theta(1)$ with respect to change in width $m_{d_{in}}$ and change in density m_{ρ} . To achieve this we can ensure the mean and variance of \mathbf{Y}_{ij} are invariant to $m_{d_{in}}$ and m_{ρ} .

Mean: As expectation is linear and \mathbf{X} and \mathbf{W} are independent at initialization:

$$\mathbb{E}[\mathbf{Y}_{ij}] = \mathbb{E} \left[\sum_{k:\mathbf{M}_{kj}=1}^{d_{in}\rho} \mathbf{Y}_{ik} \mathbf{W}_{kj} \right] = \sum_{k:\mathbf{M}_{kj}=1}^{d_{in}\rho} \mathbb{E}[\mathbf{X}_{ik} \mathbf{W}_{kj}] = \sum_{k:\mathbf{M}_{kj}=1}^{d_{in}\rho} \mathbb{E}[\mathbf{X}_{ik}] \mathbb{E}[\mathbf{W}_{kj}] \quad (5)$$

Therefore, since at initialization $\mathbb{E}[\mathbf{W}_{ij}] = 0$, $\mathbb{E}[\mathbf{Y}_{ij}] = 0$ and the mean is controlled.

Variance: As expectation is linear and each weight element is IID:

$$\text{Var}(\mathbf{Y}_{ij}) = \text{Var} \left(\sum_{k:\mathbf{M}_{kj}=1}^{d_{in}\rho} \mathbf{X}_{ik} \mathbf{W}_{kj} \right) = \sum_{k:\mathbf{M}_{kj}=1}^{d_{in}\rho} \text{Var}(\mathbf{X}_{ik} \mathbf{W}_{kj}) \quad (6)$$

Then, since \mathbf{X} and \mathbf{W} are independent at initialization:

$$\text{Var}(\mathbf{Y}_{ij}) = \sum_{k:\mathbf{M}_{kj}=1}^{d_{\text{in}}\rho} (\text{Var}(\mathbf{X}_{ik}) + \mathbb{E}[\mathbf{X}_{ik}]^2)(\text{Var}(\mathbf{W}_{kj}) + \mathbb{E}[\mathbf{W}_{kj}]^2) - (\mathbb{E}[\mathbf{X}_{ik}]\mathbb{E}[\mathbf{W}_{kj}])^2 \quad (7)$$

Finally, since at initialization $\mathbb{E}[\mathbf{W}_{kj}] = 0$ and redefining $\text{Var}(\mathbf{W}_{kj}) = \sigma_{\mathbf{W}}^2$:

$$\text{Var}(\mathbf{Y}_{ij}) = \sum_{k:\mathbf{M}_{kj}=1}^{d_{\text{in}}\rho} (\text{Var}(\mathbf{X}_{ik}) + \mathbb{E}[\mathbf{X}_{ik}]^2)\text{Var}(\mathbf{W}_{kj}) = d_{\text{in}}\rho\sigma_{\mathbf{W}}^2(\text{Var}(\mathbf{X}) + \mathbb{E}[\mathbf{X}]^2) \quad (8)$$

Rewriting in terms of multipliers for the width $m_{d_{\text{in}}} = \frac{d_{\text{in}}}{d_{\text{in,base}}}$ and the change in density $m_{\rho} = \frac{\rho}{\rho_{\text{base}}}$:

$$\text{Var}(\mathbf{Y}_{ij}) = m_{d_{\text{in}}}d_{\text{in,base}}m_{\rho}\rho_{\text{base}}\sigma_{\mathbf{W}}^2(\text{Var}(\mathbf{X}) + \mathbb{E}[\mathbf{X}]^2) \quad (9)$$

Solution: To satisfy the FLD and ensure $\text{Var}(\mathbf{Y}_{ij})$ scales independently of $m_{d_{\text{in}}}$ and m_{ρ} , we choose to set $\sigma_{\mathbf{W}}^2 = \frac{\sigma_{\mathbf{W},\text{base}}^2}{m_{d_{\text{in}}}m_{\rho}}$. This ensures typical entry size of \mathbf{Y} is invariant to changes in width $m_{d_{\text{in}}}$ and density m_{ρ} .

Note that this correction is equivalent to μP [63] when $m_{\rho} = 1$. Further, the sparsity factor in the denominator matches the correction for sparsity-aware initialization from Evci et al. [11].

D.2 Backward gradient pass at initialization

The next stage we would like to control training dynamics is in the layer's backward pass. For a random unstructured sparsity mask \mathbf{M} , since each row of \mathbf{M} has $d_{\text{out}}\rho$ non-zero elements in expectation, we can rewrite the backward pass as:

$$\nabla_{\mathbf{X}}\mathcal{L}_{ij} = [\nabla_{\mathbf{Y}}\mathcal{L}(\mathbf{W} \odot \mathbf{M})^{\top}]_{ij} = \sum_q^{d_{\text{out}}} \nabla_{\mathbf{Y}}\mathcal{L}_{iq}(\mathbf{W}_{jq} \cdot \mathbf{M}_{jq}) = \sum_{k:\mathbf{M}_{jk}=1}^{d_{\text{out}}\rho} \nabla_{\mathbf{Y}}\mathcal{L}_{ik}\mathbf{W}_{jk} \quad (10)$$

To satisfy the FLD, we desire the *typical element size* of $\nabla_{\mathbf{X}}\mathcal{L}$ is $\Theta(1)$ with respect to change in width $m_{d_{\text{out}}}$ and change in density m_{ρ} . To achieve this, we can ensure the mean and variance of $\nabla_{\mathbf{X}}\mathcal{L}$ are invariant to $m_{d_{\text{out}}}$ and m_{ρ} .

Mean: Although the gradients $\nabla_{\mathbf{Y}}\mathcal{L}$ will have some correlation with weights \mathbf{W} even at initialization, we assume for simplicity that they are fully independent. Future work could investigate this assumption more deeply. As expectation is linear:

$$\mathbb{E}[\nabla_{\mathbf{X}}\mathcal{L}_{ij}] = \mathbb{E}\left[\sum_{k:\mathbf{M}_{jk}=1}^{d_{\text{out}}\rho} \nabla_{\mathbf{Y}}\mathcal{L}_{ik}\mathbf{W}_{jk}\right] = \sum_{k:\mathbf{M}_{jk}=1}^{d_{\text{out}}\rho} \mathbb{E}[\nabla_{\mathbf{Y}}\mathcal{L}_{ik}\mathbf{W}_{jk}] = \sum_{k:\mathbf{M}_{jk}=1}^{d_{\text{out}}\rho} \mathbb{E}[\nabla_{\mathbf{Y}}\mathcal{L}_{ik}]\mathbb{E}[\mathbf{W}_{jk}] \quad (11)$$

Therefore, since at initialization $\mathbb{E}[\mathbf{W}_{jk}] = 0$, $\mathbb{E}[\nabla_{\mathbf{X}}\mathcal{L}_{ij}] = 0$, the mean is controlled.

Variance: As expectation is linear and each weight element is IID:

$$\text{Var}(\nabla_{\mathbf{X}}\mathcal{L}_{ij}) = \text{Var}\left(\sum_{k:\mathbf{M}_{jk}=1}^{d_{\text{out}}\rho} \nabla_{\mathbf{Y}}\mathcal{L}_{ik}\mathbf{W}_{jk}\right) = \sum_{k:\mathbf{M}_{jk}=1}^{d_{\text{out}}\rho} \text{Var}(\nabla_{\mathbf{Y}}\mathcal{L}_{ik}\mathbf{W}_{jk}) \quad (12)$$

From the backward pass mean derivation, we know $\mathbb{E}[\nabla_{\mathbf{Y}}\mathcal{L}_{ij}] = 0$. Then, similar to the forward pass variance derivation, we can simplify using the facts that at initialization, $\nabla_{\mathbf{Y}}\mathcal{L}$ and \mathbf{W} are (roughly) independent and $\mathbb{E}[\mathbf{W}] = 0$. Similarly we can also define $\text{Var}(\mathbf{W}_{kj}^i) = \sigma_{\mathbf{W}}^2$ and rewrite in terms of width multiplier $m_{d_{\text{out}}} = \frac{d_{\text{out}}}{d_{\text{out,base}}}$ and changes in density $m_{\rho} = \frac{\rho}{\rho_{\text{base}}}$:

$$\text{Var}(\nabla_{\mathbf{X}}\mathcal{L}_{ij}) = m_{d_{\text{out}}}d_{\text{out,base}}m_{\rho}\rho_{\text{base}}\sigma_{\mathbf{W}}^2\text{Var}(\nabla_{\mathbf{Y}}\mathcal{L}) \quad (13)$$

Solution: To satisfy the FLD and ensure $\text{Var}(\nabla_{\mathbf{X}}\mathcal{L}_{ij})$ scales independently of $m_{d_{\text{out}}}$ and m_{ρ} , we choose to set $\sigma_{\mathbf{W}}^2 = \frac{\sigma_{\mathbf{W},\text{base}}^2}{m_{d_{\text{out}}}m_{\rho}}$. This ensures the typical entry size of $\nabla_{\mathbf{X}}\mathcal{L}$ is invariant to changes

in width $m_{d_{\text{out}}}$ and density m_ρ . Typically, we scale model width such that $m_{d_{\text{out}}} = m_{d_{\text{in}}}$. This equal scaling allows the same initialization variance to correct both forward activation and backward gradient scales, making them independent of width. Further, since we assume random sparsity across layer’s weights, the sparsity initialization correction factor, m_ρ , is the same for both the forward activations and backward gradients.

D.3 Effect of weight update $\Delta \mathbf{W}$ on \mathbf{Y}

To satisfy the FLD, we desire the *typical element size* of the weight update $\Delta \mathbf{Y}$ is $\Theta(1)$ with respect to change in width $m_{d_{\text{in}}}$ and change in density m_ρ . To achieve this we examine the expected size of each element. Here, we use η to be the learning rate for this layer. For a random unstructured sparsity mask \mathbf{M} , since each *column* of \mathbf{M} has $d_{\text{in}}\rho$ non-zero elements in expectation:

$$\Delta \mathbf{Y}_{ij} = [\eta \mathbf{X}(\Delta \mathbf{W} \odot \mathbf{M})]_{ij} = \eta \sum_{q=1}^{d_{\text{in}}} \mathbf{X}_{iq}(\Delta \mathbf{W}_{qj} \cdot \mathbf{M}_{qj}) = \eta \sum_{k:\mathbf{M}_{kj}=1}^{d_{\text{in}}\rho} \mathbf{X}_{ik} \Delta \mathbf{W}_{kj} \quad (14)$$

Mean: As expectation in linear:

$$\mathbb{E}[\Delta \mathbf{Y}_{ij}] = \mathbb{E} \left[\eta \sum_{k:\mathbf{M}_{kj}=1}^{d_{\text{in}}\rho} \mathbf{X}_{ik} \Delta \mathbf{W}_{kj} \right] = \eta \sum_{k:\mathbf{M}_{kj}=1}^{d_{\text{in}}\rho} \mathbb{E}[\mathbf{X}_{ik} \Delta \mathbf{W}_{kj}] \quad (15)$$

Since $\Delta \mathbf{W}$ was derived from \mathbf{X} , there is covariance between these variables and $\mathbb{E}[\mathbf{X}_{ik} \Delta \mathbf{W}_{kj}]$ is non-zero. By the Law of Large Numbers:

$$\mathbb{E}[\Delta \mathbf{Y}_{ij}] \rightarrow \eta d_{\text{in}} \rho \mathbb{E}[\mathbf{X}_{ik} \Delta \mathbf{W}], \text{ as } (d_{\text{in}} \rho) \rightarrow \infty \quad (16)$$

Rewriting in terms of width and density multipliers:

$$\mathbb{E}[\Delta \mathbf{Y}_{ij}] \rightarrow \eta m_{d_{\text{in}}} d_{\text{in,base}} m_\rho \rho_{\text{base}} \mathbb{E}[\mathbf{X}_{ik} \Delta \mathbf{W}], \text{ as } (d_{\text{in}} \rho) \rightarrow \infty \quad (17)$$

Equation 17 will be used as intermediate result in the following sections.

D.3.1 Effect SGD weight update $\Delta \mathbf{W}$ on \mathbf{Y}

Following the formulation in [63], stochastic gradient descent (SGD) weight updates take the form:

$$\Delta \mathbf{W}_{kj}^l = \left[\frac{(\mathbf{X})^\top (\nabla_{\mathbf{Y}} \mathcal{L})}{d_{\text{in}}} \right]_{kj} = \frac{1}{d_{\text{in}}} \sum_{b=1}^B \mathbf{X}_{bk} (\nabla_{\mathbf{Y}} \mathcal{L})_{bj} \quad (18)$$

so we can rewrite Equation 17 as:

$$\mathbb{E}[\Delta \mathbf{Y}_{ij}] \rightarrow \eta m_\rho \rho_{\text{base}} \mathbb{E} \left[\mathbf{X}_{ik} \sum_{b=1}^B \mathbf{X}_{bk} (\nabla_{\mathbf{Y}} \mathcal{L})_{bj} \right], \text{ as } (d_{\text{in}} \rho) \rightarrow \infty \quad (19)$$

Solution: For SGD updates, to satisfy the FLD and ensure $\mathbb{E}[\Delta \mathbf{Y}_{ij}]$ and the typical entry size of $\Delta \mathbf{Y}$ are scale invariant to m_d and m_ρ , we choose $\eta = \eta_{\text{base}}/m_\rho$. Note this correction is equivalent to $\mu \mathbf{P}$ [63] when $\rho = 1, m_\rho = 1$.

D.3.2 Effect of Adam weight update $\Delta \mathbf{W}$ on \mathbf{Y}

Following the formulation in Yang et al. [63], Adam weight updates take the form:

$$\Delta \mathbf{W}_{kj} = \frac{\sum_t^T \gamma_t \sum_b^B \mathbf{X}_{bk}^{l,t} (\nabla_{\mathbf{Y}} \mathcal{L})_{bj}^t}{\sqrt{\sum_t^T \omega_t \sum_b^B (\mathbf{X}_{bk}^t (\nabla_{\mathbf{Y}} \mathcal{L})_{bj}^t)^2}} \quad (20)$$

where T is the current training step and γ_t, ω_t are the moving average weights at each training step. Here, we can just consider the weight update associated with an unpruned weight, since a pruned

weight will have value and update 0 (i.e., pruned weights trivially satisfy that their effect on forward activations cannot depend on width or sparsity). We can rewrite Equation 17 as:

$$\mathbb{E}[\Delta \mathbf{Y}_{ij}] \rightarrow \eta m_{d_{\text{in}}} d_{\text{in,base}} m_{\rho} \rho_{\text{base}} \mathbb{E} \left[\mathbf{X}_{ik} \left(\frac{\sum_t \gamma_t \sum_b \mathbf{X}_{bk}^t \nabla_{\mathbf{Y}} \mathcal{L}_{bj}^t}{\sqrt{\sum_t \omega_t \sum_b (\mathbf{X}_{bk}^t \nabla_{\mathbf{Y}} \mathcal{L}_{bj}^t)^2}} \right) \right], \text{ as } (d_{\text{in}} \rho) \rightarrow \infty \quad (21)$$

Solution: For Adam updates, to satisfy the FLD and ensure $\mathbb{E}[\Delta \mathbf{Y}_{ij}]$ and the typical entry size of $\Delta \mathbf{Y}$ are scale invariant to $m_{d_{\text{in}}}$ and m_{ρ} , we choose $\eta = \frac{\eta_{\text{base}}}{m_{d_{\text{in}}} m_{\rho}}$. Note that this correction is equivalent to μP [63] when $\rho = 1, m_{\rho} = 1$.

D.4 Additional notes about derivation

We make a few supplementary notes about the above derivation:

- Throughout our derivation, we use ρ to refer to the density level. Note that since this derivation is local to a single layer in the model, the density (or sparsity) level can also be parameterized independently for each layer. If a sparsity technique will use layer-wise independent sparsity levels, appropriate corrections should be made for each layer.
- Similar to the ρ notation, we use η to denote the learning rate, but this learning rate can be layer-specific depending on sparsity level. Appropriate corrections must be made if using layer-wise independent sparsities.
- The use of the Law of Large Numbers in portions of the above derivation indicate that $\text{S}\mu\text{P}$ ar is expected to provide stable training dynamics as the number of non-zero weights per neuron (WPN) tends to infinity. However, in sparse settings, the WPN can tend to be small. If WPN is small, training dynamics may be affected, and this might be a direction for future work.
- In this work, we only consider sparsifying linear projection layers. As a result, $\text{S}\mu\text{P}$ ar matches μP for other layers like input, output, bias, layer-norm, and attention logits. Depending on the sparsification technique, these other layers might need to be reviewed for their effects on training dynamics.

E Experimental details

S μ Par base hyperparameter tuning To find the optimized set of hyperparameters for $\text{S}\mu\text{P}$ ar, we actually tune μP HPs on a dense proxy model. By formulation of $\text{S}\mu\text{P}$ ar, these HPs transfer optimally to all the sparse models trained for this work. This dense proxy model is a GPT-2 model, but with small changes: ALiBi position embeddings [47] and SwiGLU nonlinearity [51]. We configure it with width: $d_{\text{model}} = d_{\text{model,base}} = 256$, number of layers: $n_{\text{layers}} = 24$, and head size: $d_{\text{head}} = 64$, resulting in 39M parameters. We trained this proxy model on 800M tokens with a batch size of 256 sequences and sequence length 2048 tokens. We randomly sampled 350 configurations of base learning rates, base initialization standard deviation, and embedding and output logits scaling factors. From this sweep we obtained the tuned hyperparameters listed in Table 3.

Table 3: Tuned hyperparameters for our dense proxy model.

Hyperparameter	Value
$\sigma_{W,\text{base}}$	0.08665602
η_{base}	1.62E-2
α_{input}	9.1705
α_{output}	1.0951835

Experimental details for all figures In Table 4, we provide extensive details on hyperparameters, model size, and training schedule for all experiments in this paper. All models in this paper were trained on the SlimPajama dataset [54], a cleaned and deduplicated version of the RedPajama dataset.

Table 4: Experimental details for all figures in this paper.

Figure	d_{model}	L	d_{head}	B	LR	Init. Stdev.	α_{input}	α_{output}	LR decay	LR warm-up steps	Steps	Tokens
Fig. 1, 6, 8, 14	4096	2	64	128	Variable	SP: 2.166E-2 μP , S μP Par: 0.087	9.1705	1.095	10x linear	116	1169	306M
Fig. 7, 13	4096	2	64	8	1.62E-2	Variable	9.1705	1.095	Constant	0	100	1.6M
Fig. 3	2048	10	64	504	SP: 2e-4 μP , S μP Par: 1.62E-2	SP: 0.02 μP , S μP Par: 0.087	9.1705	1.095	Decay to zero	1175	11752	12.13B
Fig. 5	2048	2	32	4	1.68E-02 μP , S μP Par: 1.62E-2	0.101	11.22	1	Constant	0	10	82K
Fig. 9	Variable	2	64	128	Variable	SP: 0.02 μP , S μP Par: 0.087	9.1705	1.095	10x linear	116	1169	306M
Fig. 10	Variable	10	64	256	SP: $d_{\text{model}} \leq 1088$: 6E-4 SP: $d_{\text{model}} > 1088$: 2E-4 μP , S μP Par: 1.62E-2	SP: 0.02 μP , S μP Par: 0.087	9.1705	1.095	10x linear	190	1907	1B
Fig. 11	Variable	2	64	128	Variable	0.087 for SP	N/A	N/A	10x linear	116	1169	306M
Fig. 12	1024	2	64	128	Variable	SP: 2.166E-2 μP , S μP Par: 0.087	9.1705	1.095	10x linear	116	1169	306M

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Sections 6 and A contain a discussion of limitations and broader impact.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: In Section 3 we provide a short proof sketch to provide intuition behind both $\text{S}\mu\text{P}$ and μP . Then in Section D we provide a detailed derivation for $\text{S}\mu\text{P}$ (which also reduces to μP when $\rho = 1$).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: To ensure reproducibility, in Section E we provide extensive details on all experiments contained in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All tests in this paper use GPT-like transformer language models [48, 9], trained on the SlimPajama dataset [54], which are both open. Section E contains sufficient detail to faithfully reproduce the main experimental results. A minimal implementation of SuPar is available at <https://github.com/EleutherAI/nanoGPT-mup/tree/supar>. We also provide a simple breakdown of the code changes required in Table 1.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: To ensure reproducibility, in Section E we provide extensive details on all experiments contained in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: In most of our experiments, we sample multiple seeds and plot the mean loss as a solid line, as well as the standard error of the mean as a shaded area of the same color. We disclose the number of seeds used for each experiment in the corresponding figure captions.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: In Section E we provide extensive experimental details which can be used to easily derive compute requirements based on a reader's specific hardware setting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We do not conduct any research with human participants. Also, we consider the broader impacts and limitations of our work in Sections A and 6. Finally, we provide extensive experimental details in Section E to improve reproducibility.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See Section A.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper presents a new neural network parameterization $\text{S}\mu\text{Par}$ and does not release any models or datasets.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite all datasets and architectures in accordance with their licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We do not introduce new assets in this paper.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not conduct any research involving crowdsourcing or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not conduct any research involving human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.