The Sharpness Disparity Principle in Transformers for Accelerating Language Model Pre-Training

Jinbo Wang* Mingze Wang*† Zhanpeng Zhou* Junchi Yan Weinan E Lei Wu†

* denotes equal contributions, † denotes corresponding author

1. Motivation

Can adaptive layerwise learning rates accelerate Transformer training?

We investigate this question and make the following contributions:

- uncover a clear sharpness disparity across diverse building blocks of Transformers. (not simply layerwise)
- propose Blockwise Learning Rate (LR), a strategy that tailors the LR to each block's sharpness, accelerating LLM pre-training.

2. The Sharpness Disparity Principle

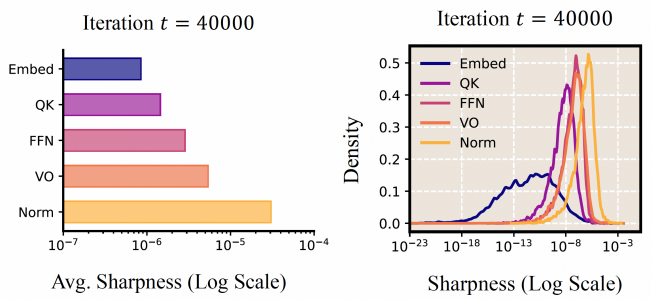


Figure. (left) The average sharpness for the five typical block types in a pre-trained LLaMA model (0.25B) ; (right) the sharpness distribution across different blocks in a pre-trained GPT-2 (small) model.

Transformers exhibit distinctive **alloy-like characteristics**, where **diverse types of blocks** synergistically combine to achieve superior performance. These blocks include: self-attention (further broken down into *query-key (QK)*, *value-output (VO)*), *point-wise feedforward networks (FFN)*, *normalization layers (Norm)*, and *embedding layers (Emb)*.

Key Finding

By measuring the average sharpness for each block type using Hessian approximation, we uncover a clear sharpness disparity across above blocks (see Figure left):

$$\mathcal{S}(\text{Emb}) \ll \mathcal{S}(\text{QK}) < \mathcal{S}(\text{FFN}) < \mathcal{S}(\text{VO}) \ll \mathcal{S}(\text{Norm})$$

This disparity *emerges early* and *persists throughout training*.

3. The Blockwise Learning Rate Strategy

Methodology

We propose a Blockwise Learning Rate (LR) strategy that adjusts the learning rate at the block-type level within the Transformer architecture. Each block type is assigned a different LR multiplier based on its characteristic sharpness:

$$\eta_{\text{Norm}} = \eta_{\text{base}}, \eta_{\text{VO}} = 4\eta_{\text{base}}, \eta_{\text{FFN}} = 6\eta_{\text{base}}, \eta_{\text{QK}} = 8\eta_{\text{base}}, \eta_{\text{Embed}} = 10\eta_{\text{base}}$$

These ratios are guided by sharpness disparity and grounded in the theoretical foundation of the fast-slow dynamics, which suggests increasing the LRs along low-sharpness directions while maintaining stability by preserving the LRs along high-sharpness directions. Each block type is assigned a different LR multiplier based on its sharpness, the specific multipliers above represent a tuned configuration.

4. Experiments

We consistently achieve **lower terminal loss** and nearly **2x speedup** in a wide range of experiments!

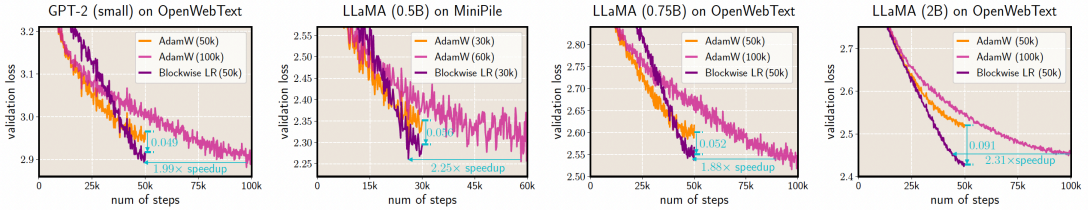
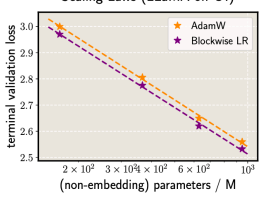


Figure. AdamW with Blockwise LR outperforms AdamW across different model types, model sizes, and datasets.

Scaling law experiment



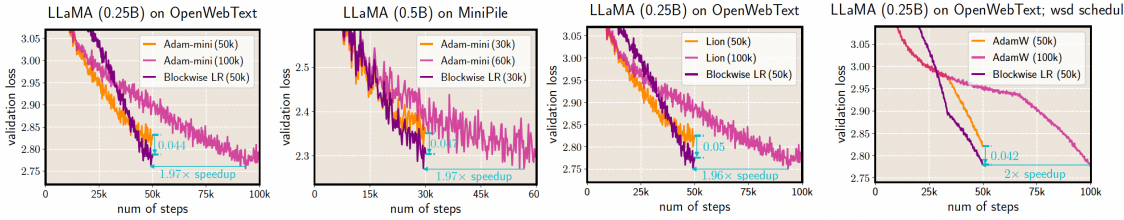


Figure. Results on other optimizers (Adam-mini, Lion) and wsd lr scheduler.

Ablation study

Blockwise LR	Terminal loss
w/o	2.834
Emb	2.818 (-0.016)
Emb & FFN	2.791 (-0.043)
All except *Norm*	2.784 (-0.050)
Norm	2.837 (+0.003)