

Adam-mini: Use Fewer Learning Rates To Gain More

Yushun Zhang^{*12}, Congliang Chen^{*12}, Ziniu Li¹², Tian Ding²,
Chenwei Wu³, Yinyu Ye⁴, Zhi-Quan Luo¹², Ruoyu Sun^{†12}

¹The Chinese University of Hong Kong, Shenzhen, China

²Shenzhen Research Institute of Big Data

³ Duke University

⁴ Stanford University

{yushunzhang, congliangchen, ziniuli}@link.cuhk.edu.cn, dingtian@sribd.cn
 cwwu@cs.duke.edu, yyye@stanford.edu, luozq@cuhk.edu.cn, sunruoyu@cuhk.edu.cn

Abstract

We propose Adam-mini, an optimizer that achieves on-par or better performance than AdamW with 45% to 50% less memory footprint. Adam-mini reduces memory by cutting down the learning rate resources in Adam (i.e., $1/\sqrt{v}$). We find that $\geq 90\%$ of these learning rates in v could be harmlessly removed if we (1) carefully partition the parameters into blocks following our proposed principle on Hessian structure; (2) assign a single but good learning rate to each parameter block. We further find that, for each of these parameter blocks, there exists a single high-quality learning rate that can outperform Adam, provided that sufficient resources are available to search it out. We then provide one cost-effective way to find good learning rates and propose Adam-mini. Empirically, we verify that Adam-mini performs on par or better than AdamW on various language models sized from 125M to 7B for pre-training, supervised fine-tuning, and RLHF. The reduced memory footprint of Adam-mini also alleviates communication overheads among GPUs, thereby increasing throughput. For instance, Adam-mini achieves 49.6% higher throughput than AdamW when pre-training Llama2-7B on $2 \times$ A800-80GB GPUs, which saves 33% wall-clock time for pre-training¹.

1 Introduction

Adam(W) [22, 28] has become the de-facto optimizer for training large language models (LLMs) (e.g., [50, 2, 49, 47]). Despite its superior performance, Adam is expensive to use. Specifically, Adam requires the memory for its optimizer states: the first-order momentum m , and second-order momentum v^2 . These in total take at least $2 \times$ the memory of the model size. This memory consumption has become a major burden in LLM training. For instance, to train a 7B model, Adam alone requires about 56 GB per card for m and v , and with the gradients included, a total of 86 GB is needed. This is expensive even for cutting-edge graphical cards (e.g., A100-80GB). To support the high-memory algorithm, CPU-offload and sharding [41] must be used in practice, which unfortunately increases the latency and slows down the training [42]. The situation further deteriorates when training larger models like PaLM with 540B parameters [6]. In this case, Adam alone occupies more than 50 GPUs and has become a major overhead for pre-training.

It is intriguing to design effective optimizers that require less memory. **First**, the reduction in memory can ease the burden on CPU offloading, and alleviate the need to shard model parameters. All these

^{*}: Equal contribution. [†]: Correspondence author.

¹Our implementation of Adam-mini is available at <https://github.com/zyushun/Adam-mini>

²We restate the update rules of Adam and AdamW in Appendix A.

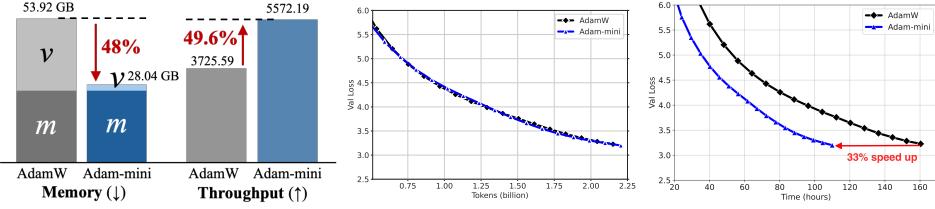


Figure 1: Results for Llama2-7B pre-training. (a) Adam-mini takes less memory and can reach higher throughput (# tokens per second). The throughput is tested on 2× A800-80GB GPUs. (b, c) Adam-mini performs on-par with AdamW, but takes 33% less time for processing the same # tokens.

reduce communication among GPUs and CPUs, which in turn, can enhance throughput and accelerate the training process. **Second**, it allows practitioners to use fewer GPUs to train a model with desired size, leading to substantial saving in both cost and energy. **Third**, it lowers the threshold for training LLMs and encourages participation from more researchers with limited GPU resources.

It is challenging to modify Adam without sacrificing performance. One primary reason is that we still lack understanding for the role of Adam’s m and v [57, 24]. It remains uncertain which components in Adam are indispensable for superior performance, and consequently, which elements could be re-designed and improved. One popular attempt is Adafactor [45], which cuts down memory by low-rank factorization on v . However, Adafactor is widely reported to suffer degenerated performance for training large language models (e.g. [30]). There are two possible reasons. First, the current v in Adam might be crucial for effectiveness and has no room to cut down. This is possible because most existing Adam variants proposed to modify v to varying extents, but they are reported to perform worse than Adam [35]. Second, it is possible to cut down v , but Adafactor did not use the most suitable way: matrix factorization is a generic approach that could be applied broadly, but it does not leverage much problem-specific structure, thus it does not work well on specific neural-net tasks.

In this work, we find it is possible to significantly reduce the usage of v by a simple trick. Currently, Adam’s v assigns an individual learning rate for each parameter, i.e., i -th parameter receives learning rate $\frac{\eta}{\sqrt{v_i}}$, where v_i is the i -th component of v . For a billion-sized model, Adam designs billions of learning rates. We argue that it is possible to achieve on-par or better performance with much fewer learning rate resources. We notice that Transformer’s Hessian has a near-block diagonal structure consisting of different sizes of dense sub-blocks. We find that, for each of these dense sub-blocks, there exists a single high-quality learning rate that outperforms Adam, provided that we have enough resources to search it out. Since the number of dense sub-blocks is much fewer than the number of parameters, our findings imply that it is possible to achieve good performance with much fewer learning rates. The remaining question is how to find them efficiently.

We then propose a cheap and simple way to find good learning rates that are sufficient to perform on-par or better than Adam. We introduce the proposed design principle here: we first partition the gradient vector into B sub-vectors according to the dense Hessian sub-block, and call it g_b for $b = [B]$. For each g_b , we calculate the quantity below.

$$v_b = (1 - \beta_2) * \text{mean}(g_b \circ g_b) + \beta_2 * v_b, \quad b = 1, \dots, B$$

We then use $\eta/\sqrt{v_b}$ as the learning rate for the parameters associated with g_b . We call the corresponding method Adam-mini. We provide a simple illustration in Figure 2 and relegate the complete form later in Algorithm 1. As a result, Adam-mini changes almost all Adam’s v to a handful of scalars and thus significantly saves memory of Adam. We summarize our main contribution as follows.

- **New optimizer.** We propose a new optimizer called Adam-mini. First, Adam-mini partitions the model parameters based on the principle we established upon the Hessian structure. Then,

it chooses a single learning rate for each block using the average of Adam’s v in that block. Adam-mini has the following advantages.

- **Lightweightness:** By design, Adam-mini largely reduces the number of learning rates used in Adam. For mainstream LLMs, Adam-mini could cut down $\geq 90\%$ proportion of Adam’s v , which saves 45% to 50% of the memory cost of Adam.
- **Effectiveness:** Despite the memory cut down, we empirically verify that Adam-mini performs on par or even better than AdamW on various language models sized from 125M to 7B, including pre-training, supervised fine-tuning (SFT), and reinforcement learning from human feedback (RLHF). Adam-mini also reaches better performance on non-LLM tasks such as training diffusion models, vision models and graph neural networks.
- **Efficiency:** Adam-mini can reach higher throughput than AdamW. We observe that Adam-mini reaches 49.6% higher throughput of AdamW when pre-training Llama2-7B on $2 \times$ A800-80GB, which saves 33.1% wall-clock time for pre-training. The efficiency comes from two factors. First, Adam-mini does not introduce extra compute in per-step updates. Second, the memory cut-down allows larger batch sizes per GPU, and at the same time, it eases the burden of communication among GPUs and CPUs, which is usually a major overhead.
- **Partition principle.** A key component in Adam-mini is the strategy for parameter partition. We propose to partition parameters based on the smallest dense sub-block in Hessian. This principle can apply to generic problems: for problems with block diagonal structure, we find that more learning rates do not necessarily bring extra gain. In particular, for the problem associated with each dense sub-block, a single (but good) learning rate suffices to bring better performance.
- **Hessian structure of Transformers.** We empirically apply the above principle to Transformers. We find that Transformer Hessian’s smallest dense blocks are: Query and Key (by heads); Value (as a whole); and MLP (by layer). We use the corresponding partition strategy in Adam-mini and find it is important for good performance. Our findings on Transformers’ Hessian structure may motivate other algorithmic designs as well.

2 Method

2.1 Motivations and Observations

Now we discuss our observations that motivate the design of Adam-mini. We start by investigating the role of Adam’s v and explore possibilities for improvement. In Adam, v provides an individual learning rate for each parameter, i.e., i -th parameter receives the learning rate $\frac{\eta}{\sqrt{v_i}}$, where v_i is the i -th component of v . Recently, Zhang et al. [59] point out that such design is crucial because Transformers need different learning rates to different blocks. They provide two evidence: First, the Hessian of Transformers and various neural nets are near block diagonal (restated in Figure 3). Second, for Transformers, each block has a dramatically different eigenvalue distribution (restated in Appendix A.2). Combining together, Transformer needs different learning rates for each block to handle the eigenvalue heterogeneity. This could be provided by Adam’s v .

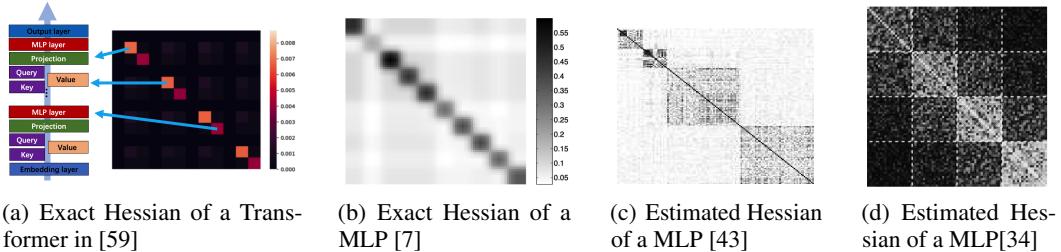


Figure 3: The Hessian of neural nets have near-block-diagonal structure. This is widely reported in the literature on Transformers (a) and various multi-layer perceptrons (MLPs) (b)(c)(d).

The findings in [59] suggest that it is necessary to use a different learning rate for *each block*. Nonetheless, Adam does much more than that: it assigns an individual learning rate not just for each block, but for *each parameter*. Note that the number of parameters (could be billions) is much larger than the number of blocks (usually hundreds). This begs the question:

Is it necessary to use an individual learning rate for each parameter? If not, how much can we save?

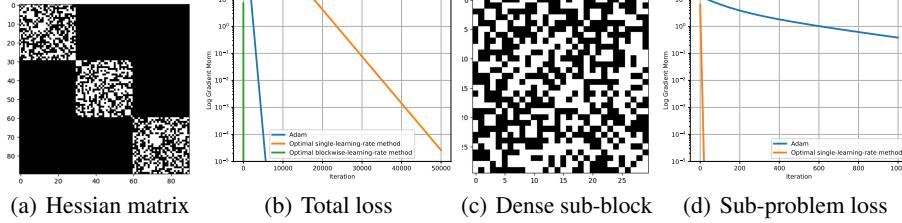


Figure 4: (a): The Hessian of a three-block random quadratic problem. (b): Training curves for the problem associated with the full Hessian in (a). The optimal single (blockwise) learning rate is chosen based on the full (blockwise) Hessian in (a). (c): The 1st dense Hessian sub-blocks in (a). (d): Training curves for the new problem associated with the Hessian in (c).

We first explore this question on generic optimization problems and then discuss Transformers. Following the findings in Figure 3, we conduct a case study on random quadratic minimization problem with block-diagonal Hessian. We visualize the Hessian in Figure 4 (a) and relegate the detailed setup in Appendix B.2. We conduct the coordinate-wise learning-rate method, i.e., Adam, and the single-learning-rate method, i.e., gradient descent (GD). We choose quadratic minimization because the optimal learning rate has a close form. We have the following findings.

- (1): as shown in Figure 4 (a) and (b), Adam outperforms the optimal single-learning-rate method. This is expected since Adam deploys different learning rates to different parameters.
- (2): as shown in Figure 4 (c) and (d), we consider a new problem whose Hessian is a dense sub-block of (a). We consider the optimal single learning-rate method for this new problem and find it outperforms Adam, even though Adam assigns much more learning rates. Similar phenomena apply to all the three sub-blocks of (a).
- (3): If we collect these optimal learning rates in (2) and apply them to a “blockwise” version of GD, it would be faster than Adam on the original problem (the green line in Figure 4 (b)).

In summary, for generic problems with block-diagonal Hessian, many learning rates do not necessarily bring extra gain. In particular, for each dense sub-block, a single (but good) learning rate suffices to bring better performance than using tens or hundreds more. Why would this happen? One possible explanation is the following. Adam can be viewed as a diagonal preconditioned method, but it may not be a good preconditioner and thus cannot effectively reduce the condition number of the dense sub-matrix [13, 55, 46].

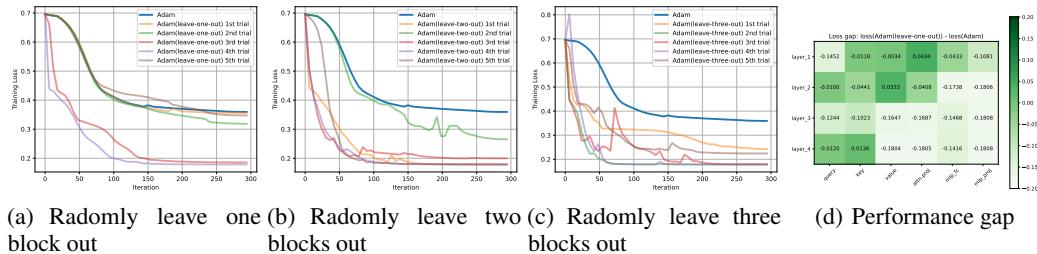


Figure 5: (a) (b) (c) Adam (leave- x -out) can reach a similar or better performance than Adam for all randomly picked left-out blocks. $x = 1, 2, 3$. (d) The performance gap between Adam and Adam (leave-one-out) for all possible blocks. We find Adam (leave-one-out) always performs on par with Adam, and for most blocks, Adam (leave-one-out) performs better.

We also observe the similar phenomena in Transformers. We consider a 4-layer Transformer in the codebase³. Under the default partition by PyTorch, We randomly choose one parameter block as the “left-out” block and change the coordinate-wise learning rate to a single-learning rate counter-part. We use Adam for the rest of the blocks. We grid-search the learning rate for the left-out block and apply the cosine decay schedule. We report the best result and call this method “Adam (leave-one-out)”. Figure 5 shows that Adam (leave-one-out) can reach similar or better performance than Adam for all randomly picked left-out blocks. A similar phenomenon is also observed when we randomly leave

³ <https://colab.research.google.com/drive/1SiF0KZJp75rUeetK0WqpsA8clmHP6jMg?usp=sharing>

out up to three blocks and search three learning rates. We cannot afford to leave more blocks out since the cost of grid search grows exponentially.

To summarize all the phenomena above, we find that: for Transformers, it is possible to reach similar or better performance with much fewer learning rates than Adam. The remaining issue is how to find them without grid-search. Following this finding, in the next part, we propose a simple and cost-effective method called Adam-mini, which could bring comparable or even better performance than Adam, but with 90% to 99% fewer learning rates.

2.2 Proposed Method: Adam-mini

Based on the discussion above, we propose Adam-mini in Algorithm 1. Adam-mini aims to cut down the learning rate resources in Adam without laboriously grid-search the learning rates as in Section 2.1. The design of Adam-mini consists of two steps. **Step 1** is only required at the initialization.

Step 1-1. We partition the model parameters into blocks. For Transformers, we use the strategy of “**Partition** for Transformers”, which partitions all Querys and Keys by heads and uses the default PyTorch partition for the rest. For other networks, we use the default PyTorch partition, and we call it “**Partition** for non-Transformers”. We discuss its design principle later in Section 2.3.

Step 1-2. We choose `embd_blocks`. For Transformers, it includes the embedding layer and output layer. For other networks, no parameters will be chosen.

Step 2. For each parameter block that is outside the `embd_blocks`, we use a single learning rate. To efficiently choose a suitable learning rate in each block, Adam-mini simply replaces $g \circ g$ in vanilla Adam by its mean value. We adopt the moving average on these mean values as in Adam.

A simple example. We use a simple example to illustrate the key design of Adam-mini. For a problem with 5 parameters $w \in \mathbb{R}^5$, Adam and Adam-mini both perform $w = w - u \circ m$, where m is the 1st-order momentum and u has different forms as follows:

- For Adam: $u_{\text{Adam}} = \left(\frac{\eta}{\sqrt{v_1}}, \frac{\eta}{\sqrt{v_2}}, \frac{\eta}{\sqrt{v_3}}, \frac{\eta}{\sqrt{v_4}}, \frac{\eta}{\sqrt{v_5}} \right)$.
- For Adam-mini: suppose the partition is (1, 2, 3) and (4, 5) then

$$u_{\text{mini}} = \left(\frac{\eta}{\sqrt{(v_1 + v_2 + v_3)/3}}, \frac{\eta}{\sqrt{(v_1 + v_2 + v_3)/3}}, \frac{\eta}{\sqrt{(v_1 + v_2 + v_3)/3}}, \frac{\eta}{\sqrt{(v_4 + v_5)/2}}, \frac{\eta}{\sqrt{(v_4 + v_5)/2}} \right).$$

Note that the number of effective elements u_{mini} equals to the number of blocks, which could be significantly smaller than that of u_{Adam} , which equals to the number of parameters. For LLMs, we will show that this would free $\geq 90\%$ elements in v .

Remark on the “embd_blocks”. In Transformers, “`embd_blocks`” refer to the embedding layer and output layer. We explain why **Step 2** does not apply to these blocks. This is because these blocks do not reconcile with the average operation in **Step 2**: By design of the embedding layer, many rows will be zero if the tokens corresponding to these rows do not appear in the current data minibatch. In other words, the only non-zero row is the row that represents the current input word [39]. As such, taking an average over the entire embedding layer will absorb a significant amount of zeros, which leads to highly biased learning rates. A similar argument also applies to the output layer. In Figure 8 (a), we show that removing the `embd_blocks` causes training instability (see the curve with label “Adam-mini (`embd_blocks_removed`)”).

2.3 Principle for the Partition Strategy

We now discuss how to choose the parameter partition for Adam-mini. Based on the analysis in Section 2.1, a general principle is shown in **Principle 1**. Building upon this principle, we now present the corresponding partitioning strategy in the context of neural networks.

Principle 1: We should partition parameters into blocks, s.t., each parameter block is associated with the smallest dense sub-blocks in Hessian.

Based on the block-diagonal structure reported in the literature (Figure 3), the default partition in PyTorch would be a reasonable candidate. In Section 3, we will show that this partition indeed works well widely on non-Transformer tasks such as ResNet, diffusion model, and graph models. We show this strategy in Algorithm 3: “Partition for non-Transformers”.

Algorithm 1 Adam-mini in Pytorch style

```

1: Input weight-decay coefficient  $\lambda$  and
   current step  $t$ 
2: Choose param_blocks from
   Algorithm 2 or 3
3: Choose embd_blocks from
   Get_embd_blocks
4: for name, param in param_blocks do
5:   g = param.grad
6:   param = param -  $\eta_t * \lambda * \text{param}$ 
7:   m =  $(1 - \beta_1) * g + \beta_1 * m$ 
8:    $\hat{m} = \frac{m}{1 - \beta_1^t}$ 
9:   if param in embd_blocks then
10:    v =  $(1 - \beta_2) * g \circ g + \beta_2 * v$ 
11:   else
12:    v =  $(1 - \beta_2) * \text{mean}(g \circ g) + \beta_2 * v$ 
13:   end if
14:    $\hat{v} = \frac{v}{1 - \beta_2^t}$ 
15:   param = param -  $\eta_t * \frac{\hat{m}}{\sqrt{\hat{v} + \epsilon}}$ 
16: end for

```

Unfortunately, we find that default PyTorch partition does not always work well on Transformers. In particular, we find that Adam-mini encounters training instability on 1B models (see Figure 6 (d)). We suspect this is because the default PyTorch partition did not fully capture the Hessian structure. To find out, we explore the Hessian the 4-layer Transformer as in Figure 5. Under the default PyTorch partition, we compute the exact Hessian at initialization for each parameter block (or equivalently, the principal sub-blocks in the whole Hessian). We report some representatives in Figure 6 and relegate the others in Appendix A. We find there are two classes of Hessian sub-blocks.

- **Class 1:** Similar to the whole Hessian, the Hessian sub-block itself further has a block-diagonal structure and consists of smaller dense matrices. This class includes Query and Key. We empirically find that the number of small dense sub-blocks equals the number of heads in multi-head attention.
- **Class 2:** The Hessian sub-blocks have dense structures and cannot be further partitioned into smaller ones. This class includes Value, attention projection, and MLP layers. Note that the Hessian structure of Value is different from that of Query and Key, although they all consist of four heads. This is probably because Value is positioned outside the softmax operator in the self-attention design, while Query and Key are not.

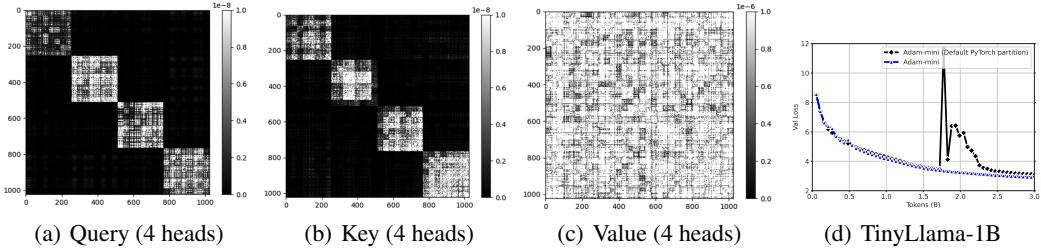


Figure 6: (a,b,c): The Hessian of different parameter blocks in a 4-layer Transformer at initialization. More figures in other blocks and detailed setup are shown in Appendix A.3 and B.2. (d) Training curves on TinyLlama-1B. When using the PyTorch default partition, Adam-mini could suffer loss spikes. The spike disappears when we change the partition strategy to “**Partition for Transformers**”.

Based on above findings, we find the default Pytorch partition is indeed not the best fit for Transformers: Querys and Keys can be further partitioned into different blocks by head. By **Principle 1**, different heads should belong to different blocks. Intuitively, different heads play a different role in understanding tokens, thus it seems reasonable that they require distinct learning rates.

We then introduce the resulting Algorithm 2: “Partition for Transformers”, which partitions Querys and Keys by heads. This procedure follows the principle based on the smallest Hessian sub-blocks. As shown in Figure 6 (d). This strategy indeed stabilizes the training and boosts the performance.

2.4 Some Characteristics of Adam-mini

Memory cut down. By design, for Transformers, Adam-mini reduces the number of learning rates from the number of total paramters to the sum of the size of embedding layer, the size of output layer, and the number of blocks in other layers according to our procedure to partition Transformers. Thus, the proportion of reduced memory depends on the fraction of non-embedding parameters in the model. On Llama2-7B, this fraction is 96.2%. For mainstreamed LLMs, this fraction is usually $\geq 90\%$ (see a summary in Table 5 in Appendix A). Note that this fraction monotonically approaches 100% as more and more layers are stacked into the architecture, e.g., for Llama3-70B, this proportion is 99.25%. As such, Adam-mini could cut down more than 90% of v , which saves up to 45% to 50% of Adam’s memory.

Table 1: Memory cost of AdamW v.s. Adam-mini on Llama2-7B. Calculation is based on float32.

Model	Optimizer	Memory (GB)
GPT2-1.5B	AdamW	12.48
GPT2-1.5B	Adam-mini	6.56 (47.4% ↓)
TinyLlama-1B	AdamW	8.80
TinyLlama-1B	Adam-mini	4.92 (44.10% ↓)
Llama2-7B	AdamW	53.92
Llama2-7B	Adam-mini	28.04 (48.04% ↓)

Table 2: Throughput (\uparrow) test on $2 \times$ A800-80GB GPUs for Llama2-7B pre-training. \times means out of memory.

Optimizer	Micro_bs	Total_bs	Throughput (\uparrow)
Adam-mini	4	256	5572.19 ($\uparrow 49.6\%$)
AdamW	2	256	\times
AdamW	1	256	3725.59

Higher throughput. Adam-mini can reach a higher throughput than AdamW, especially under limited hardware resources. There are two reasons. First, Adam-mini does not introduce extra computation in its update rules. The average operation incurs negligible cost. Additionally, it significantly reduces the number of tensor-square-root and tensor-division operations in AdamW. Second, thanks to the memory cut-down, Adam-mini can support larger batch sizes per GPU, and at the same time, Adam-mini reduces the communication among GPUs, which is known to be a major overhead [42].

Owing to these properties, Adam-mini could reduce the overall time for pre-training. We provide evidence in Table 2. Here, Micro_bs and Total_bs denote per-device and total batch size. When pre-training Llama2-7B on $2 \times$ A800-80GB GPUs, we find Adam-mini could reach 49.6% higher throughput than AdamW. Recall throughput (\uparrow) refers to # processed tokens per second, so Adam-mini could save 33.1% time on processing the same amount of tokens for pre-training.

Has room to improve. Adam-mini designs the learning rate for each dense Hessian sub-block using the average of Adam’s v in that block. Such a design achieves cheap computation, but it might not be optimal. Fortunately, the current design is sufficient to achieve the same or marginally better performance than Adam with much less memory. We believe there is great room to improve the learning rate design, yet it requires more fine-grained analysis of each dense Hessian sub-block. We leave it as an important future direction.

3 Experiments

We now verify the efficacy of Adam-mini. We primarily focus on LLM tasks, including pre-training, supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). We will also evaluate Adam-mini on non-LLM tasks. All LLM experiments are conducted on four NVIDIA A800-80GB GPUs and the rest are conducted on four V100 GPUs. All the experimental details are explained in Appendix B.1.

3.1 Pre-training

Setups. We pre-train open-sourced LLMs including GPT2 series and Llama series. We train these models on mainstream English Corpus from scratch. In particular, We train GPT2 [40] series

on Openwebtext [16]. We train TinyLlama-1B, Llama2-7B [49] on CommonCrawl⁴. We train models with 2B, 3B, and 25B tokens (as recommended by Chinchilla’s law [19]). We choose model configurations (e.g. context length) by their standard protocols. We compare Adam-mini with AdamW as well as popular memory-efficient methods including Adafactor [45], CAME [30], and SM3[4]. For Adafactor and SM3, we incorporate momentum with $\beta_1 = 0.9$ to ensure a fair comparison with other methods. We tune the learning rate for all methods within the same budget and report the best performance.

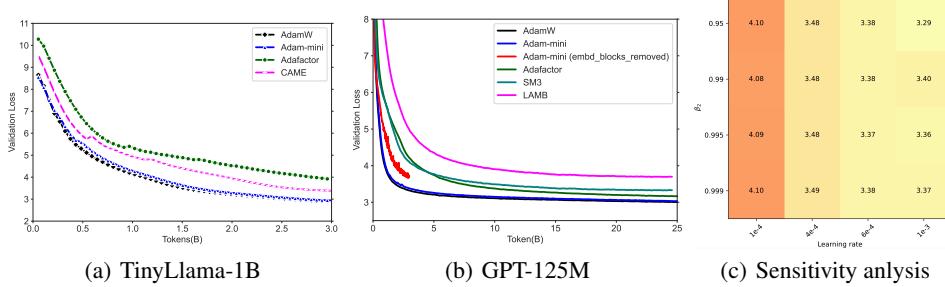


Figure 7: Training curves of (a) TinyLlama-1B. (b) GPT2-125M. Adam-mini performs on par as AdamW with less memory, while other methods perform worse on these tasks. (c): Adam-mini seems not sensitive to hyperparameters.

Llama series. Figure 7 (a) shows the validation loss curve for pre-training TinyLlama-1B. The training curve for Llama2-7B is presented in Figure 1 (c) in Section 1. For both TinyLlama-1B and Llama2-7B, we find Adam-mini performs on par with AdamW with less memory.

GPT2 series. Figure 7 (b) shows the validation loss curve for GPT2-125M and Figure 8 shows the validation loss curve for GPT2 sized from 330M to 1.5B. We find that Adam-mini matches the good performance of AdamW with less memory, while other methods perform worse. In Figure 7 (b), we further consider Adam-mini (all-packed), which sets the `embd_blocks` = \emptyset . That is, we use one single learning rate for the whole embedding (output) layer. We find that Adam-mini (all-packed) performs poorly, as expected from the analysis in Section 2.1. We stop the trial since it shows clear unstable behavior.

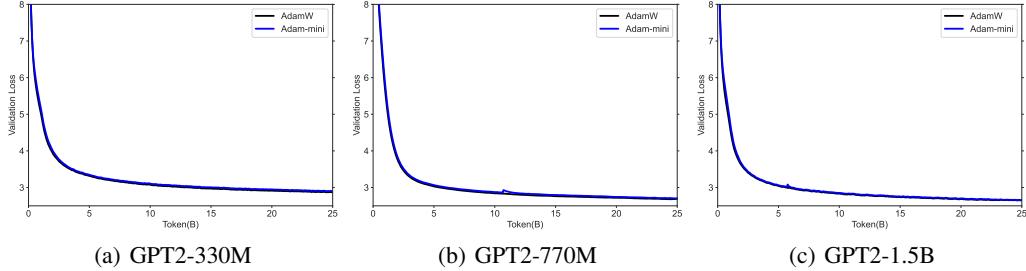


Figure 8: Training curves of GPT2 pre-training. Adam-mini reaches the same performance as AdamW with less memory.

Sensitivity analysis. On GPT2-125M pre-training task, we test the sensitivity of Adam-mini to hyperparameters. We report the validation loss after training with 2.5B tokens. As shown in Figure 7, Adam-mini seems not sensitive to hyperparameters.

3.2 Supervised Fine-tuning and RLHF

In this section, we evaluate the effectiveness of Adam-mini for downstream fine-tuning tasks. Specifically, we consider two representative tasks: supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF). We use the Llama-2-7b pretrained model (released by Meta [49]) for our study. We use the ultrafeedback dataset⁵ and implement the RLHF workflow from [36]. Specifically, we train an SFT model with 40% of the chosen data and train a reward model using the remaining 60%. Then, we apply the reinforcement learning algorithm ReMax [26], a memory-efficient alternative to PPO [44], to optimize the preference reward.

⁴<https://commoncrawl.org/>.

⁵<https://huggingface.co/datasets/argilla/ultrafeedback-binarized-preferences-cleaned>

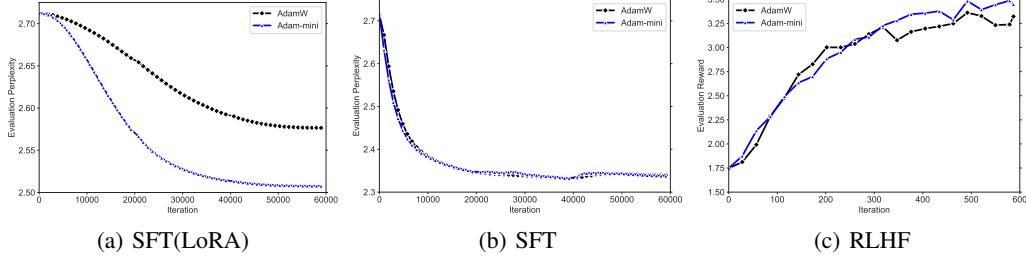


Figure 9: Training curves of SFT (LoRA), SFT and RLHF when aligning Llama2-7B. Adam-mini reaches better performance (smaller perplexity, higher reward) than AdamW with less memory.

Table 3: GPT-4 evaluation score (\uparrow) of SFT and RLHF on the MT-Bench [61].

	SFT (LoRA)		SFT		RLHF	
	AdamW	Adam-mini	AdamW	Adam-mini	AdamW	Adam-mini
MT-Bench	4.23	4.41	5.37	5.40	5.54	5.68

We evaluate the alignment performance in terms of chat ability using the MT-Bench [61], where GPT-4 assesses multi-turn chatting capabilities and assigns a score from 0 to 10 (higher is better). Our results, presented in Table 3, demonstrate that Adam-mini can outperform AdamW, even though Adam-mini uses a single learning rate and is more memory-efficient.

3.3 Non-LLM Tasks

We now evaluate Adam-mini on non-LLM tasks. Table 4 shows the results for training ResNet18 on ImagNet; Diffusion model on CelebA; Graph Convolution Net (GCN), Graph Attention Net (GAT) on OGB-arxiv. We find that Adam-mini reaches comparable or better performance than AdamW, but with less memory.

Table 4: On popular non-LLM tasks, Adam-mini performs on par or better than AdamW.

Domain	Model	Optimizer	Metric	25% steps	50% steps	75% steps	100% steps
Vision	ResNet18	AdamW	Test acc (\uparrow)	0.6149	0.6478	0.6613	0.6669
Vision	ResNet18	Adam-mini	Test acc (\uparrow)	0.6140	0.6501	0.6629	0.6667
Vision	Diffusion model	AdamW	Train loss (\downarrow)	0.0529	0.0497	0.0420	0.0394
Vision	Diffusion model	Adam-mini	Train loss (\downarrow)	0.0525	0.0495	0.0416	0.0388
Graph	GAT	AdamW	Val acc (\uparrow)	0.7277	0.7367	0.7399	0.7421
Graph	GAT	Adam-mini	Val acc (\uparrow)	0.7378	0.7394	0.7403	0.7429
Graph	GCN	AdamW	Val acc (\uparrow)	0.7347	0.7428	0.7379	0.7374
Graph	GCN	Adam-mini	Val acc (\uparrow)	0.7406	0.7427	0.7380	0.7423

4 Related works

Understanding of Adam. There is an active line of works trying to understand why Adam works well [56, 53, 58, 52, 37, 21, 24, 59, 3]. In contrast to these works, we point out that Adam’s v might not function at its full potential as effectively as we expected: sometimes fewer learning rates can reach the same or better results (due to the dense Hessian sub-blocks). Our findings might motivate stronger optimizers that better fit the neural-net Hessian structure.

Lightweight optimizers for general tasks. There are several attempts to reduce the memory cost of Adam. Adafactor [45] and its variant CAME [30] conduct nonnegative low-rank factorization over Adam’s v . SM3 [4] is a lightweight version of Adagrad [12]. SM3 chooses the learning rate of i -th parameter by taking the minimal value in a certain candidate set, and each element in the candidate set is related to the maximal squared gradient under a predetermined cover. All these aforementioned methods could release almost all memory for v and save about 48% of Adam’s memory. However, we find that their performance degenerate in various experiments, while Adam-mini maintains as effective as AdamW (Figure 7).

After completing this work, we noticed two unpublished arXiv papers with similar algorithmic designs to us: BAGM [62] and NovoGrad [15]. Both of them use block-wise or layer-wise adaptive learning rates to achieve robust performance and better generalization. We summarize their key differences with Adam-mini. BAGM designs partitions to reach minimal-norm solutions and achieve provable robustness. They then partition parameters based on their theory: (I) they used the default

PyTorch partition; (II) they did not introduce `embd_blocks`. Although these partition designs may have benefits on robustness, they overlook the Hessian structure and oversimplify the training problem (as we discussed in Section 2.2 and 2.3). As a result, these designs will lead to training instability on large-scale experiments. This is evident in our failed preliminary versions of Adam-mini in Figure 6 (d) and 7 (b). As for NovoGrad, it uses layer-wise learning rate design, which is coarser than the blockwise design in BAGM and thus would share similar drawbacks to BAGM. Further, NovoGrad introduces a different design to 1st-order momentum: instead of performing weighted-sum on the past gradients, it performs weighted-sum on “the current gradient divided by the 2nd-order momentum”. Such design is largely different from AdamW and Adam-mini. It seems unclear whether such design can work on large-scale tasks like LLMs.

In summary, these two methods have different designs and their partition strategies oversimplify the training problems. Consequently, they would cause training instability on large-scale experiments as evident in Figure 6 (d) and 7 (b). In contrast, Adam-mini carefully assigns learning rates following our proposed principle on Hessian structures. Such design principle is crucial for training stability and it works well on various LLMs including 7B models.

Lightweight methods for SFT. To fine-tune LLMs under limited resources, researchers proposed to use SGD [32] or its zeroth-order version [33]. Yet, these methods converge more slowly than AdamW on SFT tasks (e.g., see evidence in [14, 29]). Further, it seems unclear if these methods could be used for pre-training since SGD is widely considered incompetent for pre-training tasks [57].

Other orthogonal methods. A very recent work GaLore [60] conducts low-rank approximation on the gradient and then calculates m and v based on this low-rank gradient estimation. Adam-mini can be combined orthogonally with GaLore to achieve more memory reduction. LoRA [20] is a memory-efficient method for SFT tasks. This method fine-tunes the model via additive low-rank adaptors and uses Adam to update these adaptors. Note that the Adam steps in LoRA can be replaced by Adam-mini. As a result, Adam-mini brings better performance (Figure 9). In parallel to our work, BAdam [29] conducts SFT in a block-coordinate-descent (BCD) fashion. This method requires repeated Adam steps to solve the sub-problem in BCD. Similarly as in LoRA, the Adam steps in BAdam can be replaced by Adam-mini to further reduce memory. The design principle of Adam-mini can also be applied to other diagonal preconditioned methods such as Sophia [27].

There are several other tricks that ease GPU memory burden but are orthogonal to optimizer design. These tricks include gradient checkpointing [5], model offload and sharding [41, 42], quantization [11, 25], and fused update [31, 32]. Adam-mini can be implemented upon these tricks.

Finally, we discuss another popular adaptive optimizer called LAMB [54] (see Algorithm 6 in Appendix A.1). LAMB can be easily mis-understood as a similar optimizer to Adam-mini, but actually it is not. We emphasize that Adam-mini is *significantly different* from LAMB. First, LAMB uses the same coordinate-wise learning-rate design $1/\sqrt{v}$ as in Adam. Second, in addition to this $1/\sqrt{v}$, LAMB further re-scales the parameters in a layer-by-layer fashion. This re-scaling design is often known as the “layer-wise learning rates”, but to be precise, it is actually an additional “layer-wise scaling” over the “coordinate-wise learning rates $1/\sqrt{v}$ ”. As a result, LAMB does not save memory over Adam and its overall design is quite different from Adam-mini. Numerically, we find that LAMB performs worse than Adam-mini on GPT2 pre-training (Figure 7 (b)).

5 Concluding Remarks

We proposed Adam-mini, an optimizer that saves 45% to 50% memory of Adam. We remark that there is great room to improve the design of Adam-mini: currently Adam-mini uses a simple and cost-effective way to design a learning rate for each dense Hessian sub-block, but it might not be an optimal way. We leave the stronger designs as a future direction.

References

- [1] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

- [3] K. Ahn, Z. Zhang, Y. Kook, and Y. Dai. Understanding adam optimizer via online learning of updates: Adam is ftrl in disguise. In *Forty-first International Conference on Machine Learning*.
- [4] R. Anil, V. Gupta, T. Koren, and Y. Singer. Memory efficient adaptive optimization. *Advances in Neural Information Processing Systems*, 32, 2019.
- [5] T. Chen, B. Xu, C. Zhang, and C. Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.
- [6] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- [7] R. Collobert. Large scale machine learning. Technical report, Université de Paris VI, 2004.
- [8] G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.
- [9] A. B. Da Silva and M. Gazeau. A general system of differential equations to model first-order adaptive algorithms. *The Journal of Machine Learning Research*, 21(1):5072–5113, 2020.
- [10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [11] T. Dettmers, M. Lewis, S. Shleifer, and L. Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2021.
- [12] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.
- [13] G. E. Forsythe and E. G. Straus. On best conditioned matrices. *Proceedings of the American Mathematical Society*, 6(3):340–345, 1955.
- [14] T. Gautam, Y. Park, H. Zhou, P. Raman, and W. Ha. Variance-reduced zeroth-order methods for fine-tuning language models. *arXiv preprint arXiv:2404.08080*, 2024.
- [15] B. Ginsburg, P. Castonguay, O. Hrinchuk, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, H. Nguyen, Y. Zhang, and J. M. Cohen. Training deep networks with stochastic gradient normalized by layerwise adaptive second moments. 2019.
- [16] A. Gokaslan, V. Cohen, E. Pavlick, and S. Tellex. Openwebtext corpus, 2019.
- [17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [18] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [19] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- [20] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [21] K. Jiang, D. Malik, and Y. Li. How does adaptive optimization impact local neural network geometry? *Advances in Neural Information Processing Systems*, 36, 2023.
- [22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [23] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2016.

- [24] F. Kunstner, J. Chen, J. W. Lavington, and M. Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. *arXiv preprint arXiv:2304.13960*, 2023.
- [25] B. Li, J. Chen, and J. Zhu. Memory efficient optimizers with 4-bit states. *Advances in Neural Information Processing Systems*, 36, 2024.
- [26] Z. Li, T. Xu, Y. Zhang, Y. Yu, R. Sun, and Z.-Q. Luo. Remax: A simple, effective, and efficient method for aligning large language models. *arXiv preprint arXiv:2310.10505*, 2023.
- [27] H. Liu, Z. Li, D. Hall, P. Liang, and T. Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023.
- [28] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.
- [29] Q. Luo, H. Yu, and X. Li. Badam: A memory efficient full parameter training method for large language models. *arXiv preprint arXiv:2404.02827*, 2024.
- [30] Y. Luo, X. Ren, Z. Zheng, Z. Jiang, X. Jiang, and Y. You. Came: Confidence-guided adaptive memory efficient optimization. *arXiv preprint arXiv:2307.02047*, 2023.
- [31] K. Lv, H. Yan, Q. Guo, H. Lv, and X. Qiu. Adalomo: Low-memory optimization with adaptive learning rate. *arXiv preprint arXiv:2310.10195*, 2023.
- [32] K. Lv, Y. Yang, T. Liu, Q. Gao, Q. Guo, and X. Qiu. Full parameter fine-tuning for large language models with limited resources. *arXiv preprint arXiv:2306.09782*, 2023.
- [33] S. Malladi, T. Gao, E. Nichani, A. Damian, J. D. Lee, D. Chen, and S. Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.
- [34] J. Martens and R. Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.
- [35] F. Orabona. Neural networks (maybe) evolved to make adam the best optimizer. 2020. URL <https://parameterfree.com/2020/12/06/neural-network-maybe-evolved-to-make-adam-the-best-optimizer/>.
- [36] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [37] Y. Pan and Y. Li. Toward understanding why adam converges faster than sgd for transformers. *arXiv preprint arXiv:2306.00204*, 2023.
- [38] B. A. Pearlmutter. Fast exact multiplication by the hessian. *Neural computation*, 6(1):147–160, 1994.
- [39] O. Press and L. Wolf. Using the output embedding to improve language models. *arXiv preprint arXiv:1608.05859*, 2016.
- [40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [41] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.
- [42] S. Rajbhandari, O. Ruwase, J. Rasley, S. Smith, and Y. He. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–14, 2021.
- [43] N. Roux, P.-A. Manzagol, and Y. Bengio. Topmoumoute online natural gradient algorithm. *Advances in neural information processing systems*, 20, 2007.

- [44] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [45] N. Shazeer and M. Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018.
- [46] R. Sun and Y. Ye. Worst-case complexity of cyclic coordinate descent: $O(n^2 \log n)$ gap with randomized version. *Mathematical Programming*, 185:487–520, 2021.
- [47] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [48] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.
- [49] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [51] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.
- [52] B. Wang, Y. Zhang, H. Zhang, Q. Meng, Z.-M. Ma, T.-Y. Liu, and W. Chen. Provable adaptivity in adam. *arXiv preprint arXiv:2208.09900*, 2022.
- [53] Y. Wu, X. Zhu, C. Wu, A. Wang, and R. Ge. Dissecting hessian: Understanding common structure of hessian in neural networks. *arXiv preprint arXiv:2010.04261*, 2020.
- [54] Y. You, J. Li, S. Reddi, J. Hseu, S. Kumar, S. Bhojanapalli, X. Song, J. Demmel, K. Keutzer, and C.-J. Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.
- [55] D. Young. Iterative methods for solving partial difference equations of elliptic type. *Transactions of the American Mathematical Society*, 76(1):92–111, 1954.
- [56] J. Zhang, T. He, S. Sra, and A. Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.
- [57] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33: 15383–15393, 2020.
- [58] Y. Zhang, C. Chen, N. Shi, R. Sun, and Z.-Q. Luo. Adam can converge without any modification on update rules. *Advances in Neural Information Processing Systems*, 35:28386–28399, 2022.
- [59] Y. Zhang, C. Chen, T. Ding, Z. Li, R. Sun, and Z.-Q. Luo. Why transformers need adam: A hessian perspective. *arXiv preprint arXiv:2402.16788*, 2024.
- [60] J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.
- [61] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.
- [62] S. Zheng and J. T. Kwok. Blockwise adaptivity: Faster training and better generalization in deep learning. *arXiv preprint arXiv:1905.09899*, 2019.

Broader Impacts

We propose a new method for training AI models. Our method can help save energy for large AI model training. However, it would be a potential threat if the AI models are used for illegal usage.

A More Results

A.1 Preliminaries on Adam and AdamW

We introduce Adam and AdamW in the Pytorch style in Algorithm 4 and 5. These methods need to track m and v along the training. Both m and v are vectors of the same size as # model parameter.

Algorithm 4 Adam in Pytorch style

```

1: Let  $\lambda$  be the weight decay coefficient
2: for param in parameter_blocks do
3:   g = param.grad
4:   if  $\lambda > 0$  then
5:     g = g +  $\lambda * \text{param}$ 
6:   end if
7:   param = param -  $\eta_t * \lambda * g$ 
8:   m =  $(1 - \beta_1) * g + \beta_1 * m$ 
9:    $\hat{m} = \frac{m}{1 - \beta_1^t}$ 
10:  v =  $(1 - \beta_2) * g \odot g + \beta_2 * v$ 
11:   $\hat{v} = \frac{v}{1 - \beta_2^t}$ 
12:  param = param -  $\eta_t * \frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon}$ 
13: end for
```

Algorithm 5 AdamW in Pytorch style

```

1: Let  $\lambda$  be the weight decay coefficient
2: for param in parameter_blocks do
3:   g = param.grad
4:   param = param -  $\eta_t * \lambda * g$ 
5:   m =  $(1 - \beta_1) * g + \beta_1 * m$ 
6:    $\hat{m} = \frac{m}{1 - \beta_1^t}$ 
7:   v =  $(1 - \beta_2) * g \odot g + \beta_2 * v$ 
8:    $\hat{v} = \frac{v}{1 - \beta_2^t}$ 
9:   param = param -  $\eta_t * \frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon}$ 
10: end for
```

Algorithm 6 LAMB in Pytorch style

```

1: Let  $\lambda$  be the weight decay coefficient, let  $\phi$  be a scaling function.
2: for param in all_layers do
3:   g = param.grad
4:   param = param -  $\eta_t * \lambda * g$ 
5:   m =  $(1 - \beta_1) * g + \beta_1 * m$ 
6:    $\hat{m} = \frac{m}{1 - \beta_1^t}$ 
7:   v =  $(1 - \beta_2) * g \odot g + \beta_2 * v$ 
8:    $\hat{v} = \frac{v}{1 - \beta_2^t}$ 
9:   r =  $\frac{\hat{m}}{\sqrt{\hat{v}} + \epsilon}$ 
10:  param = param -  $\eta_t * \frac{\phi(\|\text{param}\|)}{\|r + \lambda * \text{param}\|} * r$ 
11: end for
```

A.2 Preliminary results in [59]

We here restate [59, Figure 3]. This figure shows that: for Transformers, different parameter blocks have different Hessian eigenvalue distributions, while for CNNs, the eigenvalue distributions are similar among blocks. This suggests that Transformers need different learning rates for different blocks to handle the heterogeneity in eigenvalue distributions.

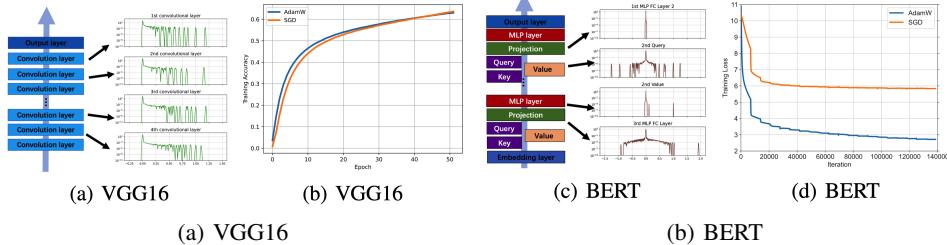


Figure 10: Figure 3 in [59]. The eigenvalues distribution are similar among blocks for CNNs, while they differ significantly across blocks for Transformers. This indicates Transformers need different learning rates for different blocks to handle the heterogeneity in eigenvalue distributions.

A.3 More Hessian sub-blocks in Figure 6

In Figure 6, we present the Hessian of Query, Key, Value and MLP_{proj} in the 2nd layer. We now comprehensively provide Hessian in all attention and MLP blocks in all layers in the 4-layer Transformer used in Figure 6. We plot the Hessian at initialization. Experimental details can be seen in Appendix B.2. As we can see below, almost all these Hessian sub-blocks belong to either **Class 1** or **Class 2**. The only exception is the MLP_{fc} block in the 4th layer, which shows a diagonal pattern.

This begs the question: does Adam significantly outperforms single-learning-rate methods on this diagonal block? We find the answer is still no. This is because the Hessian eigenvalues in this block is concentrated. We provide evidence in Figure 15. This suggests that the optimization problem in this block is well-conditioned and single-learning-rate methods can work well. As shown in Figure 15, optimal single-learning-rate still significantly outperforms Adam, similarly as what happens in other dense blocks. Another evidence lies in Figure 5 (d): when the MLP_{fc} block in the 4th layer is chosen as the left-out block, Adam (leave-one-out) achieves 0.1416 lower loss than Adam.

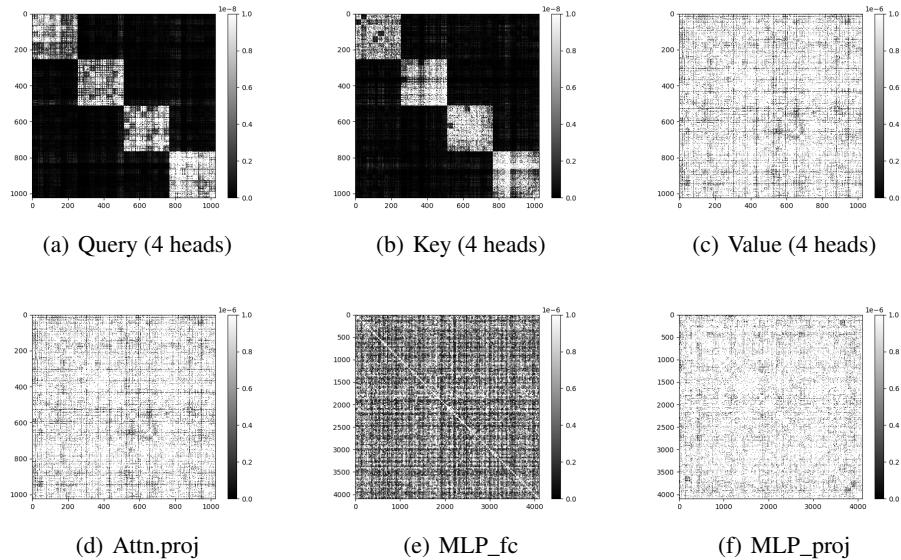


Figure 11: The initial Hessian structure of different parameter blocks in the 1st layer.

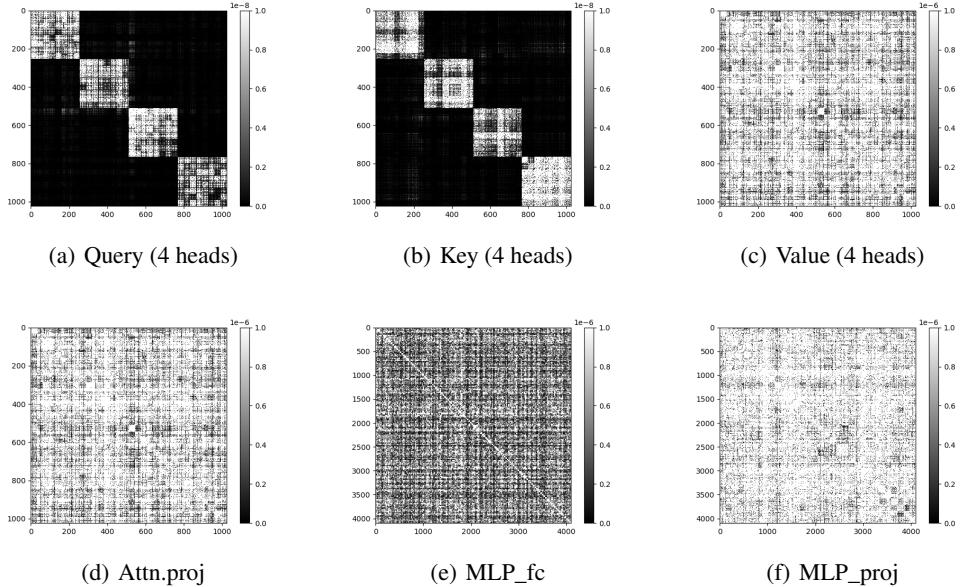


Figure 12: The initial Hessian structure of different parameter blocks in the 2nd layer.

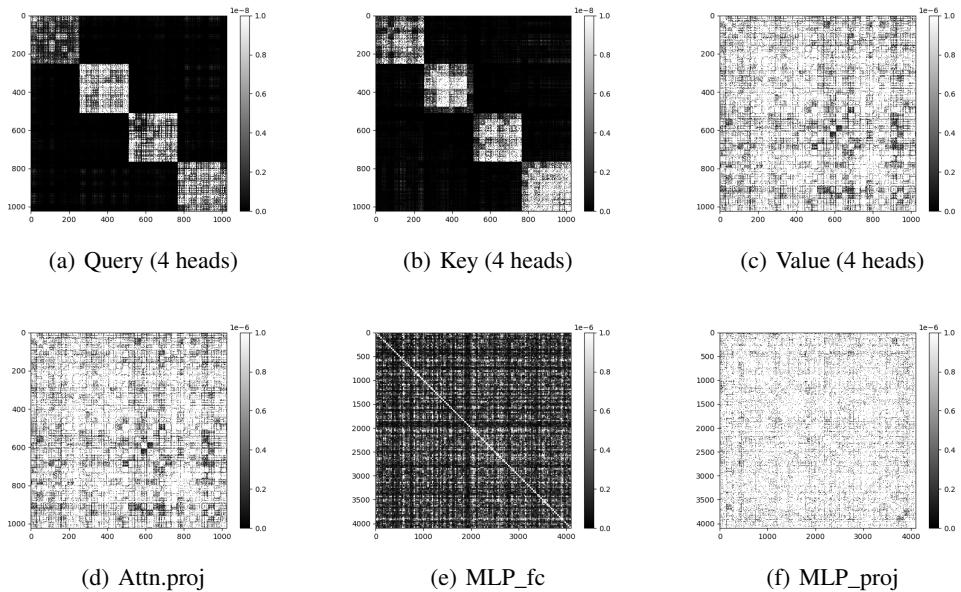


Figure 13: The initial Hessian structure of different parameter blocks in the 3rd layer.

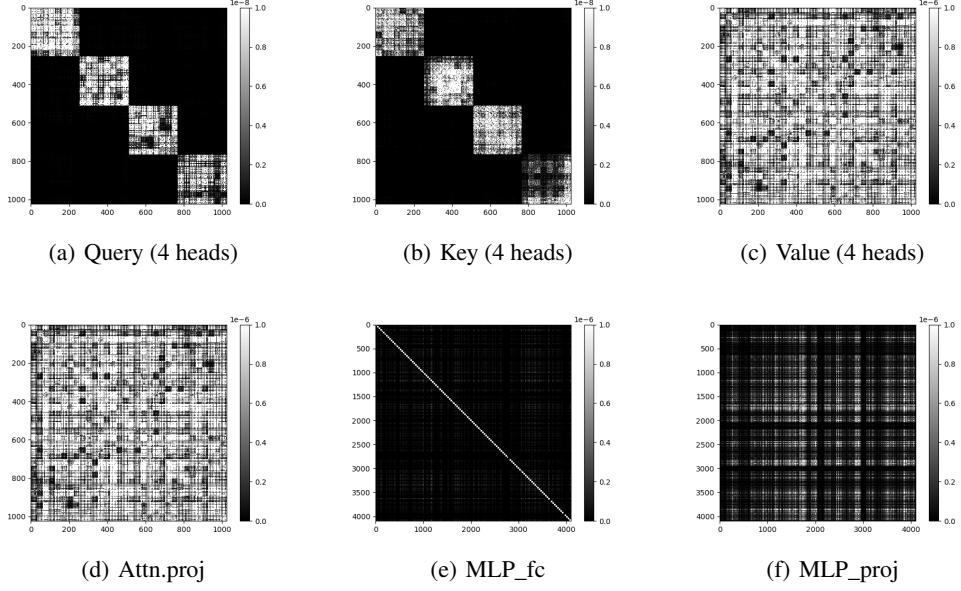


Figure 14: The initial Hessian structure of different parameter blocks in the 4th layer.

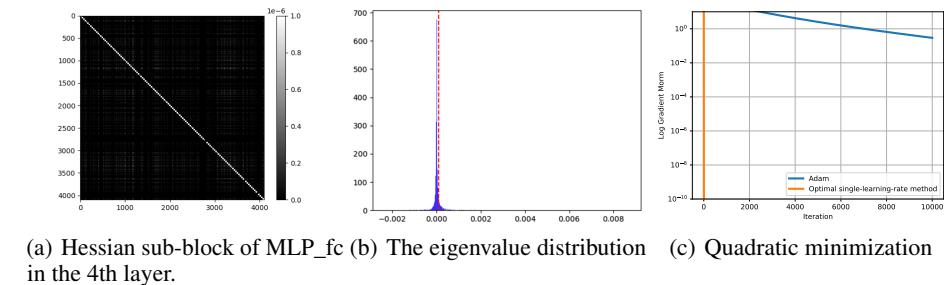


Figure 15: (a) The Hessian sub-block of MLP_fc in the 4th layer. There is a near-diagonal pattern. (b) The eigenvalues are concentrated in this Hessian sub-block. (c) When taking this matrix as the Hessian of a quadratic function, the optimal single-learning-rate method outperforms Adam.

A.4 Embedding Proportion in LLMs

Table 5 shows the proportion of the embedding & output parameters in mainstream LLMs including GPT2 [40], Llama series [49], Phi-3 [1], and Gemma [48]. We find that the embedding & output layer takes a fairly small proportion of the total parameters, usually $\leq 10\%$.

Table 5: Proportion of the embedding & output parameters in mainstream LLMs. Note that GPT2 and Gemma use weight-tying, so the parameters for the embedding & output layers are shared. We find that the embedding & output layer usually takes $\leq 10\%$ of total parameters.

Model	Vocabulary size	Embedding dimension	# Parameters for the embedding & output layer	# Total parameters	Proportion
GPT2-1.5B	50304	1600	50304 * 1600	1.5B	5.17%
TinyLlama-1B	32000	2048	32000*2048*2	1.1B	11.80%
Llama2-7B	32000	4096	32000*4096*2	6.7B	3.80%
Llama2-13B	32000	4096	32000*4096*2	13B	2.02%
Llama2-33B	32000	4096	32000*4096*2	33B	0.79%
Llama2-65B	32000	4096	32000*4096*2	65B	0.40%
Llama3-8B	128256	4096	128256*4096*2	8B	13%
Llama3-70B	128256	4096	128256*4096*2	70B	0.15%
Phi-3	32064	3072	32064*3072*2	3.8B	5.18%
Gemma-7B	256128	3072	256128*3072	8.5B	9.83%

A.5 Further Comparison with Popular Memory-Efficient Optimizers

Similar trajectory to AdamW. As a bonus finding, Adam-mini could generate similar trajectories to that of AdamW, while other popular memory-efficient methods cannot. We train a 6-layer Transformer sized 11M on Openwebtext and launch AdamW, Adam-mini, and other memory-efficient optimizers under the same random seed and same learning rate 1e-5. We save the model weights for every 250 iterations and compare their Euclidean distance to the weights along AdamW’s trajectory. As shown in Figure 16, the model weights obtained by Adam-mini is close to those by AdamW throughout training. In contrast, other optimizers deviate from AdamW. This is because Adam-mini makes fewer modifications over AdamW than other methods.

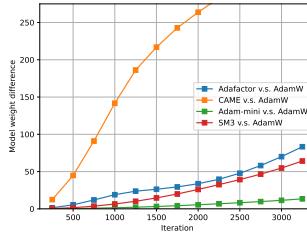


Figure 16: On a 6-layer Transformer, we compare the model checkpoints along the trajectory of different methods. We find the model weights obtained by Adam-mini is close to those by AdamW throughout training.

A.6 Sample Responses from LLMs trained by Adam-mini

The sample response in shown in Figure 17.

B More Experimental Details

B.1 Training configurations for Section 3

For all experiments, we choose the model configurations (e.g. context length) by their standard protocols. We choose the learning rates by the recommendation from open-source platforms if applicable. For instance, for GPT2 series, we use the recommended learning rates by [27], which are reported to be optimal by grid search. Unless mentioned otherwise, Adam-mini, Adafactor, SM3, and LAMB use the same learning rate as the recommended ones of AdamW. We choose the learning rate

Prompt: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

AdamW: Sure! Here's a travel blog post about a recent trip to Hawaii: Whether you're interested in history, nature, or simply soaking up the sun, Hawaii has something for everyone. Until next time, Aloha!

Adam-mini: Sure! Here's a travel blog post about a recent trip to Hawaii: I highly recommend visiting this beautiful archipelago for a truly unforgettable adventure.

Prompt: Rewrite your previous response. Start every sentence with the letter A.

AdamW: A trip to Hawaii is a must-visit destination for any traveler, offering a variety of cultural experiences and must-see attractions. From the Polynesian Cultural Center to the USS Arizona Memorial, and from Lanikai Beach to the lively night markets, every corner of the islands reveals a new surprise and a new way to connect with the spirit of Hawaii.

Adam-mini: A trip to Hawaii is a must-do for any travel enthusiast. A trip to the beautiful Hawaiian Islands is a paradise on earth. A visit to Oahu is a great way to start your adventure. A helicopter tour of the island is both educational and breathtaking. A visit to Maui is a must-see. A visit to Haleakala National Park is a highlight of your trip. A visit to the Polynesian Cultural Center is a great way to learn about the history and traditions of the islands. A traditional luau is a must-attend event. Overall, a trip to Hawaii is an unforgettable experience.

Figure 17: Response Sample of Llama2-7B finetuned by AdamW and Adam-mini to the #81 test question from the MT-Bench [61] dataset. Since there is little information in the first round response, we omit the answers from the models.

of CAME following the suggestions from the authors⁶. If there is no public recommended learning rate for AdamW, we tune the learning rate for all optimizers within the same computational budget and report the best performance. For other hyperparameters, we follow the recommendation from open-source platforms or by their default setting. For SM3 and Adafactor, we incorporate momentum with $\beta_1 = 0.9$ to offer a fair comparison with other optimizers and the rest of the hyperparameters are set as default. The detailed configurations are explained as follows.

GPT2 pre-training. We use the nanoGPT codebase⁷ to train GPT2 sized 125M (small), 330M (medium), 770M (large), and 1.5B (XL) on Openwebtext. For all models, we use context length = 1024, batch size = 480, weight decay coefficient $\lambda = 0.1$, $\epsilon = 1e-8$, $\beta_1 = 0.9$, $\beta_2 = 0.95$. We use cosine-decay learning rate schedule with 2000 iterations of warm-up. For GPT2-small, medium, and large, we use the recommended peak learning rate by [27], which are reported to be the optimal ones found by grid search. For GPT2-XL, we use the recommended peak learning rate by the Levanter⁸. The chosen peak learning rates are 6e-4, 3e-4, 2e-4, 1e-4 for GPT2-small, medium, large, XL, respectively. The minimal learning rate is chosen as 3e-5, 6e-5, 1e-5, 1e-5 for these models.

Llama pre-training. We use TinyLlama codebase⁹ to train TinyLlama-1B and Llama2-7B on CommonCrawl¹⁰. We use batch size = 40. We preprocess the directionary “2019-30” of CommonCrawl into 10850 bins, which in total gives 85GB of tokens. We leave the last 20 bins out as the validation set. For both models, we use weight decay coefficient $\lambda = 0.1$, $\epsilon = 1e-8$, $\beta_1 = 0.9$, $\beta_2 = 0.95$. We use cosine-decay learning rate schedule with 2000 iterations of warm-up. For TinyLlam-1B, we use context length = 2048, batch size = 512, peak learning rate = 2e-4, and minimal learning rate = 2e-5. For Llama2-7B, we use context length = 4096, batch size = 256, peak learning rate = 8e-5, and minimal learning rate = 8e-6.

SFT and RLHF. The implementation of SFT and RLHF code is based on the ReMax codebase¹¹. We use DeepSpeed ZeRO-2 in our training. GPT-4 evaluation template in Table 3 is from the

⁶<https://github.com/yangluo7/CAME>

⁷<https://github.com/karpathy/nanoGPT/tree/master>

⁸https://github.com/stanford-crfm/levanter/blob/e183ec80ec5971b12d4a3fb08a160268de342670/config/gpt2_xl.yaml

⁹<https://github.com/jzhang38/TinyLlama>

¹⁰<https://commoncrawl.org/>

¹¹<https://github.com/liziniu/ReMax>

codebase¹². In the reward optimization stage, We use ReMax, a memory-efficient alternative to PPO. We use UltraFeedback dataset [8] and use 40% data for SFT and 60% data for ReMax.

SFT. We use 80 samples in a batch and train the model for 3 epochs. For the full parameter tuning, we search the learning rate from {1e-6, 2e-6, 3e-6, 4e-6, 5e-6, 1e-5, 2e-5} based on validation loss, and we use 2e-6 with cosine annealing for both AdamW and Adam-mini. For LoRA, We apply LoRA for all layers except the embedding layer. The rank of LoRA is set to 128. After selecting the learning rate from the same set as the full parameter tuning, we use 5e-6 for both AdamW and Adam-mini when LoRA is applied. The weight decay coefficient is set to 0 as recommended by LlamaFactory¹³. The rest of the hyperparameters of AdamW and Adam-mini are $\epsilon = 1e-8$, $\beta_1 = 0.9$, $\beta_2 = 0.95$.

ReMax. We use 48 samples in a batch and train the model for 1 epoch. By searching the peak learning rate from {5e-7, 1e-6, 2e-6} based on validation reward, AdamW uses 1e-6 while Adam-mini selects 5e-7 as the peak learning rate. The weight decay coefficient is set to 0. The rest of the hyperparameters of AdamW and Adam-mini are $\epsilon = 1e-8$, $\beta_1 = 0.9$, $\beta_2 = 0.95$.

ResNet. We use the PyTorch official implementation codebase¹⁴ to train ResNet18 [17] on ImageNet [10]. We use cosine-decay learning rate, epoch = 90, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$. For ResNet18, we use batch size = 256, peak learning rate = 0.005. For ViT-base, we use batch size = 128, peak learning rate = 0.0001. These configurations are used for both Adam-mini and AdamW.

Diffusion models. We use the codebase¹⁵ to train diffusion models. The image size is 64 and the training objective is to predict the noise as in [18]. We use the default U-Net architecture hyperparameters and the dimension multiply in U-Net is (1, 2, 4, 8). We use the CelebA dataset¹⁶ and train the diffusion model with a learning rate 5×10^{-5} with cosine decay. The batch size is 128 and the training epoch is 50.

Graph Neural Networks. We use the DGL implementation¹⁷ of Graph Convolution Networks (GCN) [23] and Graph Attention Networks (GAT) [51] for OGBN-arxiv¹⁸ dataset. All configurations as default. For both Adam-mini and AdamW, we use the default learning rate = 0.005 for GCN and the default learning rate = 0.002 for GAT.

B.2 Detailed Setup for Other Experiments

Configuration for Figure 4. For each dense sub-block $H_l, l = 1, 2, 3$, we use random positive definite matrices. We fix the choose the eigenvalues of each H_l as follows: for $l = 1$, we independently sample from {1,2,3} for 30 times; for $l = 2$, we repeat this procedure for {99,100,101}; for $l = 3$, we repeat this procedure for {4998 ,4999, 5000}. For the single (blockwise) learning rate method, we use GD with optimal constant learning rate $2/(L + \mu)$, where L, μ are the largest and smallest eigenvalue of the (blockwise) Hessian. We use Adam with $\beta_1 = 0$. This helps us focus on the effect of coordinatewise learning rate in Adam. We also set $\beta_2 = 1$ to the time-varying learning rate. This is necessary because, for any $\beta_2 < 1$, Adam with constant learning rate will oscillate on quadratic functions. This is theoretically proved in [9, Proposition 12, Figure 1] and empirically observed in [59, Section 3.3].

Configuration for Figure 6. We use the codebase of¹⁹. We consider a 4-layer Transfomer with $n_emb = 32$, $n_head = 4$. The dataset is a randomly generated binary sequence of length 42. With the help of auto-differentiation framework, we calculate the Hessian on the whole dataset with two passes of backpropagation [38] and the calculation is exact. In Figure 6, we first set the largest value

¹²https://github.com/lm-sys/FastChat/tree/main/fastchat/llm_judge

¹³<https://github.com/hiyouga/LLaMA-Factory>

¹⁴<https://github.com/pytorch/examples/blob/main/imagenet/main.py>

¹⁵<https://github.com/lucidrains/denoising-diffusion-pytorch>

¹⁶<https://cseweb.ucsd.edu/~weijian/static/datasets/celeba/>

¹⁷<https://github.com/dmlc/dgl/tree/master/examples/pytorch/ogb/ogbn-arxiv>

¹⁸<https://ogb.stanford.edu/docs/nodeprop/>

¹⁹<https://colab.research.google.com/drive/1SiF0KZJp75rUeetKOWqpsA8clmHP6jMg?usp=sharing>

of color bar v_{max} as $1e-6$ for all blocks, which is roughly the maximal value in most blocks. But then we find the whole query and key blocks are black, so we decrease v_{max} to $1e-8$ for these blocks and then the block diagonal pattern becomes clear.

Throughput Comparison in Table 2. The results are tested on $2 \times$ A800-80GB GPUs using Tinyllama codebase. We did not turn on CPU offload. We report the throughput from the summary file of Wandb log.