SPARSE GRADIENT COMPRESSION FOR FINE-TUNING LARGE LANGUAGE MODELS

David H. Yang[1], **Mohammad Mohammadi Amiri**[1]
Tejaswini Pedapati[2], **Subhajit Chaudhury**[2], **Pin-Yu Chen**[2]
[1]Rensselaer Polytechnic Institute, [2]IBM Research

ABSTRACT

Fine-tuning large language models (LLMs) for downstream tasks has become increasingly crucial due to their widespread use and the growing availability of open-source models. However, the high memory costs associated with fine-tuning remain a significant challenge, especially as models increase in size. To address this, parameter efficient fine-tuning (PEFT) methods have been proposed to minimize the number of parameters required for fine-tuning LLMs. However, these approaches often tie the number of optimizer states to dimensions of model parameters, limiting flexibility and control during fine-tuning. In this paper, we propose sparse gradient compression (SGC), a training regime that leverages inherent sparsity in gradients to compress optimizer states by projecting them onto a low-dimensonal subspace, with dimensionality independent of the original model's parameters. By enabling optimizer state updates in an arbitrary low-dimensional subspace, SGC offers a flexible tradeoff between memory efficiency and performance. By fine-tuning LLMs on downstream tasks, we show that SGC can deliver superior performance while substantially lowering optimizer state memory requirements, particularly in both data-limited and memory-limited settings.

1 INTRODUCTION

As LLMs scale to billions of parameters, fine-tuning all the parameters of a model becomes increasingly challenging, demanding substantial memory resources (Touvron et al., 2023; Dubey et al., 2024). Full fine-tuning requires not only storing billions of model weights, but also maintaining the gradients and optimizer states needed during training, which can drastically increase memory consumption (Chowdhery et al., 2022; Bai et al., 2023). For example, the Adam optimizer requires storing both the first-and second-order moments of the gradients, doubling the memory needed compared to storing the model's trainable parameters (Kingma & Ba, 2017).

To address this problem, parameter efficient fine-tuning (PEFT) techniques have been introduced, to train a model using a significantly smaller number of parameters (Ding et al., 2023; Han et al., 2024). However, many existing methods lack the ability to provide both *flexible* and *granular* control over the number of optimizer states used for fine-tuning (see Appendix A for related works). Flexibility refers to the capacity to accommodate a broad range in the number of optimizer states, while granular control refers to the precision with which the number of optimizer states can be adjusted in small increments. This limitation may hinder the realization of a broader range of memory-performance tradeoffs, thereby restricting the potential of PEFT methods to achieve further efficiency gains.

On one end, we have approaches like BitFit (Zaken et al., 2022), which fine-tune on only the bias terms, using a minimal number of parameters, but is neither flexible nor offers granular control. On the other hand, the popular low-rank adaptation (LoRA) is a more flexible approach that provides some control over the number of trainable parameters (Hu et al., 2021). However, there still exists limitations to both flexibility and granularity. LoRA reparameterizes the fine-tuned weight matrices $\boldsymbol{W}^{(1)} \in \mathbb{R}^{m\times n}$ into $\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(0)} + \boldsymbol{B}\boldsymbol{A}$, where $\boldsymbol{W}^{(0)} \in \mathbb{R}^{m\times n}$ is the frozen pre-trained weight matrix, and $\boldsymbol{A} \in \mathbb{R}^{r\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times r}$ are two low-rank matrices of rank r ($r \ll \min\{m, n\}$) to be trained. However, with LoRA, the number of optimizer states is a function of the dimensions of $\boldsymbol{A}$ and $\boldsymbol{B}$, which are dependent on n and m, respectively. The minimum number of trainable parameters (achieved when $r = 1$) is equal to $n+m$, limited by the dimensions of $\boldsymbol{W}^{(0)}$. Therefore, there exists a bound dependent on $n + m$ in which we cannot reduce the number of optimizer states

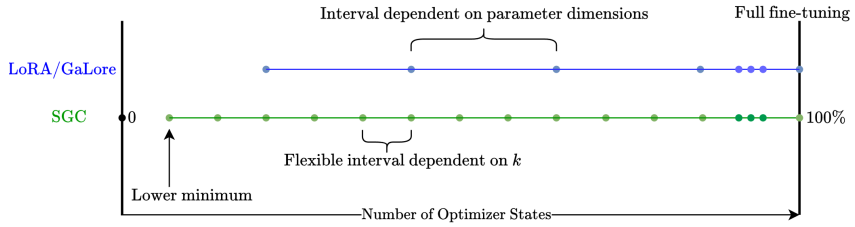


Figure 1: Diagram comparing SGC (green) and PEFT methods LoRA and GaLore (blue) in terms of the dimension of optimizer states compared to full fine-tuning. SGC enables a lower minimum and finer granularity for the number of optimizer states since it is independent of parameter dimensions.

during fine-tuning any further. A similar limitation exists with other approaches using prefix-tuning (Li & Liang, 2021) and gradient compression, such as GaLore (Zhao et al., 2024) (see Appendix B).

To address the above limitation, we propose sparse gradient compression (SGC), a training regime that enables more flexible and granular control over the number of parameters to train during fine-tuning. SGC updates the optimizer states in a k-dimensional subspace, where k is independent of the original parameters dimension. This allows SGC to significantly reduce the number of optimizer states, irrespective of the pretrained model's size, with k providing flexibility to balance performance and memory efficiency (see Figure 1). Importantly, this memory saving comes without sacrificing performance, as we will demonstrate in our experimental results.

The key idea behind SGC is leveraging the inherent sparsity of gradients during fine-tuning. By linearly projecting the optimizer states onto an arbitrarily lower-dimensional subspace, we can perform updates in this compressed space instead of the original space. A sparse recovery algorithm is then used to project the result of the optimizer function back into the original space, estimating the full-dimensional sparse vector from its lower dimensional representation, with sparsity originating from the gradients. By fine-tuning LLMs on downstream tasks, we show that SGC yields improved fine-tuning performance in both data-limited and optimizer state limited scenarios.

2 Problem Formulation

We investigate the task of updating the parameters of a neural network, $\boldsymbol{W} \in \mathbb{R}^d$, focusing specifically on *fine-tuning*, and without introducing any new weights into the model's architecture. The objective is to adapt pretrained weights $\boldsymbol{W}^{(0)} \in \mathbb{R}^d$ to $\boldsymbol{W}^{(1)} \in \mathbb{R}^d$ for a particular task.[1] The transition from $\boldsymbol{W}^{(0)}$ to $\boldsymbol{W}^{(1)}$ is defined as follows:

$$\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(0)} + \Delta \boldsymbol{W}. \tag{1}$$

To improve fine-tuning performance, an optimizer function $\rho_t(\cdot)$ is typically applied to the gradient $\boldsymbol{G}_t \in \mathbb{R}^d$, where ρ_t requires storing and updating additional optimizer states, each with the same dimensions as $\boldsymbol{G}_t$. Therefore, the memory requirements of applying the optimizer function is directly dependent on d. In LLMs, d can be substantially large, making the execution of the optimizer function $\rho_t(\cdot)$ highly resource-intensive. To address this, we define a transformation function that reduces the dimension of $\boldsymbol{G}_t$ before being used in the optimizer function ρ_t. Specifically, we define $f : \mathbb{R}^d \to \mathbb{R}^k$ as the transformation applied to gradient $\boldsymbol{G}_t$ as $\hat{\boldsymbol{G}}_t = f(\boldsymbol{G}_t)$ for some $k \ll d$. Now we use $\hat{\boldsymbol{G}}_t$ as the input to the optimizer function ρ_t, reducing the dimension of the operations in the optimizer from a d-dimensional space to a k-dimensional space. We denote the total changes in the parameters $\boldsymbol{W}$ after T time steps as:

$$\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(0)} - \eta \sum_t g(\rho_t(\hat{\boldsymbol{G}}_t)), \tag{2}$$

where η is the learning rate, and $g : \mathbb{R}^k \to \mathbb{R}^d$ is a transformation that brings the output of ρ_t back into the original d-dimensional space. This formulation allows us to perform the optimizer state updates in a smaller subspace $\mathbb{R}^k$ instead of the original space $\mathbb{R}^d$, where $k \ll d$. In practice, tracking the optimizer states in ρ_t can be memory intensive if k is large. Thus, the goal is to reduce k as much as possible while maintaining a reasonable performance.

[1] Without loss of generality, we represent model parameters as vectors instead of matrices.

3 METHODOLOGY

Sparse Gradient Compression (SGC). In full fine-tuning, the gradients input to optimizer functions like the well-known AdamW (Kingma & Ba, 2017), outlined in Algorithm 1 (see Appendix C), can have a large dimension d. We would like to modify AdamW to update $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$ on a k-dimensional subspace rather than the d-dimensional space, for some $k \ll d$, while retaining performance. This would significantly enhance the memory and compute efficiency of the optimizer, improving the efficiency of fine-tuning.

We first sparsify $\boldsymbol{G}_t \in \mathbb{R}^d$ by keeping only s non-zero elements corresponding to s entries with largest magnitudes, and set all other elements to zero which is denoted by $\text{Sparsify}_s(\cdot)$. The sparsified gradient is then projected onto a lower dimensional subspace of an arbitrary dimension k using a projection matrix $\boldsymbol{A} \in \mathbb{R}^{k \times d}$ that is initialized before fine-tuning:

$$\tilde{\boldsymbol{G}}_t = \text{Sparsify}_s(\boldsymbol{G}_t) \in \mathbb{R}^d, \quad \boldsymbol{p}_t = \boldsymbol{A}\tilde{\boldsymbol{G}}_t \in \mathbb{R}^k, \quad \boldsymbol{q}_t = \boldsymbol{A}\tilde{\boldsymbol{G}}_t^2 \in \mathbb{R}^k, \tag{3}$$

where we make use of the fact that elementwise squares retain the sparsity pattern of $\boldsymbol{G}_t$. With $\boldsymbol{G}_t$ and $\boldsymbol{G}_t^2$ represented in a compressed form with dimension k as $\boldsymbol{p}_t$ and $\boldsymbol{q}_t$, respectively, we modify Algorithm 1 by representing $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$ in this k-dimensional subspace as follows:

$$\boldsymbol{M}_t \leftarrow \beta_1 \boldsymbol{M}_{t-1} + (1-\beta_1)\boldsymbol{p}_t, \quad \boldsymbol{V}_t \leftarrow \beta_1 \boldsymbol{V}_{t-1} + (1-\beta_1)\boldsymbol{q}_t. \tag{4}$$

Accordingly, we can perform the updates on optimizer states $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$ on a k-dimensional subspace. However, we need to go back to the original d-dimensional space to perform the weight updates by using the function $g : \mathbb{R}^k \to \mathbb{R}^d$. Rewriting equation 2, this problem is equivalent to finding a function $g(\cdot)$ to perform the update $\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(0)} - \eta \sum_t g(\rho_t(\boldsymbol{p}_t, \boldsymbol{q}_t))$. Thus, this approach enables performing the updates on a k-dimensional subspace instead of the d-dimensional space using AdamW. The only missing part is how to define $g(\cdot)$ that enables going from a k-dimensional subspace back to the original d-dimensional space for the parameter updates.

Compressed Sensing of Optimizer States. We use a recovery algorithm from compressive sensing (CS) to achieve the function $g(\cdot)$, which aims to estimate a sparse vector from its compressed form, compressed through linear projection. CS is a signal processing technique used to recover signals using fewer measurements than the Nyquist rate, when the signals are sparse (Candes et al., 2004). Consider an s-sparse signal $\boldsymbol{x} \in \mathbb{R}^d$ with s non-zero entries. We can reconstruct $\boldsymbol{x}$ from a set of linear measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, if the measurement matrix $\boldsymbol{A} \in \mathbb{R}^{k \times d}$ satisfies the restricted isometry property (RIP) for some number of measurements $k \leq d$ (Candes & Tao, 2005).

There exist various recovery algorithms to recover the d-dimensional s-sparse signal $\boldsymbol{x}$ from measurements $\boldsymbol{y}$ (Marques et al., 2018). In this paper, we use a greedy algorithm named orthogonal matching pursuit (OMP) (Pati et al., 1993). To enhance efficiency, inspired by Zhu et al. (2020), we have developed a GPU optimized version of OMP, enabling its seamless integration with fine-tuning (see Appendix D for details). The OMP algorithm reconstructs an s-sparse vector $\boldsymbol{x}$ from the measurements $\boldsymbol{y}$ having knowledge about the measurement matrix $\boldsymbol{A}$ denoted as follows:

$$\hat{\boldsymbol{x}} = \text{OMP}_{\boldsymbol{A}}(\boldsymbol{y}). \tag{5}$$

We now apply the recovery algorithm OMP to map the updates $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$, given in 4, from the k-dimensional subspace back to the original d-dimensional space. With the initialization $\boldsymbol{M}_0 = \boldsymbol{0}$ and $\boldsymbol{V}_0 = \boldsymbol{0}$, we can rewrite the updates $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$ as:

$$\boldsymbol{M}_t = \boldsymbol{A}\sum_{i=1}^{t} h_i(\beta_1)\tilde{\boldsymbol{G}}_i, \quad \boldsymbol{V}_t = \boldsymbol{A}\sum_{i=1}^{t} h_i(\beta_2)\tilde{\boldsymbol{G}}_i^2, \tag{6}$$

where $h_i(\cdot)$ is a constant only a function of β_1 or β_2. We observe that $\sum_{i=1}^{t} h_i(\beta_1)\tilde{\boldsymbol{G}}_i$ and $\sum_{i=1}^{t} h_i(\beta_2)\tilde{\boldsymbol{G}}_i^2$ are linear combinations of the first and second moments of the sparsified gradients, respectively. Assuming that the total changes in the sparsity of $\boldsymbol{G}_t$ over all t can be bounded by some constant $\tilde{s} \ll d$, we can use the OMP algorithm as in 5 to almost accurately recover the original d-dimensional representations of $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$. After applying OMP to $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$ separately, we obtain the output of the optimizer function $\boldsymbol{N}_t$ as follows:

$$\boldsymbol{N}_t = \alpha \frac{\text{OMP}_{\boldsymbol{A}}(\boldsymbol{M}_t)}{\sqrt{\text{OMP}_{\boldsymbol{A}}(\boldsymbol{V}_t)} + \epsilon}, \tag{7}$$

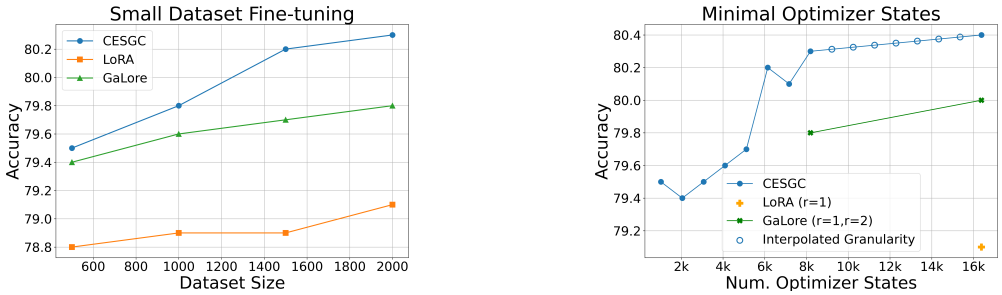
where α is a scaling factor. We note that the feasibility of obtaining $\boldsymbol{N}_t$ is ensured by the fact that $\tilde{\boldsymbol{G}}_t$ and $\tilde{\boldsymbol{G}}_t^2$, and thus $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$, share the same sparsity pattern. Consequently, the indices of the non-zero entries in $\text{OMP}_{\boldsymbol{A}}(\boldsymbol{M}_t)$ and $\text{OMP}_{\boldsymbol{A}}(\boldsymbol{V}_t)$ are identical. Furthermore, the sparsity level s provides a tradeoff between performance and efficiency. Clearly, a larger s leads to better performance since $\tilde{\boldsymbol{G}}_t$ provides a better estimate for $\boldsymbol{G}_t$; however, it increases the computational overhead with the OMP algorithm in recovering an s-sparse vector.

Following compression, the optimizer states $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$ are now k-dimensional vectors. Setting $k = \kappa s$ leads to a reasonable recovery using OMP, where κ is an algorithm dependent constant. Optimizer states in AdamW becomes purely a function of k, and can be controlled at a granular level. We refer to our proposed method as SGC, which uses the AdamW optimizer and is presented in Algorithm 3 (Appendix E). Incorporating this into our formulation in 2 yields $\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(0)} - \eta \sum_t \text{SGC}(\boldsymbol{G}_t)$. Memory and convergence analysis for SGC can be found in Appendix G and H.

4 Experiments

In this section, we analyze our compute efficient SGC, CESGC (Appendix F) approach in extreme scenarios of small datasets and optimizer states. To evaluate CESGC's effectiveness on small datasets, we fine-tune LLaMA2-7B on subsets of the BoolQ (Clark et al., 2019) dataset while using a minimal number of optimizer states (details can be found in Appendix J.4). Figure 2(a) shows that CESGC performs strictly better using small dataset sizes, suggesting that our approach can provide a more targeted method for fine-tuning by utilizing sparsity of the gradients. On the other hand, LoRA at the lowest rank ($r = 1$) struggles to learn under the limited dataset scenario, while GaLore with $r = 1$ underperforms CESGC.

By being independent of hidden dimension size, our approach enables fine-tuning using a smaller number of optimizer states than possible compared to both GaLore and LoRA (see Figure 2(b)). The increments that the CESGC optimizer states can increase is 1024, which is significantly less than both GaLore (8192) and LoRA (16384). This enables a finer sweep in the number of optimizer states to search for best hyperparameters to use. For instance, as shown in the figure, CESGC achieves 80.2% accuracy with using just over 6000 optimizer states, whereas both GaLore and LoRA are unable to obtain results since it is below the minimum number of optimizer state parameters they can support. Additional experiment results and ablation studies can be found in Appendix I.



(a) Small Dataset study (b) Optimizer State study

Figure 2: **(a)**. CESGC outperforms GaLore and LoRA when fine-tuning with limited data on BoolQ. **(b)**. CESGC, GaLore, and LoRA while using a minimal number of optimizer states. Hollow blue points are interpolated values that indicate the granularity of CESGC across optimizer states.

5 Conclusion

In this work, we proposed a novel fine-tuning method, SGC, that enables flexible and granular control over the number of optimizer states. The key idea is to leverage the sparsity of gradients and perform optimizer state updates in a lower subspace of an arbitrary dimension k. This allows SGC to have significantly smaller and more granular number of parameters to train during fine-tuning, achieving superior performance in data-limited settings compared to LoRA and GaLore. Our approach is also orthogonal to many gradient compression methods, opening opportunities for future work to integrate them and explore SGC's generalizability in domains like vision and audio.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and et. al. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Emmanuel Candes and Terence Tao. Decoding by linear programming, 2005. URL https://arxiv.org/abs/math/0502327.

Emmanuel Candes, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, 2004. URL https://arxiv.org/abs/math/0409186.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and et. al. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions, 2019. URL https://arxiv.org/abs/1905.10044.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv.org/abs/2305.14314.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, and et. al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL https://arxiv.org/abs/2403.14608.

Yongchang Hao, Yanshuai Cao, and Lili Mou. Flora: Low-rank adapters are secretly gradient compressors, 2024. URL https://arxiv.org/abs/2402.03293.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models, 2024. URL https://arxiv.org/abs/2402.12354.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning, 2022. URL https://arxiv.org/abs/2110.04366.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019. URL https://arxiv.org/abs/1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Ka-Wei Lee. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models, 2023. URL https://arxiv.org/abs/2304.01933.

Uijeong Jang, Jason D. Lee, and Ernest K. Ryu. Lora training in the ntk regime has no spurious local minima, 2024. URL https://arxiv.org/abs/2402.11867.

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

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. Vera: Vector-based random matrix adaptation, 2024. URL https://arxiv.org/abs/2310.11454.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning, 2021. URL https://arxiv.org/abs/2104.08691.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021. URL https://arxiv.org/abs/2101.00190.

Kaizhao Liang, Bo Liu, Lizhang Chen, and Qiang Liu. Memory-efficient llm training with online subspace descent, 2024. URL https://arxiv.org/abs/2408.12857.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation, 2024. URL https://arxiv.org/abs/2402.09353.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks, 2022. URL https://arxiv.org/abs/2110.07602.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks, 2021. URL https://arxiv.org/abs/2106.04489.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on lora of large language models, 2024. URL https://arxiv.org/abs/2407.11046.

Elaine Crespo Marques, Nilson Maciel, Lirida Naviner, Hao Cai, and Jun Yang. A review of sparse recovery algorithms. *IEEE access*, 7:1300–1322, 2018.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pp. 40–44. IEEE, 1993.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning, 2021. URL https://arxiv.org/abs/2005.00247.

Haotong Qin, Xudong Ma, Xingyu Zheng, Xiaoyang Li, Yang Zhang, Shouda Liu, Jie Luo, Xianglong Liu, and Michele Magno. Accurate lora-finetuning quantization of llms via information retention, 2024. URL https://arxiv.org/abs/2402.05445.

Weixi Song, Zuchao Li, Lefei Zhang, Hai Zhao, and Bo Du. Sparse is enough in fine-tuning pre-trained large language models, 2024. URL https://arxiv.org/abs/2312.11875.

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory, 2018. URL https://arxiv.org/abs/1809.07599.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Huiwen Wu, Xiaohan Li, Deyi Zhang, Xiaogang Xu, Jiafei Wu, Puning Zhao, and Zhe Liu. Cg-fedllm: How to compress gradients in federated fune-tuning for large language models, 2024. URL https://arxiv.org/abs/2405.13746.

Wenhan Xia, Chengwei Qin, and Elad Hazan. Chain of lora: Efficient fine-tuning of language models via residual learning, 2024. URL https://arxiv.org/abs/2401.04151.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022. URL https://arxiv.org/abs/2106.10199.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning, 2023. URL https://arxiv.org/abs/2303.10512.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection, 2024. URL https://arxiv.org/abs/2403.03507.

Hufei Zhu, Wen Chen, and Yanpeng Wu. Efficient implementations for orthogonal matching pursuit. *Electronics*, 9(9):1507, 2020.

Algorithm 1 AdamW at timestep t

Inputs: $\boldsymbol{G}_t, \beta_1, \beta_2, \epsilon$
$\boldsymbol{M}_t \leftarrow \beta_1 \boldsymbol{M}_{t-1} + (1-\beta_1)\boldsymbol{G}_t$
$\boldsymbol{V}_t \leftarrow \beta_2 \boldsymbol{V}_{t-1} + (1-\beta_2)\boldsymbol{G}_t^2$
$\boldsymbol{M}_t \leftarrow \frac{\boldsymbol{M}_t}{1-\beta_1^t}$
$\boldsymbol{V}_t \leftarrow \frac{\boldsymbol{V}_t}{1-\beta_2^t}$
$\boldsymbol{N}_t = \frac{\boldsymbol{M}_t}{\sqrt{\boldsymbol{V}_t}+\epsilon}$
return $\boldsymbol{N}_t$

A Related Works

Parameter Efficient Fine-tuning. PEFT methods are used to reduce the expensive memory requirements for fine-tuning large models. Existing techniques can be split into several categories. Adapter-based methods introduce additional trainable modules that are inserted into the original frozen model (Houlsby et al., 2019; Pfeiffer et al., 2021; He et al., 2022; Mahabadi et al., 2021). However, these approaches can introduce latency during inference. Prompt tuning, on the other hand, adapts a model by adding learnable prefix tokens to the input (Li & Liang, 2021; Lester et al., 2021; Liu et al., 2022). Despite their simplicity, these methods have structural limitations since they only train additional input tokens. LoRA is a widely used PEFT method that does not introduce additional inference latency (Hu et al., 2021). LoRA employs low-rank matrices to approximate the updates in the parameters during fine-tuning. Several variants of LoRA have been developed to either improve performance or further reduce the number of trainable parameters (Zhang et al., 2023; Xia et al., 2024; Liu et al., 2024; Kopiczko et al., 2024). Due to LoRA's popularity, extensive research has been conducted on both its theoretical foundations and empirical performance (Jang et al., 2024; Hayou et al., 2024; Mao et al., 2024). Additionally, quantization-based methods have been proposed to further reduce memory overhead Dettmers et al. (2023); Qin et al. (2024).

Gradient Compression. An area that has been relatively underexplored but is now gaining attention is gradient compression (Zhao et al., 2024; Hao et al., 2024; Liang et al., 2024; Wu et al., 2024; Song et al., 2024). These approaches selectively compress gradient information to reduce the size of optimizer states during training. One category of methods uses projection matrices to obtain a lower-rank gradients (Zhao et al., 2024; Hao et al., 2024; Liang et al., 2024). For instance, GaLore uses singular value decomposition (SVD) to obtain projection matrices (Zhao et al., 2024), while FLoRA utilizes random projection matrices (Hao et al., 2024). Liang et al. (2024) propose a method that updates the projection matrix in an online fashion using principal component analysis. Alongside projection matrices, gradient sparsity is another emerging factor. SIFT shows that gradients are approximately sparse, and achieves efficient fine-tuning by selecting parameters corresponding to the largest gradient magnitudes (Song et al., 2024). However, a significant limitation of this approach is that the selected parameters remain static, failing to fully capture the dynamic nature of gradient sparsity patterns during training.

B GaLore Analysis

Rather than operating on the parameter space, GaLore saves memory by reducing the number of parameters in the optimizer states (Zhao et al., 2024). Specifically, it projects the gradient $\boldsymbol{G}_t \in \mathbb{R}^{m\times n}$ at each time step t to a lower-dimensional representation $\hat{\boldsymbol{G}}_t = \boldsymbol{P}_t\boldsymbol{G}_t \in R^{r\times n}$ by using a projection matrix $\boldsymbol{P}_t \in \mathbb{R}^{r\times m}$ that is set to the first r columns of the left singular vectors of SVD of $\boldsymbol{G}_t$. The size of the optimizer states, which are equal to the dimensions of the projected gradient $\hat{\boldsymbol{G}}_t$ is then reduced, providing memory savings. However, observe that $\hat{\boldsymbol{G}}_t$ is still dependent on n, meaning that, similar to LoRA, there exists a bound dependent on n that we cannot reduce the number of optimizer states any further. Likewise, granularity over parameters is a function of n, and tied to the model's weight dimensions.

C AdamW Algorithm

AdamW optimizer function updates (see Algorithm 1).

D EFFICIENT ORTHOGONAL MATCHING PURSUIT

Our implementation of OMP is based on the inverse Cholesky factorization method (Zhu et al., 2020), see Algorithm 2. We perform pre-calculation of the gram matrix $\boldsymbol{G}$, to reduce computational costs, but at the same time introduce additional implementation memory requirements. For pure memory efficiency, $\boldsymbol{G}$ should not be pre-computed or alternatively, it is possible to implement a more memory efficient Algorithm 2 at the expense of additional runtime.

Algorithm 2 OMP by Inverse Cholesky Factorization

Inputs: measurements $\boldsymbol{y}$, projection matrix $\boldsymbol{A}$, sparsity value s
Initialize: $\Lambda_0 = \emptyset$, the residual $\boldsymbol{r}^{(0)} = \boldsymbol{y}$, gram matrix $\boldsymbol{G} = \boldsymbol{A}^H\boldsymbol{A}$, and the iteration counter $k = 1$.
while $k \leq s$ **do**
Projection: if $k = 1$, compute $\boldsymbol{p}^0 = \boldsymbol{A}^H\boldsymbol{r}^0$, else

$$\boldsymbol{p}^{(k-1)} = \boldsymbol{p}^{(k-2)} - \boldsymbol{b}_{:(k-1)}a_{k-1},$$

where $\boldsymbol{b}_{:(k-1)}$ is the $(k-1)$-th column of $\boldsymbol{B}_{k-1}$, and $\boldsymbol{a}_{k-1}$ is the $(k-1)$-th entry of $\boldsymbol{a}_{k-1}$.
Select $i^{(k)} = \arg\max_{i=1,2,\ldots,d}\left(\frac{|p_i^{(k-1)}|}{\|\boldsymbol{A}_{:i}\|}\right)$, where $p_i^{(k-1)}$ is the i-th entry of $\boldsymbol{p}^{(k-1)}$.

$$\text{Let } \Lambda_k = \Lambda_{k-1} \cup \{i^{(k)}\}, \quad \text{i.e.,} \quad \lambda_k = i^{(k)} \text{ is the k-th entry of the set } \Lambda_k.$$

Obtain

$$\boldsymbol{c}_{k-1} = \left(\boldsymbol{b}^H_{\lambda_k,1:\Lambda_{k-1}}\right)^H,$$

where $\boldsymbol{b}_{\lambda_k,1:\Lambda_{k-1}}$ is the λ_k-th row of $\boldsymbol{B}_{k-1}$. Then compute $\gamma_k = \frac{1}{\sqrt{g_{\lambda_k,\lambda_k} - \boldsymbol{c}^H_{k-1}\boldsymbol{c}_{k-1}}}$,

$$a_k = \gamma_k p^{k-1}_{\lambda_k},$$

$$\boldsymbol{a}_k = \begin{bmatrix}\boldsymbol{a}^T_{k-1} & a_{:k}\end{bmatrix}^T,$$
$$\boldsymbol{b}_{:k} = \gamma_k\left(\boldsymbol{g}_{:\lambda_k} - \boldsymbol{B}_{k-1}\boldsymbol{c}_{k-1}\right),$$
$$\boldsymbol{B}_k = \begin{bmatrix}\boldsymbol{B}^T_{k-1} & b_{:k}\end{bmatrix},$$

where $p^{k-1}_{\lambda_k}$ is the λ_k-th entry of $\boldsymbol{p}^{k-1}$, $\boldsymbol{g}^k_{:\lambda_k}$ is the λ_k-th column of $\boldsymbol{G}$, and $\boldsymbol{c}_0 = \boldsymbol{B}_0 = \boldsymbol{a}_0 = \emptyset$ is assumed for $k = 1$. Finally, if $k = 1$, compute $\boldsymbol{F}_1 = \sqrt{g_{\lambda_1,\lambda_1}}$, else

$$\boldsymbol{F}_k = \begin{bmatrix}\boldsymbol{F}_{k-1} & -\gamma_k\boldsymbol{F}_{k-1}\boldsymbol{c}_{k-1} \\ \boldsymbol{0}_{k-1} & \gamma_k\end{bmatrix},$$

$k := k + 1$.
end while
Output: Compute $\hat{\boldsymbol{x}}_s = \boldsymbol{F}_s\boldsymbol{a}_s$, $\boldsymbol{r}^{(s)} = y - \boldsymbol{A}_{\Lambda_s}\hat{\boldsymbol{x}}_s$, and return $\boldsymbol{r}^{(s)}, \Lambda_s, \hat{\boldsymbol{x}}_s$.

E SGC ALGORITHM

Sparse Gradient Compression (SGC) algorithm sparsifies and compresses gradients to a lower dimensional subspace to perform optimizer state updates, and uses Orthogonal Matching Pursuit (OMP) to perform weight updates in the original dimension (see Algorithm 3).

F EFFICIENT SGC

Here, we propose two efficient alternatives of the SGC algorithm.

Memory Efficient SGC (MESGC). Based on our observations, size of the projection matrix $\boldsymbol{A} \in \mathbb{R}^{k\times d}$ may significantly contribute to the computation overhead. Although it is initialized only once before fine-tuning, the memory requirements can become substantial depending on the value of s, the sparsity level of $\tilde{\boldsymbol{G}}_t$, particularly when applying the OMP algorithm. To address this issue, we

Algorithm 3 SGC at timestep t

Inputs: $\boldsymbol{G}_t, \boldsymbol{A}, s, \beta_1, \beta_2, \epsilon$
$\boldsymbol{p}_t = \boldsymbol{A}$ Sparsify$_s(\boldsymbol{G}_t)$, $\boldsymbol{q}_t = \boldsymbol{A}$ Sparsify$_s(\boldsymbol{G}_t^2)$
$\boldsymbol{M}_t \leftarrow \beta_1 \boldsymbol{M}_{t-1} + (1-\beta_1)\boldsymbol{p}_t$
$\boldsymbol{V}_t \leftarrow \beta_2 \boldsymbol{V}_{t-1} + (1-\beta_2)\boldsymbol{q}_t$
$\boldsymbol{M}_t \leftarrow \frac{\boldsymbol{M}_t}{1-\beta_1^t}$
$\boldsymbol{V}_t \leftarrow \frac{\boldsymbol{V}_t}{1-\beta_2^t}$
$\boldsymbol{N}_t = \alpha \frac{\text{OMP}_{\boldsymbol{A}}(\boldsymbol{M}_t)}{\sqrt{\text{OMP}_{\boldsymbol{A}}(\boldsymbol{V}_t)}+\epsilon}$
return $\boldsymbol{N}_t$

introduce the idea of chunking the gradient signals prior to applying a projection matrix. Specifically, we split $\boldsymbol{G}_t$ into c equal sized chunks before sparsifying and projecting each chunk. This enables the projection matrix $\boldsymbol{A}$ to be much smaller in size from $k \times d$ to $(k \times d)/c$. We split $\boldsymbol{G}_t$ to c equal-size chunks $\boldsymbol{G}_t = \left[\boldsymbol{G}_t^1, \ldots, \boldsymbol{G}_t^c\right]$ and apply the SGC algorithm to each $\boldsymbol{G}_t^i$. Accordingly, we have $\boldsymbol{N}_t^i = \text{SGC}(\boldsymbol{G}_t^i) \in \mathbb{R}^{\frac{d}{c}}$, and we concatenate all these outputs to obtain $\boldsymbol{N}_t$ as $\boldsymbol{N}_t = \left[\boldsymbol{N}_t^1, \ldots, \boldsymbol{N}_t^c\right]$. We select $s_c = s/c$ non-zero elements per chunk to ensure s non-zero entries overall. Since the projection matrix $\boldsymbol{A}$ is the same for each chunk, we obtain efficiency by a factor of c for storing $\boldsymbol{A}$. However, we may not achieve an exact estimate of $\tilde{\boldsymbol{G}}_t$ and $\tilde{\boldsymbol{G}}_t^2$ when sparsifying and concatenating $\boldsymbol{G}_t^i$'s because the sparsity pattern in $\boldsymbol{G}_t$ is not truly uniform. This performance loss would be more severe with increasing c, while it enhances efficiency by reducing the dimension of the projection matrix $\boldsymbol{A}$. We note that the chunking technique introduces more flexibility with the proposed SGC approach in realizing a more diverse spectrum of performance-efficiency tradeoff.

Compute Efficient SGC (CESGC). The main tradeoff for our memory efficient approach is increased runtime attributed to OMP, which scales with d, the size of gradients $\boldsymbol{G}_t$. Here, we present a computationally efficient alternative at the expense of slightly increased memory usage. For ease of presentation here, consider $\boldsymbol{G}_t \in \mathbb{R}^{m \times n}$ to be in a matrix form. The main idea is to perform double compression, where we first compress $\boldsymbol{G}_t$ once using a projection matrix $\boldsymbol{B}_t \in \mathbb{R}^{r \times m}$, and then apply SGC to this compressed gradient of dimension $(r \times n) \ll d$, therefore reducing time complexity. The intuition behind this approach is that the resultant vector after the first compression is still quasi-sparse. The projection matrix $\boldsymbol{B}_t$ should be selected such that as much information is retained after projection. For this purpose, we use the fact that SGC is orthogonal to many PEFT methods. Thus, we apply one of these methods, GaLore, to obtain $\boldsymbol{B}_t$, which reduces the dimension of the vector entering the SGC algorithm. Specifically, we initialize the projection matrix $\boldsymbol{B}_t$ every fixed number of iterations by applying truncated SVD on $\boldsymbol{G}_t$:

$$\boldsymbol{U}, \Lambda, \boldsymbol{V} = \text{SVD}(\boldsymbol{G}_t), \quad \boldsymbol{B}_t = \boldsymbol{U}[:, :r] \in \mathbb{R}^{r \times m},$$

where $\boldsymbol{B}_t$ is set to be the first r columns of the left-singular vectors of SVD of $\boldsymbol{G}_t$. We then project the gradients $\boldsymbol{G}_t$ using $\boldsymbol{B}_t$ and apply SGC to the resultant vector, i.e., SGC($\boldsymbol{B}_t\boldsymbol{G}_t$). Finally, we project back the resultant updates from SGC($\boldsymbol{B}_t\boldsymbol{G}_t$) onto the original d-dimensional space using $\boldsymbol{B}_t^T$ to update the parameters in $\boldsymbol{W}$. Incorporating this into our formulation in 2 yields:

$$\boldsymbol{W}^{(1)} = \boldsymbol{W}^{(0)} - \eta \sum_t \boldsymbol{B}_t^T \text{SGC}(\boldsymbol{B}_t\boldsymbol{G}_t). \tag{8}$$

We note that the dimension of the vector entering SGC is $r \times n$ rather than d, thus improving the compute efficiency with OMP. CESGC can be combined with our memory efficient implementation, where chunking is performed after the projection of $\boldsymbol{G}_t$, and we assume this is performed by default for experiments using CESGC.

G Memory Analysis

Here, we analyze the memory requirements of our efficient SGC implementations and compare it with popular PEFT methods, specifically GaLore and LoRA. The memory requirements of our approach, Galore, and LoRA to perform weight updates for a single vector are shown in Table 1. Observe that the number of optimizer states in both Galore and LoRA are a function of d. On the other hand, the size of optimizer states for our memory efficient approach is independent of the

Table 1: Comparison between our approach, GaLore, and LoRA for storing the trainable parameters during fine-tuning with AdamW. For simplicity, assume weight dimensions d can be reshaped to 2D matrix of size $\sqrt{d} \times \sqrt{d}$, $r \ll d$ is the chosen rank, $k \ll d$ is the dimension we want to compress each optimizer state to. The projection matrices refer to the costs of storing $\boldsymbol{B}_t$ during fine-tuning.

	MESGC	**CESGC**	**GaLore**	**LoRA**
Weights	d	d	d	$d + 2r\sqrt{d}$
Optimizer States	$2k$	$2k$	$2r\sqrt{d}$	$4r\sqrt{d}$
Projection Matrices	-	$r\sqrt{d}$	$r\sqrt{d}$	-

weight dimensions, and only depends on $k = \kappa c s_c$, where s_c is sparsity per chunk, c is the number of chunks, and the constant κ is to satisfy the RIP conditions for the OMP algorithm. This enables our approach to be significantly more memory efficient in the optimizer states.

H Convergence Analysis

Following Stich et al. (2018), it is possible to show that top-k sparsification leads to convergence at the same rate as vanilla SGD. The key difference in our algorithm is the use of chunking and sparsification applied to every chunk. Thus, the proof of convergence boils down to bounding the distance between the sparse form of gradient vector $\boldsymbol{G}$ and the sparse form of every sub-vector after chunking the gradient vector $\boldsymbol{G}$. [Chunk-based s-sparsification] Let $\boldsymbol{G} \in \mathbb{R}^d$ be a gradient vector, partitioned into c equally sized chunks:

$$\boldsymbol{G} = \big[\boldsymbol{G}^1, \ldots, \boldsymbol{G}^c\big], \quad \boldsymbol{G}^i \in \mathbb{R}^{\frac{d}{c}}, \quad i = 1, \ldots, c.$$

We define the *chunk-based s-sparsified* vector $\tilde{\boldsymbol{G}}'$ by applying an s_c-sparsification to each chunk, where $s = \sum_{i=1}^{c} s_c$. Concretely,

$$\tilde{\boldsymbol{G}}' = \big[\tilde{\boldsymbol{G}}^1, \ldots, \tilde{\boldsymbol{G}}^c\big], \quad \tilde{\boldsymbol{G}}^i = \mathrm{Sparsify}_{s_c}\big(\boldsymbol{G}^i\big).$$

That is, within each chunk $\boldsymbol{G}_i$, we keep exactly the top s_c magnitude entries and set the rest to zero.

Separately, we define the *global s-sparsified* vector

$$\tilde{\boldsymbol{G}} = \mathrm{Sparsify}_s\big(\boldsymbol{G}\big),$$

which keeps the top-s entries from the entire vector $\boldsymbol{G}$ rather than chunk-by-chunk.

[Worst-case bound on chunk-based vs. global sparsification] Let $\boldsymbol{G}, \tilde{\boldsymbol{G}}'$ and $\tilde{\boldsymbol{G}}$ be as in Definition H. Then, it holds that

$$\mathbb{E}\big[\|\tilde{\boldsymbol{G}}' - \tilde{\boldsymbol{G}}\|_2^2\big] \;\le\; 2\Big(1 - \frac{s}{d}\Big)\, G_{\max},$$

where $G_{\max}$ is an upper bound on $\mathbb{E}\big[\|\tilde{\boldsymbol{G}}'\|_2^2\big]$.

Proof. The worst-case scenario corresponds to when all s non-zero entries of $\tilde{\boldsymbol{G}}'$ are contiguous, and without loss of generality, located in indices 1 to s. Let $l = \big\lceil \frac{s}{d/c} \big\rceil$ be the number of chunks spanning these s non-zero indices of $\tilde{\boldsymbol{G}}'$. Decompose the total error:

$$D_1 = \mathbb{E}\Big[\sum_{i=1}^{l} \|\tilde{\boldsymbol{G}}'^i - \tilde{\boldsymbol{G}}^i\|_2^2\Big] \quad \text{and} \quad D_2 = \mathbb{E}\Big[\sum_{i=l+1}^{c} \|\tilde{\boldsymbol{G}}'^i - \tilde{\boldsymbol{G}}^i\|_2^2\Big].$$

Intuitively, D_1 captures missing entries in the first l chunks not selected by $\tilde{\boldsymbol{G}}'$, while D_2 captures "extra" entries in the other $c - l$ chunks that are selected but should be zero.

By bounding each term via

$$D_1 \;\le\; (s - l s_c)\, \mathbb{E}\Big[\tfrac{\|\tilde{\boldsymbol{G}}'\|_2^2}{s}\Big] \quad \text{and} \quad D_2 \;\le\; (c - l)\, s_c\, \mathbb{E}\Big[\tfrac{\|\tilde{\boldsymbol{G}}'\|_2^2}{s}\Big],$$

we obtain

$$\mathbb{E}\big[\|\tilde{\boldsymbol{G}}' - \tilde{\boldsymbol{G}}\|_2^2\big] \;=\; D_1 + D_2 \;\le\; 2\Big(1 - \tfrac{s}{d}\Big)\, \mathbb{E}\big[\|\tilde{\boldsymbol{G}}'\|_2^2\big] \;\le\; 2\Big(1 - \tfrac{s}{d}\Big)\, G_{\max},$$

which completes the proof. □

Table 2: LLaMA2-7B, LLaMA3-8B, and LLaMA2-13B on fine-tuning eight commonsense benchmarks (5 shots) using various PEFT methods. Average accuracy is reported in the final column. Note that # Params refers to percentage of optimizer states, $\boldsymbol{M}_t$ and $\boldsymbol{V}_t$, relative to full fine-tuning.

Model	Method	# Params (%)	ARC-e	ARC-c	BoolQ	HellaSwag	OBQA	PIQA	SIQA	WinoGrande	Average
LLaMA2-7B	Full Fine-tuning	100	82.5	55.4	83.8	77.8	45.8	80.1	55.4	77.8	69.8
	CESGC	0.08	82.9	53.9	82.9	77.5	44.8	79.9	54.2	74.5	68.7
	GaLore	0.10	82.3	54.1	81.7	78.2	45.8	80.6	53.5	75.3	**68.9**
	LoRA	0.20	82.1	53.2	84.3	76.2	44.0	80.4	54.0	76.5	68.8
LLaMA3-8B	Full Fine-tuning	100	85.8	62.5	86.6	81.2	51.4	82.3	59.5	81.9	73.9
	CESGC	0.08	83.9	57.8	85.2	81.0	46.2	82.0	53.4	77.8	**70.9**
	GaLore	0.10	84.3	57.2	82.6	81.2	46.2	82.3	52.9	78.0	70.6
	LoRA	0.20	82.3	56.2	83.8	79.5	48.0	81.7	52.8	74.4	69.9
LLaMA2-13B	Full Fine-tuning	100	86.2	60.9	87.4	81.0	51.8	82.0	60.3	82.9	74.1
	CESGC	0.07	84.1	57.2	85.3	80.0	49.4	82.0	54.6	78.6	71.4
	GaLore	0.08	83.8	56.2	85.3	81.2	47.4	81.7	55.5	79.0	71.3
	LoRA	0.16	83.4	57.1	86.3	81.3	48.0	81.7	56.5	79.6	**71.7**

Table 3: Mistral-7B performance on the MMLU evaluation across various domains using different PEFT methods. Average accuracy is reported in the final column.

Method	STEM	Social Science	Humanities	Other	Average
CESGC	52.3	72.6	56.0	69.2	**61.9**
GaLore	52.3	72.6	56.0	69.0	61.8
LoRA	52.1	72.8	55.9	68.9	61.8

We note that for the uniform case where the non-zero entries of $\tilde{\boldsymbol{G}}'$ are uniformly distributed among the d indices, each chunk $\boldsymbol{G}_i$ is likely to contain about s_c of those entries. Thus, $\tilde{\boldsymbol{G}}' \approx \tilde{\boldsymbol{G}}$ in expectation, and

$$\mathbb{E}\big[\|\tilde{\boldsymbol{G}}' - \tilde{\boldsymbol{G}}\|_2^2\big] = 0.$$

Using these results, it is possible as future work to formulate the theoretical analysis for convergence, and we leave this as part of future work.

I Experiments

We evaluate our approach on fine-tuning various large languages models, specifically on LLaMA2-7B, LLaMA3-8B, and LLaMA2-13B, and Mistral-7B. The results are compared with full fine-tuning, LoRA, and GaLore as baseline for all the setups. In addition, we demonstrate the memory savings of the proposed MESGC algorithm and perform a set of ablation studies to illustrate the tradeoff between chunk size c and sparsity s. The results demonstrate how our approach can enable more granular control over the number of optimizer states and achieve comparable accuracy to LoRA and GaLore while using a significantly smaller number of optimizer states.

I.1 Commonsense and Knowledge Evaluation

We evaluate LLaMA2-7B, LLaMA3-8B, and LLaMA2-13B on a set of commonsense reasoning tasks to demonstrate CESGC's effectiveness in fine-tuning. Commonsense reasoning tasks involve 8 subtasks and we follow Hu et al. (2023) to combine the training sets into a single dataset and evaluate on each of the individual tasks separately. Details of hyperparameters and training settings can be found in Appendix J.1. Results from Table 2 show that our approach achieves a comparable average accuracy compared to both GaLore and LoRA, while using a smaller number of optimizer state parameters. Notably, in the LLaMA3-8B model, CESGC performs the best, achieving a superior accuracy of 1% over LoRA, while using less than half the number of optimizer state parameters. To further demonstrate the consistency of our approach, we fine-tune Mistral-7B on a subset of the cleaned Alpaca dataset Taori et al. (2023), and evaluate its performance on the MMLU benchmark (details can be found in Appendix J.2). These results indicate that our approach achieves competitive performance across different model types and tasks.

I.2 Memory Efficiency and Throughput

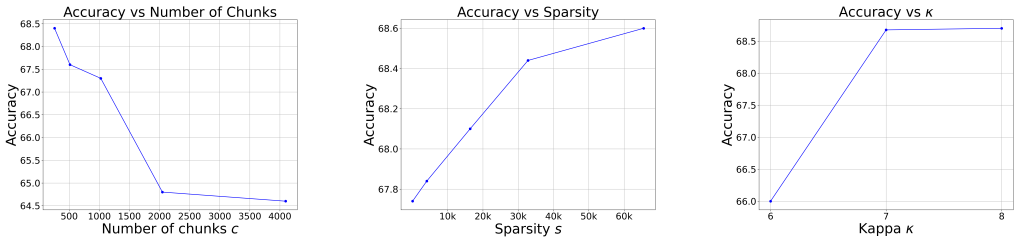
Consider $r = 1$, the minimum rank used for GaLore and LoRA. Based on Table 1, we can calculate that GaLore and LoRA require 8192 and 16384 optimizer states, respectively. With $s_c = 1$, $c =$

Table 4: Comparison of wall clock time per iteration between methods.

Method	Time per iteration (s)
Full Fine-tuning	1.69
LoRA	1.51
GaLore	1.88
MESGC	7.52
CESGC	2.82

Table 5: Fine-tuning results using a minimum number of optimizer states. MESGC conducted with $c = 256$, $s_c = 1$, $\kappa = 8$, while both GaLore and LoRA use rank $r = 1$.

Method	# Params	Accuracy
MESGC	4096	68.0
GaLore	8192	67.4
LoRA	16384	67.7



(a) Number of chunks study (b) Sparsity study (c) κ study

Figure 3: Ablation study for effects of number of chunks c, sparsity s, and constant κ. **(a)**. Average accuracy with varying c and constant s. **(b)**. Average accuracy with varying s and constant c. **(c)**. Average accuracy with varying κ.

64, and $\kappa = 7$, MESGC requires only 896 optimizer states, reducing the number of parameters by around 10 times. To demonstrate how MESGC performs using a significantly lower number of optimizer states, we fine-tune LLaMA2-7B on a subset of the commonsense reasoning dataset, setting $k = 2048$ (see Appendix J.3 for details). Table 5 shows that MESGC achieves 0.6% higher average accuracy than GaLore when fine-tuning LLaMA2-7B on commonsense reasoning while using only half the number of optimizer states. We also measure the throughput using wall clock time per iteration with the same fine-tuning task and compare our approaches with other methods. In particular, MESGC introduces some additional latency, but CESGC is optimized to be competitive with the baseline approaches.

I.3 Ablation Study

Here, we investigate the effects of number of chunks c, total sparsity s, and the constant κ on fine-tuning performance (details in Appendix J.5). First, we set the total sparsity s, to be constant and vary c. Figure 3(a) shows that increasing the number of chunks, while keeping the total s constant decreases average accuracy across the commonsense reasoning evaluation. We attribute this to the uniform chunking, where the number of non-zero elements selected per chunk is $s_c = s/c$. However, in practice, the sparsity pattern of gradients may vary across the chunks, with certain parameter regions potentially requiring more attention than others. Therefore, we see higher accuracy corresponding to smaller chunk sizes.

For sparsity, there is a general increasing trend, as seen in Figure 3(b). As the number of non-zero elements selected increases, so does the number of optimizer states k, we expect the accuracy to improve until s is equal to the number of parameters, as in full fine-tuning. We observe that increasing s after a certain point results in diminished returns seeing as the slope is most steep when s is increased initially and is less steep afterwards. This can be explained by how a small percentage of parameters account for the majority of the gradient norms during fine-tuning, which is supported by the observations in Song et al. (2024).

Finally, we investigate the effect of κ, the constant to satisfy the RIP condition, with the goal of finding a lower bound such that performance is not negatively affected. Based on Figure 3(c), we see that if κ is set to 6, performance drops significantly. However, there is minimal gain from increasing κ from 7 to 8, indicating a κ value of 7 should be sufficient.

Table 6: Hyperparameters used for commonsense reasoning experiments.

Model	Method	learning rate	rank r	num. chunks c	sparsity s	κ	α
LLaMA2-7B	Full Finetuning	1e-5	-	-	-	-	-
	CESGC	2e-5	32	64	1984	7	2
	GaLore	2e-5	4	-	-	-	2
	LoRA	1e-4	4	-	-	-	-
LLaMA3-8B	Full Finetuning	1e-5	-	-	-	-	-
	CESGC	2e-5	32	64	1984	7	2
	GaLore	2e-5	4	-	-	-	-
	LoRA	1e-4	4	-	-	-	-
LLaMA2-13B	Full Finetuning	1e-5	-	-	-	-	-
	CESGC	3e-5	32	64	2496	7	2
	GaLore	3e-5	4	-	-	-	2
	LoRA	1e-4	4	-	-	-	-

Table 7: Hyperparameters used for knowledge evaluation experiment.

Model	Method	learning rate	rank r	num. chunks c	sparsity s	κ	α
Mistral-7B	CESGC	2e-5	32	64	1984	7	2
	GaLore	2e-5	4	-	-	-	2
	LoRA	1e-4	4	-	-	-	-

J Experiment Details

J.1 Commonsense Reasoning

We fine-tune pretrained LLaMA2-7B, LLaMA2-13B, and LLaMA3-8B models obtained from Hugging Face. We trained each model for 1 epoch on the full commonsense dataset consisting of 170k examples. For consistency, we used a batch size of 16 across all experiments and train for 1 epoch. Since the goal is to observe performance improvements with only training a limited number of parameters, we only fine-tune on two of the attention matrices, keeping everything else frozen. For LlaMA2-7B and LLaMA-2-13B, we target the query and value matrices, whilst for LLaMA3-8B, we targeted the query output matrices. For LLaMA3-8B, we select the output matrix instead of the value matrix to keep the dimensions consistent for comparison. Full details of hyperparameters can be found in Table 6.

J.2 Knowledge Evaluation

We fine-tune Mistral-7B model obtained from Hugging face using 1 epoch on a 10k subset of the cleaned Alpaca dataset. We only target the the query and value matrices and follow a similar selection policy as the commonsense reasoning task for the remaining hyperparameters (see Table 7 for details).

J.3 Memory Efficiency

For this experiment, we apply the MESGC algorithm. First, we select a subset of 10k examples from the full commonsense dataset and fine-tune the LLaMA2-7B model, evaluating on all commonsense reasoning tasks. We used a batch size of 16 across all experiments and train for 1 epoch is used. The full results can be found in Table 8 and hyperparameters in Table 9.

J.4 Fine-tuning on Small Datasets

We first obtain a subset consisting of 2000 samples from the BoolQ dataset. We then create four partitions of data ranging in size from 500 to 2000 examples, in increments of 500. For this experiment, we are interested in comparing performance between our approach and baselines given equal optimizer state sizes. Thus, we set the total number of optimizer states to 8192, and perform fine-

Table 8: LLaMA2-7B results on commonsense reasoning for MESGC.

Method	**ARC-e**	**ARC-c**	**BoolQ**	**HellaSwag**	**OBQA**	**PIQA**	**SIQA**	**WinoGrande**	**Average**
CESGC	80.9	53.4	82.4	78.4	43.8	79.9	52.3	73.2	**68.0**
GaLore	80.2	52.2	79.0	78.4	43.0	80.5	51.6	74.0	67.4
LoRA	80.9	52.2	79.5	78.5	44.6	80.0	51.7	73.9	67.7

Table 9: Hyperparameters used for commonsense reasoning for MESGC.

Method	learning rate	rank r	num. chunks c	sparsity s	κ	α
MESGC	2e-5	-	256	256	8	2
GaLore	2e-5	1	-	-	-	2
LoRA	1e-4	1	-	-	-	-

Table 10: Hyperparameters used for fine-tuning BoolQ.

Method	learning rate	rank r	num. chunks c	sparsity s	κ	α
CESGC	2e-5	8	64	64	8	2
GaLore	2e-5	1	-	-	-	2
LoRA	1e-4	1	-	-	-	-

Table 11: Hyperparameters used for ablation study.

Study	Method	rank r	num. chunks c	sparsity s	κ
Chunks c	MESGC	-	256, 512, 1024, 2048, 4096	4096	7
Sparsity s	CESGC	32	64	64, 4096, 16384, 32768, 65536	7
Kappa κ	CESGC	32	64	1984	6, 7, 8

tuning with batch size 16 over 2 epochs using LLaMA2-7B based on the settings shown in Table 10.

J.5 Ablation Study

For chunks c and sparsity s studies, we fine-tuned on the LLaMA2-7B model fine-tuned on a subset of 30k examples using commonsense reasoning dataset. For the chunk size study, we performed the experiment based on our MESGC approach, while for sparsity, we used CESGC. Finally, different values of κ was tested on the full commonsense dataset using CESGC. The same batch size of 16, training epochs of 1, learning rate, $\eta = 2e^{-5}$ and alpha, $\alpha = 2$ is used for all three studies. Other hyperparameter details are shown in Table 11.