DIFFERENTIABLE ATTENTION SPARSITY VIA STRUCTURED D-GATING

Chris Kolb, Bernd Bischl, David Rügamer
Department of Statistics, LMU Munich
Munich Center for Machine Learning (MCML)

ABSTRACT

A core component of modern large language models is the attention mechanism, but its immense parameter count necessitates structured sparsity for resource-efficient optimization and inference. Traditional sparsity penalties, such as the group lasso, are non-smooth and thus incompatible with standard stochastic gradient descent methods. To address this, we propose a deep gating mechanism that reformulates the structured sparsity penalty into a fully differentiable optimization problem, allowing effective and principled norm-based group sparsification without requiring specialized non-smooth optimizers. Our theoretical analysis and empirical results demonstrate that this approach enables structured sparsity with simple stochastic gradient descent or variants while maintaining predictive performance.

1 INTRODUCTION

Modern large language model (LLM) applications typically include some form of attention layer (Vaswani et al., 2017). These architectures comprise up to hundreds of billions of weights (Baktash & Dawodi, 2023), posing the question of whether there exist sparser architectures with similar performance. One common strategy in deep learning is to define networks that are likely much more expressive than necessary and subsequently regularize or prune the network (Han et al., 2015; Hoefler et al., 2021). To effectively reduce the computation in attention-based architectures and LLMs, structured sparsity approaches are necessary. In contrast to pruning methods (He & Xiao, 2023), employing structured sparsity *penalties* during training or model adaptation typically requires specialized non-smooth optimization routines to make them effective. While frequently used in smaller applications, the non-smoothness of these methods will cause stochastic gradient descent (SGD) optimization routines to oscillate and usually fail to provide exact sparse solutions. This raises the question of whether structured sparsity penalties can be effectively used in modern LLMs.

Differentiable sparsity regularization To circumvent the non-differentiability of sparsity penalties while sticking to prominent and successful optimization techniques based on SGD, recent work (Ziyin & Wang, 2023; Kolb et al., 2025) therefore focuses on a differentiable reparametrization to obtain smooth surrogate penalties for $L_{2/D}$-norm penalties for $D \geq 2, D \in \mathbb{N}$. Approaches such as Kolb et al. (2025), however, only provide unstructured sparsity, i.e., not directly allowing for reducing the computational overhead. While some ideas for differentiable structured sparsity have been discussed in the literature (Ziyin & Wang, 2023; Kolb et al., 2023), neither their practical nor theoretical effectiveness in attention-based models have been explicitly addressed in previous literature. In this work, we show both theoretically and empirically that a product of gating parameters with appropriate differentiable L_2 regularization can induce effective non-smooth group penalization in attention-based network architectures and thus offer a viable solution for sparse learning in LLMs.

2 METHODOLOGY

In this work, we assume data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $n \in \mathbb{N}$ independent samples $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathbb{R}^c, c \in \mathbb{N}$. We denote $[p] \coloneqq \{1, \ldots, p\}$ and $\mathcal{P}([p])$ the power set of $[p]$. We further define $f(\mathbf{w}, \boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^c$ as the neural network of interest and stack all of its weights into one vector $\mathbf{w} \in \mathbb{R}^p$ with elements

$w_j, j \in [p]$. The goal is to minimize a loss function $\ell(\cdot,\cdot) : \mathbb{R}^c \times \mathbb{R}^c \to \mathbb{R}_0^+$. We denote the total loss including potential penalties as $\mathcal{L}_{\mathbf{w}}(\mathbf{w})$ and its minimizer $\hat{\mathbf{w}} := \arg\min_{\mathbf{w}} \mathcal{L}_{\mathbf{w}}(\mathbf{w})$. We further define $\tilde{D} := D - 1$.

2.1 Problem definition

In the following, without loss of generality, we assume that the weights are decomposed into $\mathbf{w} = (\mathbf{w}_{\neg A}, \mathbf{w}_A)$, where $\mathbf{w}_{\neg A}$ are weights that correspond to unpenalized (non-attention) structures in f. In order to regularize groups or certain structures in f, we focus on structured penalties $\Omega(\mathbf{w})$ of the form

$$\Omega_A(\mathbf{w}) = ||\mathbf{w}_A||_{2,2/D}^{2/D} := \sum_{h=1}^{H} ||\mathbf{w}_{A_h}||_2^{2/D}, H \in \mathbb{N}, \tag{1}$$

where $D \geq 2$ is an integer and $A \in \mathcal{P}([p])$ is a subset of the attention-weight indices in f to be penalized. This subset is further partitioned into H arbitrary groups $A = \{A_1, \ldots, A_H\}$, e.g., the attention heads. Note that other naturally or otherwise arising structures such as rows or columns in the weight matrices of attention layers can be arbitrarily designated as groups. For $D > 2$, the $L_{2,2/D}$ penalty in Equation (1) is a non-convex extension of the $L_{2,1}$ group lasso penalty (Yuan & Lin, 2006) and known to achieve superior sparsity tradeoffs (Hu et al., 2017). The resulting optimization problem we investigate is

$$\min_{\mathbf{w}\in\mathbb{R}^p} \sum_{i=1}^{n} \ell(y_i, f((\mathbf{w}_{\neg A}, \mathbf{w}_A), \boldsymbol{x}_i)) + \lambda ||\mathbf{w}_A||_{2,2/D}^{2/D}, \tag{2}$$

where $\lambda > 0$ is the tuning parameter controlling the structured sparsity.

The difficulty of optimizing Equation (2) lies in the non-smoothness of Equation (1). While there have been proposals to combine (simpler forms of) the structured sparsity penalty above and neural networks (see, e.g., Wen et al., 2016; Scardapane et al., 2017; Zhang et al., 2019; Wang et al., 2020; Bui et al., 2021b), these approaches either naïvely use vanilla SGD on the non-smooth objective or resort to specific non-smooth optimization routines to adequately account for the non-differentiability of the structure-inducing regularizer. In the following, we suggest an optimization transfer that allows using SGD-type optimizers such as Adam without any modification while provably minimizing Equation (2).

2.2 Regularized gating for structured sparsity

Definition 2.1 (D-Gating)**.** Let $\boldsymbol{\Gamma} = \{\boldsymbol{\gamma}_d\}_{d=1}^{\tilde{D}} \in \mathbb{R}^{H\cdot\tilde{D}}$ be a collection of $\tilde{D}$ gating parameters $\boldsymbol{\gamma}_d \in \mathbb{R}^H$ each with elements $\gamma_{d,h}$, and $(\boldsymbol{v}_{A_1}, \ldots, \boldsymbol{v}_{A_H}) = \boldsymbol{v}_A \in \mathbb{R}^{|A|}$ be a vector of the same size and grouping structure as $\mathbf{w}_A$. Further, define the elementwise Hadamard product of gating parameters as $\boldsymbol{\gamma}^{\odot} := \bigodot_{d\in[\tilde{D}]} \boldsymbol{\gamma}_d$. We then call

$$\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot} := \left(\mathrm{v}_{A_h,j} \cdot \prod_{d=1}^{\tilde{D}} \gamma_{h,d} \right)_{h\in[H], j\in A_h} \tag{3}$$

a *D-Gating Operation* and call $\mathbf{w}_A$ *D-gated* when replacing $\mathbf{w}_A$ in f with $\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}$.

Simply put, D-gating $\mathbf{w}_A$ means that we replace every weight group $\mathbf{w}_{A_h}$ by a weight $\boldsymbol{v}_{A_h}$ of the same size, multiplied by $\tilde{D}$ scalar factors $\gamma_{h,d}$ — the gating variables.

Equivalent differentiable formulation The reason we split the original weight $\mathbf{w}_A = \boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}$ into D parts is that it allows the use of an optimization transfer trick by replacing the original penalty on $\mathbf{w}_A$ in Equation (1) with a smooth L_2 or weight-decay penalty

$$\boldsymbol{\Phi}_A(\boldsymbol{v}_A, \boldsymbol{\Gamma}) = D^{-1} \left(\|\boldsymbol{v}_A\|_2^2 + \sum_{d=1}^{\tilde{D}} \|\boldsymbol{\gamma}_d\|_2^2 \right), \tag{4}$$

and obtain the following result:

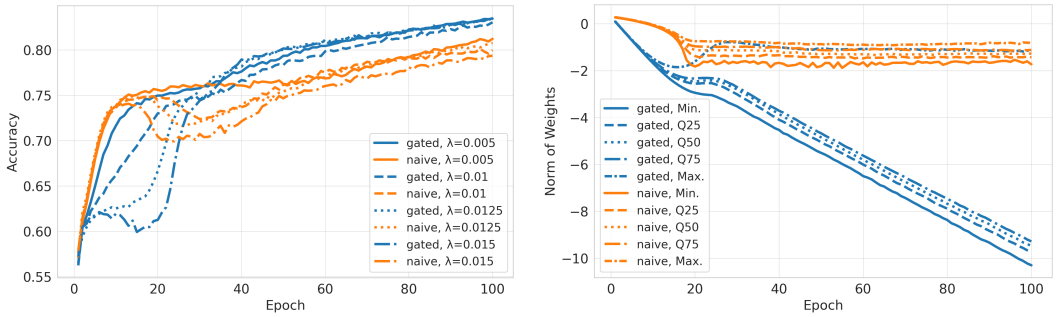

Figure 1: Left: Test accuracy of different methods (colors) and λ values (line type) vs. epochs. Right: Quantiles of the weight norms (log scale) across the attention heads for $\lambda = 0.015$ ("Max", e.g., refers to the head with the largest norm among all groups and "Q25" is the 0.25-quantile of the norms). "naïve" refers to direct optimization of the $L_{2,1}$ penalty with SGD; "gated" refers to optimization of Equation (5) with $D = 2$.

Theorem 2.2. *The optimization problems in Equation* (2) *and*

$$\min_{\mathbf{w}_{\neg A}, \boldsymbol{v}_A, \boldsymbol{\Gamma}} \sum_{i=1}^{n} \ell\left(y_i, f\left((\mathbf{w}_{\neg A}, \boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}), \boldsymbol{x}_i\right)\right) + \lambda \boldsymbol{\Phi}_A(\boldsymbol{v}_A, \boldsymbol{\Gamma}) \tag{5}$$

are equivalent in the sense of having the same minimum, despite (5) being fully differentiable. Secondly, for every local minimizer $(\hat{\mathbf{w}}_{\neg A}, \hat{\mathbf{w}}_A)$ *of Equation* (2)*, there is a local minimizer* $(\hat{\mathbf{w}}_{\neg A}, \hat{\boldsymbol{v}}_A, \hat{\boldsymbol{\Gamma}})$ *of Equation* (5) *such that* $\hat{\mathbf{w}}_A = \hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}$ *and both minima coincide.*

This means we can transfer our original sparsity-regularized problem into the smooth objective in Equation (5), optimize the over the parameters $(\mathbf{w}_{\neg A}, \boldsymbol{v}_A, \boldsymbol{\Gamma})$ using conventional SGD-based methods, and obtain the collapsed solution to the original non-smooth problem in Equation (2) as $\hat{\mathbf{w}}_A = \hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}$. Since Theorem 2.2 does not impose any restrictions on the separation of penalized $(\mathbf{w}_A)$ and unpenalized groups $(\mathbf{w}_{\neg A})$ or the specific loss function, our result implies that we can arbitrarily impose group sparse $L_{2,2/D}$ regularization on any structure in an attention layer (or in fact any other layer), and effectively optimize the objective with a standard SGD-routine.

3 Numerical Experiments

To demonstrate our findings in attention-based models, we compare our proposed approach (D-gating with $D = 2$) against a naïve optimization of the $L_{2,1}$ group lasso penalty, a widely-used regularizer for structured sparsity that is in practice often optimized using SGD without adapting the optimization routine to its non-smoothness (see, e.g., Wen et al., 2016; Scardapane et al., 2017; Liu et al., 2017). For comparisons, we use the IMDb Movie Reviews dataset (Lhoest et al., 2021) with 50,000 movie reviews and binary sentiment outcomes. For this, each review is tokenized using BERT's subword tokenizer (bert-base-uncased; Devlin et al., 2019), truncated to a maximum length of 128 tokens, and padded where necessary. The two regularization approaches (naïve and ours) are compared w.r.t. their prediction performance on a test data set as well as the induced sparsity.

Architecture and regularization To not confound the sparsity effects with other training dynamics, we use a simple transformer-style network architecture using an embedding layer with embedding size 768, followed by a multi-head attention layer with 12 heads, a pooling layer, and a fully-connected layer with sigmoid activation. The loss is chosen to be the binary cross-entropy and optimization is done with SGD using a learning rate of 0.01. For the sparsity penalty, we define the groups A as the parameters associated with the different model heads $h \in [H]$. Hence, for increased regularization, we expect the regularization to remove whole heads of the attention layer. For our gating approach, we initialize all gating parameters in $\boldsymbol{\Gamma}$ with ones while using a standard initialization scheme for all other parameters. We run our experiment for different penalty strengths $\lambda \in \{0.005, 0.01, 0.0125, 0.015\}$ and evaluate sparsity and performance across these values.

Results Figure 1 visualizes the result for both performance and sparsity across training epochs. While the naïve $L_{2,1}$-regularization shows better performance in the beginning, our approach even-

tually overtakes the vanilla implementation and results in slightly better performance. The quite astonishing finding, however, is how the model can achieve this performance. As becomes apparent on the sparsity plots, the naïve approach fails to shrink the regularized norms below a certain large threshold and does not induce actual sparsity. This behavior can be expected due to the incompatibility of differentiable SGD and the non-smooth sparsity penalty. However, our equivalent differentiable sparsity approach constantly shrinks weights until values become numerically zero. This can be observed for at least the bottom 75% of heads, while the non-sparse head with max norm maintains a large norm throughout training. Upon closer examination, it turns out that our approach has reduced all heads except one.

In summary, the results clearly show that our model can achieve similar performance as naïve regularization while resulting in a structurally sparse model that requires only a fraction of the existing weights.

4 Related Literature

Structured sparsity in deep learning promises significant computational benefits by removing entire components, such as neurons, filters, or here, attention heads. Although this generally produces inferior sparsity–accuracy tradeoffs compared to unstructured sparsity, the resulting efficiency gains are considerably higher. A widely adopted method for structured sparsity is magnitude-based pruning, where group importance is approximated by L_2 or Frobenius norm of weight components, proposed, e.g., in Li et al. (2022). For a survey over different methods, see He & Xiao (2023); Zhu et al. (2024). Many studies have applied SGD to naïvely optimize L_1 or structured $L_{2,1}$ penalized objectives (Wen et al., 2016; Scardapane et al., 2017), or applied non-convex penalties using a specialized optimization routine to handle the non-smooth and non-convex penalties (Bui et al., 2021b;a), complicating a broader practical application.

Particularly related to our D-gating are works proposing masking or gating to promote sparsity. One branch of methods implements binary stochastic gates or continuously approximated versions thereof, including, e.g., ProbMask (Zhou et al., 2021), and the stochastic L_0 type approach proposed by Louizos et al. (2018). Additionally, L_1 penalized gates have been proposed to induce sparse learning (Liu et al., 2017; Yang et al., 2019). Other approaches close to our work are differentiable pruning methods, where soft masks or gates are trained jointly with network weights to simultaneously learn weights and importances. Variations of this idea are based on using the parameters or groups themselves as masks to approximate importance by applying simple transformations of the original network parameters (Yasuda et al., 2024; Cho et al., 2024).

Pruning in large-scale models such as LLMs and other attention-based architectures additionally poses a set of unique challenges due to their size and complexity (Zhu et al., 2024). In these settings, compression is typically applied post-hoc (Ma et al., 2023; Frantar & Alistarh, 2023; He et al., 2024), and employing alternative strategies, such as low-rank approximations (Hu et al., 2021), is more attractive in many use cases.

Finally, sparsity-inducing parameterizations have been investigated in several works, starting with Hoff (2017) and recently investigated by Schwarz et al. (2021); Ziyin & Wang (2023); Yasuda et al. (2024); Kolb et al. (2025). None of these works, however, focused on differentiable structured sparsity with potentially non-convex penalizers in neural networks, and attention layers in particular.

5 Discussion

In this paper, we derived a gating mechanism that induces a smooth variational form of the original $L_{2,2/D}$-penalty. This gating operator overcomes differentiability problems with the optimization of sparsity-inducing regularization terms in neural networks and, in particular, attention layers. Our approach thereby allows for the optimization of LLM architectures and for finding sparse structures therein while not requiring adapting the optimizer for training.

While our experiments confirm the effectiveness of our approach, it remains to show how the incorporation of these gating operators affects LLMs and if, e.g., these can also be used for fine-tuning such as Low-Rank Adaptation (Hu et al., 2021) with structured sparsification.

REFERENCES

Jawid Ahmad Baktash and Mursal Dawodi. Gpt-4: A review on advancements and opportunities in natural language processing. *arXiv preprint arXiv:2305.03195*, 2023.

Kevin Bui, Fredrick Park, Shuai Zhang, Yingyong Qi, and Jack Xin. Improving network slimming with nonconvex regularization. *IEEE Access*, 9:115292–115314, 2021a.

Kevin Bui, Fredrick Park, Shuai Zhang, Yingyong Qi, and Jack Xin. Structured sparsity of convolutional neural networks via nonconvex sparse group regularization. *Frontiers in applied mathematics and statistics*, 6:529564, 2021b.

Minsik Cho, Saurabh Adya, and Devang Naik. Pdp: parameter-free differentiable pruning is all you need. *Advances in Neural Information Processing Systems*, 36, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pp. 10323–10337. PMLR, 2023.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. What matters in transformers? not all attention is needed. *arXiv preprint arXiv:2406.15786*, 2024.

Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

Peter D Hoff. Lasso, fractional norm and structured sparse estimation using a hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Yaohua Hu, Chong Li, Kaiwen Meng, Jing Qin, and Xiaoqi Yang. Group sparse optimization via lp, q regularization. *Journal of Machine Learning Research*, 18(30):1–52, 2017.

Chris Kolb, Christian L. Müller, Bernd Bischl, and David Rügamer. Smoothing the edges: A general framework for smooth optimization in sparse regularization using hadamard overparametrization. 2023.

Chris Kolb, Tobias Weber, Bernd Bischl, and David Rügamer. Deep weight factorization: Sparse learning through the lens of artificial symmetries. In *The Thirteenth International Conference on Learning Representations*, 2025.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, 2021.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2022.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In *Proceedings of the IEEE international conference on computer vision*, pp. 2736–2744, 2017.

Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. In *International Conference on Learning Representations*, 2018.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-pruner: On the structural pruning of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=J8Ajf9WfXP.

Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

Jonathan Schwarz, Siddhant Jayakumar, Razvan Pascanu, Peter E Latham, and Yee Teh. Powerpropagation: A sparsity inducing weight reparameterisation. *Advances in neural information processing systems*, 34:28889–28903, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Jian Wang, Huaqing Zhang, Junze Wang, Yifei Pu, and Nikhil R Pal. Feature selection using a neural network with group lasso regularization and controlled redundancy. *IEEE transactions on neural networks and learning systems*, 32(3):1110–1123, 2020.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Chen Yang, Zhenghong Yang, Abdul Mateen Khattak, Liu Yang, Wenxin Zhang, Wanlin Gao, and Minjuan Wang. Structured pruning of convolutional neural networks via l1 regularization. *IEEE Access*, 7:106385–106394, 2019.

Taisuke Yasuda, Kyriakos Axiotis, Gang Fu, MohammadHossein Bateni, and Vahab Mirrokni. Sequentialattention++ for block sparsification: Differentiable pruning meets combinatorial optimization. *arXiv preprint arXiv:2402.17902*, 2024.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Huaqing Zhang, Jian Wang, Zhanquan Sun, Jacek M Zurada, and Nikhil R Pal. Feature selection for neural networks using group lasso regularization. *IEEE Transactions on Knowledge and Data Engineering*, 32(4):659–673, 2019.

Xiao Zhou, Weizhong Zhang, Hang Xu, and Tong Zhang. Effective sparsification of neural networks with global sparsity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3599–3608, 2021.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12: 1556–1577, 2024.

Liu Ziyin and Zihao Wang. spred: Solving l1 penalty with sgd. In *International Conference on Machine Learning*, pp. 43407–43422. PMLR, 2023.

A PROOF OF THEOREM 2.2

Proof. For clarity, we will denote the unregularized objective $\sum_{i=1}^{n} \ell(y_i, f(\cdot, \boldsymbol{x}_i))$ as $\mathcal{L}_0(\cdot)$. Further, since the presence of $\mathbf{w}_{\neg A}$ in $\mathbf{w}$ does not affect the proof, we can assume $A = [p]$ without loss of generality and omit $\mathbf{w}_{\neg A}$. First note that the D-gating operation

$$\mathcal{O} : \mathbb{R}^{|A|} \times \prod_{d=1}^{\tilde{D}} \mathbb{R}^{H} \to \mathbb{R}^{|A|}, (\boldsymbol{v}_A, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{\tilde{D}}) = (\boldsymbol{v}_A, \boldsymbol{\Gamma}) \mapsto \boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot} \tag{6}$$

is surjective and continuously differentiable. Due to the multiplicative nature of the gating operation, for any $\mathbf{w}_A \in \mathbb{R}^{|A|}$, there exist infinitely many $(\boldsymbol{v}_A, \boldsymbol{\Gamma})$ such that $\mathcal{O}(\boldsymbol{v}_A, \boldsymbol{\Gamma}) = \mathbf{w}_A$. Because $\mathcal{L}_0(\cdot)$ only depends on the gated $\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot} = \mathbf{w}_A$, every minimizer $(\hat{\boldsymbol{v}}_A, \hat{\boldsymbol{\Gamma}})$ of the gated objective must also minimize $\boldsymbol{\Phi}(\boldsymbol{v}_A, \boldsymbol{\Gamma})$ over the feasible set $F_{\hat{\mathbf{w}}_A} = \{(\boldsymbol{v}_A, \boldsymbol{\Gamma}) : \boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot} = \hat{\mathbf{w}}_A\}$, where $\hat{\mathbf{w}}_A := \hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}$. Otherwise, the objective can always be strictly decreased by choosing $(\boldsymbol{v}'_A, \boldsymbol{\Gamma}') \in F_{\hat{\mathbf{w}}_A}$ so that $\boldsymbol{\Phi}(\boldsymbol{v}'_A, \boldsymbol{\Gamma}') < \boldsymbol{\Phi}(\hat{\boldsymbol{v}}_A, \hat{\boldsymbol{\Gamma}})$ while $\mathcal{L}_0(\boldsymbol{v}'_A \triangleright \boldsymbol{\gamma}'^{\odot}) = \mathcal{L}_0(\hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot})$. The following result provides the necessary and sufficient conditions for this constrained minimization problem and the resulting minimum value:

Lemma A.1. *(Smooth Variational Form of Ω_A) The minimum of $\boldsymbol{\Phi}(\boldsymbol{v}_A, \boldsymbol{\Gamma})$ over the feasible set $F_{\mathbf{w}_A}$ of gating parameters that leave $\mathcal{L}_0$ unchanged is given by*

$$\min_{(\boldsymbol{v}_A, \boldsymbol{\Gamma}) : \boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot} = \mathbf{w}_A} \boldsymbol{\Phi}(\boldsymbol{v}_A, \boldsymbol{\Gamma}) = \Omega_A(\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}) = \Omega_A(\mathbf{w}_A) \quad \forall\, \mathbf{w}_A \in \mathbb{R}^{|A|}.$$

Furthermore, the minimum is attained if and only if $\|\boldsymbol{v}_{A_h}\|_2 = |\gamma_{h,1}| = \ldots = |\gamma_{h,\tilde{D}}| \,\forall\, h \in [H]$.

Proof. Group-wise separating the penalty $\boldsymbol{\Phi}$ and applying the inequality of arithmetic and geometric means (AM-GM), we obtain:

$$\boldsymbol{\Phi}_A(\boldsymbol{v}_A, \boldsymbol{\Gamma}) = \sum_{h=1}^{H} D^{-1} \left(\|\boldsymbol{v}_{A_h}\|_2^2 + \sum_{d=1}^{\tilde{D}} \gamma_{h,d}^2 \right) \geq \sum_{h=1}^{H} \left(\|\boldsymbol{v}_{A_h}\|_2^2 \cdot \prod_{d=1}^{\tilde{D}} \gamma_{h,d}^2 \right)^{1/D} \tag{7}$$

$$= \sum_{h=1}^{H} \left(\left(\|\boldsymbol{v}_{A_h}\|_2 \cdot \prod_{d=1}^{\tilde{D}} \gamma_{h,d} \right)^2 \right)^{1/D} \tag{8}$$

$$= \sum_{h=1}^{H} \left(\left| \|\boldsymbol{v}_{A_h}\|_2 \cdot \textstyle\prod_{d=1}^{\tilde{D}} \gamma_{h,d} \right| \right)^{2/D}$$

$$= \sum_{h=1}^{H} \left(\|\boldsymbol{v}_{A_h}\|_2 \cdot \left| \textstyle\prod_{d=1}^{\tilde{D}} \gamma_{h,d} \right| \right)^{2/D}$$

$$= \sum_{h=1}^{H} (\| \underbrace{\boldsymbol{v}_{A_h} \cdot \textstyle\prod_{d=1}^{\tilde{D}} \gamma_{h,d}}_{=\mathbf{w}_{A_h}} \|_2)^{2/D}$$

$$= \sum_{h=1}^{H} (\|\mathbf{w}_{A_h}\|_2)^{2/D} =: \|\mathbf{w}_A\|_{2,2/D}^{2/D}$$

$$= \Omega_A(\mathbf{w}_A)$$

Finally, the AM-GM inequality holds with equality if and only if $\|\boldsymbol{v}_{A_h}\|_2^2 = \gamma_{h,1}^2 = \ldots = \gamma_{h,\tilde{D}}^2$ for all $h \in [H]$. □

Lemma A.1 shows that at every potential minimizer of the gated objective, the magnitudes of the gating factors $|\gamma_{h,d}|$ must equal $\|\boldsymbol{v}_{A_h}\|_2$ for all $d \in [\tilde{D}]$ and $h \in [H]$. Given this balancedness

condition, the smooth penalty $\mathbf{\Phi}_A(\boldsymbol{v}_A, \mathbf{\Gamma})$ reduces to its minimal value $\Omega_A(\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}) = \Omega_A(\mathbf{w}_A)$. To show that

$$\min_{\boldsymbol{v}_A, \mathbf{\Gamma}} \mathcal{L}_0(\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}) + \lambda \mathbf{\Phi}_A(\boldsymbol{v}_A, \mathbf{\Gamma}) = \min_{\mathbf{w}_A \in \mathbb{R}^{|A|}} \mathcal{L}_0(\mathbf{w}_A) + \lambda \Omega_A(\mathbf{w}_A)\,,$$

we use that $\mathcal{L}_0$ is constant over all possible gating parameters $(\boldsymbol{v}_A, \mathbf{\Gamma}) \in F_{\mathbf{w}_A}$ mapping to some $\mathbf{w}_A$:

$$\min_{\boldsymbol{v}_A, \mathbf{\Gamma}} \left\{ \mathcal{L}_0(\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}) + \lambda \mathbf{\Phi}_A(\boldsymbol{v}_A, \mathbf{\Gamma}) \right\} = \min_{\mathbf{w}_A} \{ \mathcal{L}_0(\mathbf{w}_A) + \lambda \min_{\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot} = \mathbf{w}_A} \{ \mathbf{\Phi}_A(\boldsymbol{v}_A, \mathbf{\Gamma}) \} \}\,. \tag{9}$$

Inserting the result of Lemma A.1 for the constrained inner minimum, we finally obtain the result:

$$\min_{\boldsymbol{v}_A, \mathbf{\Gamma}} \mathcal{L}_0(\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}) + \lambda \mathbf{\Phi}_A(\boldsymbol{v}_A, \mathbf{\Gamma}) = \min_{\mathbf{w}_A} \mathcal{L}_0(\mathbf{w}_A) + \lambda \Omega_A(\mathbf{w}_A)\,. \tag{10}$$

This shows that both objectives have identical globally optimal values.

In the second step, we prove that for every local minimizer $\hat{\mathbf{w}}_A$ of $\mathcal{L}_{\mathbf{w}}(\mathbf{w}_A) := \mathcal{L}_0(\mathbf{w}_A) + \lambda \Omega_A(\mathbf{w}_A)$, there is a corresponding local minimizer $(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}})$ of the gated objective $\mathcal{L}_0(\hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}) + \lambda \mathbf{\Phi}_A(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}})$, related as $\hat{\mathbf{w}}_A = \hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}$.
Assume $\hat{\mathbf{w}}_A$ is a local minimizer of the non-smooth penalized objective $\mathcal{L}_{\mathbf{w}}(\mathbf{w}_A)$, then $\exists \varepsilon_0 > 0$: $\mathcal{L}_{\mathbf{w}}(\hat{\mathbf{w}}_A) \leq \mathcal{L}_{\mathbf{w}}(\mathbf{w}'_A) \, \forall \, \mathbf{w}'_A \in \mathcal{B}(\hat{\mathbf{w}}_A, \varepsilon_0)$, where $\mathcal{B}$ is an ε_0-ball around $\hat{\mathbf{w}}_A \in \mathbb{R}^{|A|}$. Due to the multiplicative nature of the gating operation $\mathcal{O}$, we can pick balanced gating parameters $(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}})$ (sometimes abbreviated as $\hat{\boldsymbol{\psi}}$ from now on) so that $\mathcal{O}(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}}) = \hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}^{\odot}} = \hat{\mathbf{w}}_A$. As the gating variables are balanced, applying Lemma A.1 shows that $\mathcal{L}_{\boldsymbol{\psi}}(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}}) := \mathcal{L}_0(\hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}) + \lambda \mathbf{\Phi}_A(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}})$ reduces to

$$\mathcal{L}_{\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}) = \mathcal{L}_{\boldsymbol{\psi}}(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}}) := \mathcal{L}_0(\hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}) + \lambda \Omega_A(\hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}) = \mathcal{L}_0(\hat{\mathbf{w}}_A) + \lambda \Omega_A(\hat{\mathbf{w}}_A) = \mathcal{L}_{\mathbf{w}}(\hat{\mathbf{w}}_A)\,. \tag{11}$$

Using again Lemma A.1, we can further relate $\mathcal{L}_{\mathbf{w}}$ and $\mathcal{L}_{\boldsymbol{\psi}}$ as follows:

$$\mathcal{L}_{\boldsymbol{\psi}}(\boldsymbol{v}_A, \mathbf{\Gamma}) = \mathcal{L}_{\mathbf{w}}(\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot}) + \lambda (\underbrace{\mathbf{\Phi}(\boldsymbol{v}_A, \mathbf{\Gamma}) - \Omega_A(\boldsymbol{v}_A \triangleright \boldsymbol{\gamma}^{\odot})}_{:= M(\boldsymbol{\psi}) \geq 0}) \tag{12}$$

where $M(\boldsymbol{\psi})$ quantifies the non-negative distance of $\mathbf{\Phi}(\boldsymbol{v}_A, \mathbf{\Gamma})$ to its minimum value.

By the continuity of $\mathcal{O}$, $\exists \delta_0 : \mathcal{O}(\mathcal{B}(\hat{\boldsymbol{\psi}}, \delta_0)) \subseteq \mathcal{B}(\mathcal{O}(\hat{\boldsymbol{\psi}}), \varepsilon_0) = \mathcal{B}(\hat{\mathbf{w}}_A, \varepsilon_0)$, implying that all $\boldsymbol{\psi}' \in \mathcal{B}(\hat{\boldsymbol{\psi}}, \delta_0)$ map to some $\mathcal{O}(\boldsymbol{\psi}') = \mathbf{w}'_A \in \mathcal{B}(\hat{\mathbf{w}}_A, \varepsilon_0)$. Then it holds

$$\forall \boldsymbol{\psi}' \in \mathcal{B}(\hat{\boldsymbol{\psi}}, \delta_0) : \mathcal{L}_{\boldsymbol{\psi}}(\hat{\boldsymbol{\psi}}) \underbrace{=}_{\text{Eq.11}} \mathcal{L}_{\mathbf{w}}(\hat{\mathbf{w}}_A) \leq \mathcal{L}_{\mathbf{w}}(\underbrace{\mathcal{O}(\boldsymbol{\psi}')}_{\mathbf{w}'_A}) \leq \mathcal{L}_{\mathbf{w}}(\mathcal{O}(\boldsymbol{\psi}')) + \underbrace{\lambda M(\boldsymbol{\psi}')}_{\geq 0} \underbrace{=}_{\text{Eq.12}} \mathcal{L}_{\boldsymbol{\psi}}(\boldsymbol{\psi}'),$$

where we have the first inequality because $\hat{\mathbf{w}}_A$ is a local minimizer of $\mathcal{L}_{\mathbf{w}}$. This chain of inequalities shows that $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}}) \in \arg\min_{\boldsymbol{\psi}} \mathcal{L}_{\boldsymbol{\psi}}(\boldsymbol{\psi})$. As $\hat{\mathbf{w}}_A$ was arbitrary, it is shown that for all local minimizers $\hat{\mathbf{w}}_A$ of the non-smooth original objective $\mathcal{L}_{\mathbf{w}}(\hat{\mathbf{w}}_A) = \mathcal{L}_0(\mathbf{w}_A) + \lambda \Omega_A(\mathbf{w}_A)$, there are corresponding local minimizers $(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}})$ of $\mathcal{L}_{\boldsymbol{\psi}}(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}}) = \mathcal{L}_0(\hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot}) + \lambda \mathbf{\Phi}_A(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}})$, so that $\hat{\boldsymbol{v}}_A \triangleright \hat{\boldsymbol{\gamma}}^{\odot} = \hat{\mathbf{w}}_A$ and $\mathcal{L}_{\boldsymbol{\psi}}(\hat{\boldsymbol{v}}_A, \hat{\mathbf{\Gamma}}) = \mathcal{L}_{\mathbf{w}}(\hat{\mathbf{w}}_A)$.

□

B FURTHER RESULTS

The following figures Figures 2 to 4 show the resulting weight norms for the other tested λ values.

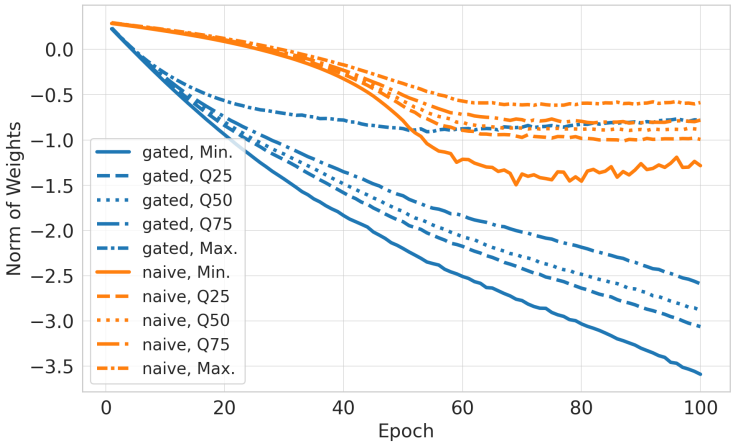

Figure 2: Quantiles of the weight norms (log scale) across attention heads for $\lambda = 0.005$.

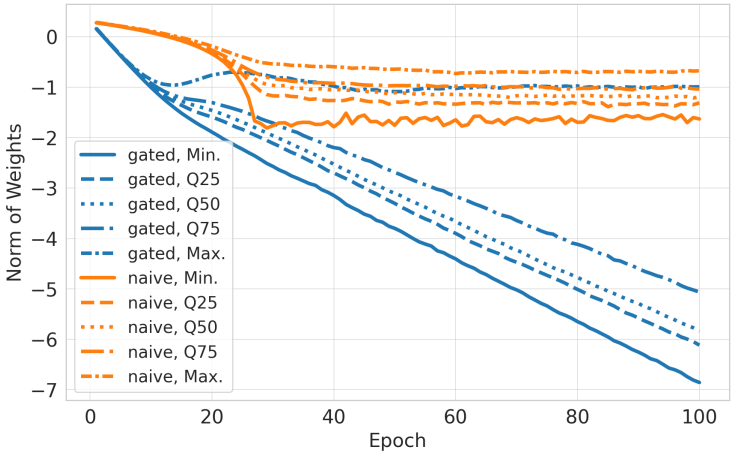

Figure 3: Quantiles of the weight norms (log scale) across the attention heads for $\lambda = 0.01$.

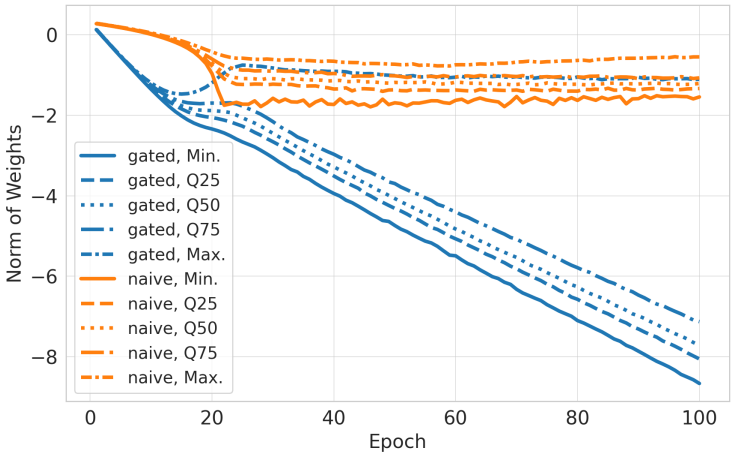

Figure 4: Quantiles of the weight norms (log scale) across the attention heads for $\lambda = 0.0125$.