

# DIFFERENTIABLE ATTENTION SPARSITY VIA STRUCTURED $D$ -GATING

**Anonymous authors**

Paper under double-blind review

## 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  $\{(x_i, y_i)\}_{i=1}^n$  from  $n \in \mathbb{N}$  independent samples  $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}^c, c \in \mathbb{N}$ . We denote  $[p] := \{1, \dots, p\}$  and  $\mathcal{P}([p])$  the power set of  $[p]$ . We further define  $f(\mathbf{w}, \mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}^c$  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  $\ell(\cdot, \cdot) : \mathbb{R}^c \times \mathbb{R}^c \rightarrow \mathbb{R}_0^+$ . 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  $\mathbf{w} = (\mathbf{w}_{-A}, \mathbf{w}_A)$ , where  $\mathbf{w}_{-A}$  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

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

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  $A = \{A_1, \dots, A_H\}$ , 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

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

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

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

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

Simply put, *D-gating*  $\mathbf{w}_A$  means that we replace every weight group  $\mathbf{w}_{A_h}$  by a weight  $\mathbf{v}_{A_h}$  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  $\mathbf{w}_A = \mathbf{v}_A \triangleright \gamma^\odot$  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

$$\Phi_A(\mathbf{v}_A, \Gamma) = D^{-1} \left( \|\mathbf{v}_A\|_2^2 + \sum_{d=1}^{\tilde{D}} \|\gamma_d\|_2^2 \right), \quad (4)$$

and obtain the following result:

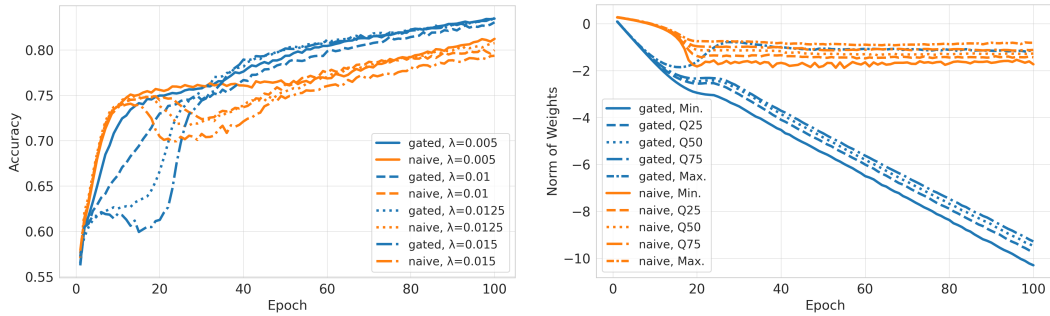


Figure 1: Left: Test accuracy of different methods (colors) and  $\lambda$  values (line type) vs. epochs. Right: Quantiles of the weight norms (log scale) across the attention heads for  $\lambda = 0.015$  (“Max”, e.g., refers to the head with the largest norm among all groups and “Q25” is the 0.25-quantile of the norms). “naive” 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*

$$\min_{\mathbf{w}_{-A}, \mathbf{v}_A, \Gamma} \sum_{i=1}^n \ell(y_i, f((\mathbf{w}_{-A}, \mathbf{v}_A \triangleright \gamma^\odot), \mathbf{x}_i)) + \lambda \Phi_A(\mathbf{v}_A, \Gamma) \quad (5)$$

are equivalent in the sense of having the same minimum, despite (5) being fully differentiable. Secondly, for every local minimizer  $(\hat{\mathbf{w}}_{-A}, \hat{\mathbf{w}}_A)$  of Equation (2), there is a local minimizer  $(\hat{\mathbf{w}}_{-A}, \hat{\mathbf{v}}_A, \hat{\Gamma})$  of Equation (5) such that  $\hat{\mathbf{w}}_A = \hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\odot$  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  $(\mathbf{w}_{-A}, \mathbf{v}_A, \Gamma)$  using conventional SGD-based methods, and obtain the collapsed solution to the original non-smooth problem in Equation (2) as  $\hat{\mathbf{w}}_A = \hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\odot$ . Since Theorem 2.2 does not impose any restrictions on the separation of penalized  $(\mathbf{w}_A)$  and unpenalized groups  $(\mathbf{w}_{-A})$  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  $\Gamma$  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-

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

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

## 174 4 RELATED LITERATURE

175  
176  
177  
178 Structured sparsity in deep learning promises significant computational benefits by removing entire  
179 components, such as neurons, filters, or here, attention heads. Although this generally produces infe-  
180 rior sparsity–accuracy tradeoffs compared to unstructured sparsity, the resulting efficiency gains are  
181 considerably higher. A widely adopted method for structured sparsity is magnitude-based pruning,  
182 where group importance is approximated by  $L_2$  or Frobenius norm of weight components, proposed,  
183 e.g., in Li et al. (2022). For a survey over different methods, see He & Xiao (2023); Zhu et al. (2024).  
184 Many studies have applied SGD to naïvely optimize  $L_1$  or structured  $L_{2,1}$  penalized objectives (Wen  
185 et al., 2016; Scardapane et al., 2017), or applied non-convex penalties using a specialized optimiza-  
186 tion routine to handle the non-smooth and non-convex penalties (Bui et al., 2021b;a), complicating  
187 a broader practical application.

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

196 Pruning in large-scale models such as LLMs and other attention-based architectures additionally  
197 poses a set of unique challenges due to their size and complexity (Zhu et al., 2024). In these set-  
198 tings, compression is typically applied post-hoc (Ma et al., 2023; Frantar & Alistarh, 2023; He et al.,  
199 2024), and employing alternative strategies, such as low-rank approximations (Hu et al., 2021), is  
200 more attractive in many use cases.

201 Finally, sparsity-inducing parameterizations have been investigated in several works, starting with  
202 Hoff (2017) and recently investigated by Schwarz et al. (2021); Ziyin & Wang (2023); Yasuda et al.  
203 (2024); Kolb et al. (2025). None of these works, however, focused on differentiable structured spar-  
204 sity with potentially non-convex penalizers in neural networks, and attention layers in particular.

## 205 5 DISCUSSION

206  
207  
208  
209 In this paper, we derived a gating mechanism that induces a smooth variational form of the original  
210  $L_{2,2/D}$ -penalty. This gating operator overcomes differentiability problems with the optimization of  
211 sparsity-inducing regularization terms in neural networks and, in particular, attention layers. Our  
212 approach thereby allows for the optimization of LLM architectures and for finding sparse structures  
213 therein while not requiring adapting the optimizer for training.

214 While our experiments confirm the effectiveness of our approach, it remains to show how the incor-  
215 poration 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

- 216  
217  
218 Jawid Ahmad Baktash and Mursal Dawodi. Gpt-4: A review on advancements and opportunities in  
219 natural language processing. *arXiv preprint arXiv:2305.03195*, 2023.
- 220  
221 Kevin Bui, Fredrick Park, Shuai Zhang, Yingyong Qi, and Jack Xin. Improving network slimming  
222 with nonconvex regularization. *IEEE Access*, 9:115292–115314, 2021a.
- 223  
224 Kevin Bui, Fredrick Park, Shuai Zhang, Yingyong Qi, and Jack Xin. Structured sparsity of convo-  
225 lutional neural networks via nonconvex sparse group regularization. *Frontiers in applied mathe-  
226 matics and statistics*, 6:529564, 2021b.
- 227  
228 Minsik Cho, Saurabh Adya, and Devang Naik. Pdp: parameter-free differentiable pruning is all you  
229 need. *Advances in Neural Information Processing Systems*, 36, 2024.
- 230  
231 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of  
232 deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and  
233 Tamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of  
234 the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long  
235 and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Compu-  
236 tational Linguistics.
- 237  
238 Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in  
239 one-shot. In *International Conference on Machine Learning*, pp. 10323–10337. PMLR, 2023.
- 240  
241 Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for  
242 efficient neural network. *Advances in neural information processing systems*, 28, 2015.
- 243  
244 Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. What matters in transformers? not all attention  
245 is needed. *arXiv preprint arXiv:2406.15786*, 2024.
- 246  
247 Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey.  
248 *IEEE transactions on pattern analysis and machine intelligence*, 2023.
- 249  
250 Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep  
251 learning: Pruning and growth for efficient inference and training in neural networks. *Journal of  
252 Machine Learning Research*, 22(241):1–124, 2021.
- 253  
254 Peter D Hoff. Lasso, fractional norm and structured sparse estimation using a hadamard product  
255 parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.
- 256  
257 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang,  
258 and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint  
259 arXiv:2106.09685*, 2021.
- 260  
261 Yaohua Hu, Chong Li, Kaiwen Meng, Jing Qin, and Xiaoyi Yang. Group sparse optimization via  
262 lp, q regularization. *Journal of Machine Learning Research*, 18(30):1–52, 2017.
- 263  
264 Chris Kolb, Christian L. Müller, Bernd Bischl, and David Rügamer. Smoothing the edges: A general  
265 framework for smooth optimization in sparse regularization using hadamard overparametrization.  
266 2023.
- 267  
268 Chris Kolb, Tobias Weber, Bernd Bischl, and David Rügamer. Deep weight factorization: Sparse  
269 learning through the lens of artificial symmetries. In *The Thirteenth International Conference on  
Learning Representations*, 2025.
- 270  
271 Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen,  
272 Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario  
273 Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Can-  
274 wen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément  
275 Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer,  
276 Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community  
277 library for natural language processing. In *Proceedings of the 2021 Conference on Empirical  
278 Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, 2021.

- 270 Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for  
271 efficient convnets. In *International Conference on Learning Representations*, 2022.  
272
- 273 Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learn-  
274 ing efficient convolutional networks through network slimming. In *Proceedings of the IEEE*  
275 *international conference on computer vision*, pp. 2736–2744, 2017.
- 276 Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through  
277 l<sub>0</sub> regularization. In *International Conference on Learning Representations*, 2018.  
278
- 279 Xinyin Ma, Gongfan Fang, and Xinchao Wang. LLM-pruner: On the structural pruning of large  
280 language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.  
281 URL <https://openreview.net/forum?id=J8Ajf9WfXP>.
- 282 Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regu-  
283 larization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.  
284
- 285 Jonathan Schwarz, Siddhant Jayakumar, Razvan Pascanu, Peter E Latham, and Yee Teh. Power-  
286 propagation: A sparsity inducing weight reparameterisation. *Advances in neural information*  
287 *processing systems*, 34:28889–28903, 2021.
- 288 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,  
289 Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg,  
290 S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural*  
291 *Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- 292 Jian Wang, Huaqing Zhang, Junze Wang, Yifei Pu, and Nikhil R Pal. Feature selection using a  
293 neural network with group lasso regularization and controlled redundancy. *IEEE transactions on*  
294 *neural networks and learning systems*, 32(3):1110–1123, 2020.  
295
- 296 Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in  
297 deep neural networks. *Advances in neural information processing systems*, 29, 2016.
- 298 Chen Yang, Zhenghong Yang, Abdul Mateen Khattak, Liu Yang, Wenxin Zhang, Wanlin Gao, and  
299 Minjuan Wang. Structured pruning of convolutional neural networks via l1 regularization. *IEEE*  
300 *Access*, 7:106385–106394, 2019.  
301
- 302 Taisuke Yasuda, Kyriakos Axiotis, Gang Fu, MohammadHossein Bateni, and Vahab Mirrokni. Se-  
303 quentialattention++ for block sparsification: Differentiable pruning meets combinatorial opti-  
304 mization. *arXiv preprint arXiv:2402.17902*, 2024.
- 305 Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal*  
306 *of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.  
307
- 308 Huaqing Zhang, Jian Wang, Zhanquan Sun, Jacek M Zurada, and Nikhil R Pal. Feature selection  
309 for neural networks using group lasso regularization. *IEEE Transactions on Knowledge and Data*  
310 *Engineering*, 32(4):659–673, 2019.
- 311 Xiao Zhou, Weizhong Zhang, Hang Xu, and Tong Zhang. Effective sparsification of neural networks  
312 with global sparsity constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision*  
313 *and Pattern Recognition*, pp. 3599–3608, 2021.
- 314 Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression  
315 for large language models. *Transactions of the Association for Computational Linguistics*, 12:  
316 1556–1577, 2024.  
317
- 318 Liu Ziyin and Zihao Wang. spred: Solving l1 penalty with sgd. In *International Conference on*  
319 *Machine Learning*, pp. 43407–43422. PMLR, 2023.  
320  
321  
322  
323

## A PROOF OF THEOREM 2.2

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

$$\mathcal{O} : \mathbb{R}^{|A|} \times \prod_{d=1}^{\bar{D}} \mathbb{R}^H \rightarrow \mathbb{R}^{|A|}, (\mathbf{v}_A, \gamma_1, \dots, \gamma_{\bar{D}}) = (\mathbf{v}_A, \Gamma) \mapsto \mathbf{v}_A \triangleright \gamma^\odot \quad (6)$$

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  $(\mathbf{v}_A, \Gamma)$  such that  $\mathcal{O}(\mathbf{v}_A, \Gamma) = \mathbf{w}_A$ . Because  $\mathcal{L}_0(\cdot)$  only depends on the gated  $\mathbf{v}_A \triangleright \gamma^\odot = \mathbf{w}_A$ , every minimizer  $(\hat{\mathbf{v}}_A, \hat{\Gamma})$  of the gated objective must also minimize  $\Phi(\mathbf{v}_A, \Gamma)$  over the feasible set  $F_{\hat{\mathbf{w}}_A} = \{(\mathbf{v}_A, \Gamma) : \mathbf{v}_A \triangleright \gamma^\odot = \hat{\mathbf{w}}_A\}$ , where  $\hat{\mathbf{w}}_A := \hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\odot$ . Otherwise, the objective can always be strictly decreased by choosing  $(\mathbf{v}'_A, \Gamma') \in F_{\hat{\mathbf{w}}_A}$  so that  $\Phi(\mathbf{v}'_A, \Gamma') < \Phi(\hat{\mathbf{v}}_A, \hat{\Gamma})$  while  $\mathcal{L}_0(\mathbf{v}'_A \triangleright \gamma'^\odot) = \mathcal{L}_0(\hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\odot)$ . 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  $\Omega_A$* ) *The minimum of  $\Phi(\mathbf{v}_A, \Gamma)$  over the feasible set  $F_{\mathbf{w}_A}$  of gating parameters that leave  $\mathcal{L}_0$  unchanged is given by*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\min_{\mathbf{v}_A, \Gamma} \{ \mathcal{L}_0(\mathbf{v}_A \triangleright \gamma^\circ) + \lambda \Phi_A(\mathbf{v}_A, \Gamma) \} = \min_{\mathbf{w}_A} \{ \mathcal{L}_0(\mathbf{w}_A) + \lambda \min_{\mathbf{v}_A \triangleright \gamma^\circ = \mathbf{w}_A} \{ \Phi_A(\mathbf{v}_A, \Gamma) \} \}. \quad (9)$$

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

$$\min_{\mathbf{v}_A, \Gamma} \mathcal{L}_0(\mathbf{v}_A \triangleright \gamma^\circ) + \lambda \Phi_A(\mathbf{v}_A, \Gamma) = \min_{\mathbf{w}_A} \mathcal{L}_0(\mathbf{w}_A) + \lambda \Omega_A(\mathbf{w}_A). \quad (10)$$

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  $(\hat{\mathbf{v}}_A, \hat{\Gamma})$  of the gated objective  $\mathcal{L}_0(\hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\circ) + \lambda \Phi_A(\hat{\mathbf{v}}_A, \hat{\Gamma})$ , related as  $\hat{\mathbf{w}}_A = \hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\circ$ .

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  $\varepsilon_0$ -ball around  $\hat{\mathbf{w}}_A \in \mathbb{R}^{|\mathcal{A}|}$ . Due to the multiplicative nature of the gating operation  $\mathcal{O}$ , we can pick balanced gating parameters  $(\hat{\mathbf{v}}_A, \hat{\Gamma})$  (sometimes abbreviated as  $\hat{\psi}$  from now on) so that  $\mathcal{O}(\hat{\mathbf{v}}_A, \hat{\Gamma}) = \hat{\mathbf{v}}_A \triangleright \hat{\Gamma} = \hat{\mathbf{w}}_A$ . As the gating variables are balanced, applying to Lemma A.1 shows that  $\mathcal{L}_{\psi}(\hat{\mathbf{v}}_A, \hat{\Gamma}) := \mathcal{L}_0(\hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\circ) + \lambda \Phi_A(\hat{\mathbf{v}}_A, \hat{\Gamma})$  reduces to

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

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

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

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

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

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

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  $\hat{\psi} = (\hat{\mathbf{v}}_A, \hat{\Gamma}) \in \arg \min_{\psi} \mathcal{L}_{\psi}(\psi)$ . 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  $\mathcal{L}_{\mathbf{w}}(\hat{\mathbf{w}}_A) = \mathcal{L}_0(\hat{\mathbf{w}}_A) + \lambda \Omega_A(\hat{\mathbf{w}}_A)$ , there are corresponding local minimizers  $(\hat{\mathbf{v}}_A, \hat{\Gamma})$  of  $\mathcal{L}_{\psi}(\hat{\mathbf{v}}_A, \hat{\Gamma}) = \mathcal{L}_0(\hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\circ) + \lambda \Phi_A(\hat{\mathbf{v}}_A, \hat{\Gamma})$ , so that  $\hat{\mathbf{v}}_A \triangleright \hat{\gamma}^\circ = \hat{\mathbf{w}}_A$  and  $\mathcal{L}_{\psi}(\hat{\mathbf{v}}_A, \hat{\Gamma}) = \mathcal{L}_{\mathbf{w}}(\hat{\mathbf{w}}_A)$ .

□



## B FURTHER RESULTS

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

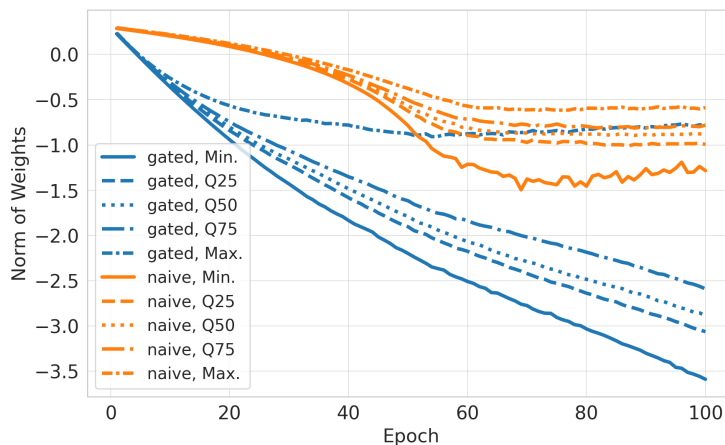


Figure 2: Quantiles of the weight norms (log scale) across attention heads for  $\lambda = 0.005$ .

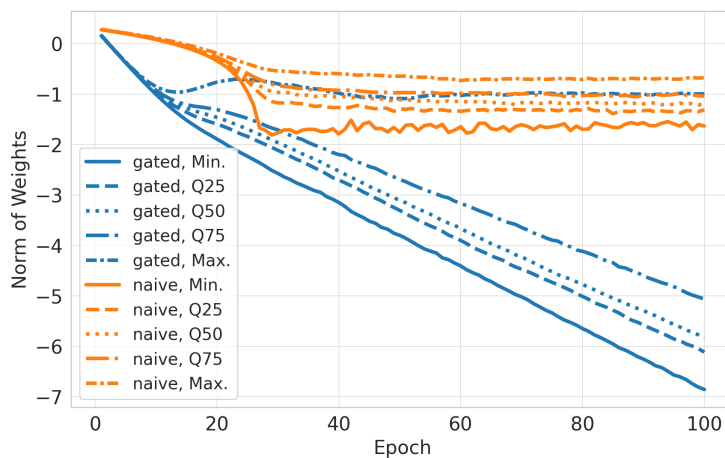


Figure 3: Quantiles of the weight norms (log scale) across the attention heads for  $\lambda = 0.01$ .

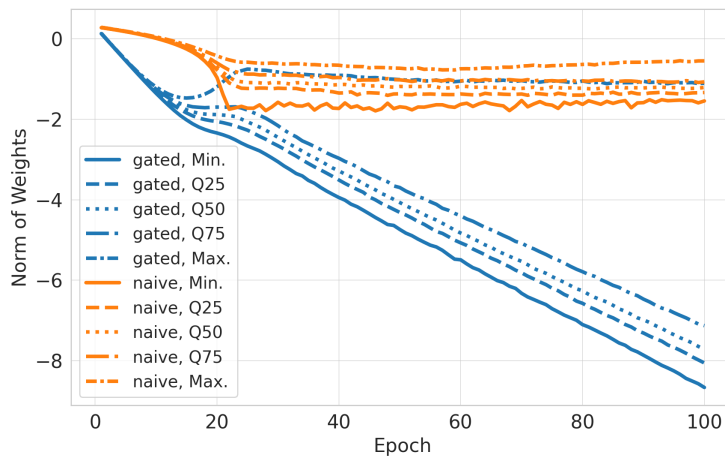


Figure 4: Quantiles of the weight norms (log scale) across the attention heads for  $\lambda = 0.0125$ .