HOW SPARSE ATTENTION APPROXIMATES EXACT ATTENTION?YOUR ATTENTION IS NATURALLY n^C-SPARSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse Attention is a technique that approximates standard attention computation with sub-quadratic complexity. This is achieved by selectively ignoring smaller entries in the attention matrix during the softmax function computation. Variations of this technique, such as pruning KV cache, sparsity-based fast attention, and Sparse Transformer, have been extensively utilized for efficient Large Language Models (LLMs) deployment. Despite its widespread use, a theoretical understanding of the conditions under which sparse attention performs on par with traditional attention remains elusive. This work aims to **bridge this gap by examining the inherent sparsity of standard attention processes.** Our theoretical framework reveals several brand-new key insights:

- Attention is n^C-sparse, implying that considering only the largest $\Omega(n^C)$ entries out of all n entries is sufficient for sparse attention to approximate the exact attention matrix with decreasing loss. Here, n represents the input length and $C \in (0, 1)$ is a constant.
- Stable $o(\log(n))$-sparse attention, which approximates attention computation with $\log(n)$ or fewer entries, may not be feasible since the error will persist at a minimum of $O(1)$.
- An adaptive strategy ($\alpha \cdot n^C, \alpha \in \mathbb{R}$) for the window size of efficient attention methods rather than a fixed one is guaranteed to perform more accurately and efficiently in a task for inference on flexible context lengths.

1 INTRODUCTION

Large Language Models (LLMs) Vaswani et al. (2017); Radford et al. (2018); Devlin et al. (2018); Radford et al. (2019); Brown et al. (2020); Chowdhery et al. (2022); Zhang et al. (2022); ChatGPT (2022) have emerged as a cornerstone of contemporary artificial intelligence, exhibiting remarkable capabilities across a plethora of AI domains. Their prowess is grounded in their ability to comprehend and generate human language with a level of sophistication that is unprecedented. This has catalyzed transformative applications in natural language processing, including machine translation He et al. (2021), content creation ChatGPT (2022); OpenAI (2023), and beyond, underscoring the profound impact of LLMs on the field of AI.

However, the architectural backbone of these models, particularly those built on the transformer framework Vaswani et al. (2017), presents a significant challenge: computational efficiency Tay et al. (2022). The essence of the transformer architecture, the Attention mechanism, necessitates a computational and memory complexity of $O(n^2)$, where n represents the sequence length. This quadratic dependency limits the scalability of LLMs, especially as we venture into processing longer sequences or expanding model capacities.

In an effort to mitigate this bottleneck, the AI research community has pivoted towards innovative solutions, one of which is sparse attention Child et al. (2019); Correia et al. (2019). Sparse attention mechanisms aim to approximate the results of the full attention computation by selectively focusing on a subset of the input data points. This is typically achieved by omitting certain interactions in the Query and Key multiplications within the attention mechanism, thereby inducing sparsity in the attention matrix. In order to arrive at the goal of preserving the model's performance while alleviating the computational and memory demands, prior works, including pruning KV cache, sparsity-based fast attention, and sparse transformer modeling, demonstrate outstanding efficiencies with $O(n^{1+o(1)})$ (sub-quadratic) complexity and sub-linear memory cache with competitive performance compared with standard attention across various tasks Liu et al. (2023b); Zhang et al. (2024); Kacham et al. (2023); Addanki et al. (2023); Lee et al. (2024); Xiong et al. (2021); Zandieh et al. (2023); Alman & Song (2023; 2024b;a); Han et al. (2023).

Despite these advancements, the theoretical underpinnings of sparse attention mechanisms and their implications on model performance and behavior remain an area of active inquiry. In detail, it's not clear when and when not sparse attention can approximate standard attention with a stable error. Also, the sparsity that attention naturally processes, which we call attention sparsity, lacks a strict confirmation of its existence and measurement. Especially, we would like to ask:

How Sparse Attention Approximates Exact Attention?

Our Contributions. In this work, we explore the theory of the sparse attention computation problem. Particularly, we first provide a analysis framework that first theoretically confirms the sparsity appears in standard attention. In detailed, our analysis describes the relationships between attention sparsity and input boundary, weights of attention networks and context length. Therefore, we derive several incremental insights based on this framework.

2 Preliminary

Assumption. In this work, we consider one-layer self-attention computation both in standard form and sparsity-based approximate form. To begin with, we give the assumption of the input matrix of attention computation, denoted as $X \in \mathbb{R}^{n \times d}$ where n is the context length and d stands the dimension, as follows (refer to Definition C.1 for the formal and detailed version of assumption):

- **Independent Entries.** For any two entries X_{i_1,j_1} and X_{i_2,j_2} in matrix X, $\forall i_1, i_2 \in [n]$ and $j_1, j_2 \in [d]$, they are independent.
- **Bounded Entries.** For failure probability $\delta \in (0, 0.1)$. With a probability $1-\delta$, the entry $X_{i,j}$ in matrix X, $\forall i \in [n]$ and $j \in [d]$, we have $|X_{i,j}| \leq B$ for some positive constant $B > 0$.

Attention Computation. Hence, we are about to introduce the standard attention computation, which occupies the main time and space complexity $O(n^2)$ in LLMs inference. First, we denote the weights of query, key and value projection as $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$. Thus, we let query, key and value state matrices be computed by $Q := XW_Q, K := XW_K, V := XW_V \in \mathbb{R}^{n \times d}$. We state the following definition:

Definition 2.1 (Attention computation)**.** *Given Query, Key and Value states matrices $Q, K, V \in \mathbb{R}^{n \times d}$. We then define $A := \exp(QK^\top/\sqrt{d})$, $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n \times n}$. The attention computation $\mathsf{Attn}(Q, K, V) \in \mathbb{R}^{n \times d}$ is given by: $\mathsf{Attn}(Q, K, V) := D^{-1}AV \in \mathbb{R}^{n \times d}$. Specially, we denote $D^{-1} = \mathrm{diag}(1/(A\mathbf{1}_n)) \in \mathbb{R}^{n \times n}$.*

Attention Sparsity. In Definition 2.1, $D^{-1}A \in \mathbb{R}^{n \times n}$ represents the attention matrix, indicating how much the model focuses on each vector. In much of the sparse attention literature, $D^{-1}A$ is assumed to be sufficiently sparse, allowing sparsity-based efficient attention methods to disregard some zero entries in order to achieve a balance between accuracy and efficiency. In this paper, we introduce a threshold, denoted as ϵ, and define **attention sparsity** as **the number of entries in each row of** $D^{-1}A \in \mathbb{R}^n$ **that are smaller than** ϵ. Specifically, for a softmax vector $u \in \mathbb{R}^n$, if there are at least $n - k$ entries in u that are not greater than ϵ for all integers $k \in [n]$, we say that u is (ϵ, k)-sparse. Since ϵ is intended to be a very small value, we will simply refer to u as being k-sparse. The formal definition is provided below:

Definition 2.2 ((ϵ, k)-sparsity)**.** *For a vector $u \in \mathbb{R}^n$ and error $\epsilon > 0$, we define sparse set $\mathcal{S}_\epsilon(u)$ as: $\mathcal{S}_\epsilon(u) := \{i \in [n] \mid |u_i| \leq \epsilon\}$. Hence, we say u is at least (ϵ, k)-sparse when it holds that $|\mathcal{S}_\epsilon(u)| \geq n - k$.*

Problem Definition I: Estimating ϵ. In practical implementations, establishing a clear relationship between ϵ and the sparsity k proves to be challenging. Therefore, we first analyze how to estimate a boundary for ϵ based on a given sparsity integer $k \in [n]$. Naively, given k, we would like to find a guaranteed value for ϵ that satisfies $|\mathcal{S}_\epsilon(u)| \geq n - k$. Address this problem will enable us to assess the loss associated with approximating standard attention using sparse attention, ultimately guiding us in finding the optimal trade-off between ϵ (where lower values yield greater accuracy) and k (where lower values lead to higher efficiency).

Sparse Attention and Approximation. Here we state an ideal mathematical definition for the sparse attention in this paper. Initially, we define a set, $\mathcal{T}_k(u)$, to filter out the greatest k entries in a vector $u \in \mathbb{R}^n$. The integer $k \in [1, n]$ is also called window size in some sparse attention works.

Definition 2.3. *For a vector $u \in \mathbb{R}^n$, given a sparsity integer k, we denote a top-k set $\mathcal{T}_k(u) := \{i \in [n] \mid \mathcal{S}_{u_i}(u) \geq n - k\}$, then we define vector $\mathsf{topk}(u) := [u_1 \cdot \mathbf{1}_{1 \in \mathcal{T}_k(u)}, \cdots, u_n \cdot \mathbf{1}_{n \in \mathcal{T}_k(u)}]^\top \in \mathbb{R}^n$.*

Note that $\mathbf{1}_{i \in \mathcal{T}_k(u)}$ is an indicator where when $i \in \mathcal{T}_k(u)$, it equals 1, otherwise, 0. We utilize $\mathsf{topk}(u)$ to compute a sparsity-based approximating version of $A = \exp(QK^\top)$ in Definition 2.1, we denote it A_{spar}. Accordingly, we provide a universal version for all sparsity-based attention as follows:

Definition 2.4 (Sparse attention)**.** *Given Query, Key and Value state matrices $Q, K, V \in \mathbb{R}^{n \times d}$. We then define $A := \exp(QK^\top/\sqrt{d}) \in \mathbb{R}^{n \times n}$. Especially, we define $A_{\mathrm{spar}} := [\mathsf{topk}(A_{1,*}), \cdots, \mathsf{topk}(A_{n,*})]^\top$, $D_{\mathrm{spar}} := \mathrm{diag}(A_{\mathrm{spar}} \mathbf{1}_n) \in \mathbb{R}^{n \times n}$. The sparse attention computation $\mathsf{SparseAttn}(Q, K, V) \in \mathbb{R}^{n \times d}$ is given by:*

$$\mathsf{SparseAttn}(Q, K, V) := D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} V \in \mathbb{R}^{n \times d}$$

Specially, we denote $D_{\mathrm{spar}}^{-1} = \mathrm{diag}(1/(A_{\mathrm{spar}} \mathbf{1}_n)) \in \mathbb{R}^{n \times n}$.

It should be noted that directly accessing top k entries in the attention matrix without any extra computational cost is overly ideal for efficient LLMs in real-world cases. Prior works usually utilize some additional approximate algorithm to meet this condition, e.g. Locality-Sensitive Hashing (LSH) for retrieving larger query-key pairs, but this also brings more approximating errors. We only focus on the part of approximating attention computation in this study and leave the part of pre-approximating top-k entries in $D^{-1}A$ as a future direction.

Problem Definition II: Sparse Attention Approximation. The variations of sparse attention, including pruning KV Cache Liu et al. (2021); Xiao et al. (2023) and sparsity-based attention Kitaev et al. (2020); Zandieh et al. (2023); Han et al. (2023), focus on solving the approximation of the attention matrix, where we call it sparse attention approximation. In particular, we emphasize the importance of *stable sparse attention approximation*, which directly affects the extensibility of sparse attention under long context scenes. We denote $f : \mathbb{N}^+ \to \mathbb{N}^+$ as the strategy to choose a suitable window size due to different input lengths. Hence, we give:

Definition 2.5 (*Stable sparse attention approximation* $\mathsf{SSAA}(f)$)**.** *For some strategy $f : \mathbb{N}^+ \to \mathbb{N}^+$ to choose the sparsity $k = f(n)$ in sparse attention (Definition 2.4), the problem of stable sparse attention approximation $\mathsf{SSAA}(f)$ is to solve: $L(f, n) = \|D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} - D^{-1} A\|_p$, where $\|\cdot\|_p$ denotes some norm. We say this sparse attention approximation is* **stable** *iff:*

- *$L(f, n)$ is monotonically decreasing with growing n.*
- *$\lim_{n \to +\infty} L(f, n) = 0$.*

3 Insights Overview

Definition 3.1. *Denote $W := W_Q W_K^\top / \sqrt{d} \in \mathbb{R}^{d \times d}$. We define $R := B^2 \cdot \|W\|_F$.*

We estimate the lower bound on the requirement for (ϵ, k)-sparse softmax vector, proving the **vanilla attention computation is naturally sparse**.

Theorem 3.2. *Let $R \geq 0$ be defined as Definition 3.1. Given sparsity integer $k \leq n$. Denote $T := \exp(\sqrt{\log(n(n-k)d/\delta)})$. Let $\mathcal{S}_\epsilon$ be defined as Definition 2.2. $\delta \in (0, 0.1)$. If we choose $\epsilon \geq \frac{T^{O(R)}}{n}$, then with a probability at least $1 - \delta$, for all $i \in [n]$, we have $\left|\mathcal{S}_\epsilon(D_{i,i}^{-1} A_{i,*})\right| \geq n - k$.*

Proof sketch of Theorem 3.2. The complete proof is provided in Appendix D and Theorem D.1. □

We introduce the concept of *attention collapse*, which demonstrates the number of effective entries in attention matrix provably decrease to 1 or some constant inevitably.

Theorem 3.3. *Consider a fixed ϵ with a very small value, $\delta \in (0, 0.1)$. Then with a probability at least $1 - \delta$, there is:*

- *Part 1. If $R = o(\sqrt{\log(n)})$, then we have $\lim_{n \to +\infty} |\mathcal{S}_\epsilon(u)| \geq n - 1$.*
- *Part 2. If $R = O(\sqrt{\log(n)})$, then we have $\lim_{n \to +\infty} |\mathcal{S}_\epsilon(u)| \geq O(1)$.*

Proof sketch of Theorem 3.3. Refer to Theorem D.2 for the detailed proof. □

We give the sufficient lower bound on the window size of stable sparse attention approximating exact attention computation, $\Omega(n^C)$ for constant $C \in (0, 1)$. This further indicates that sparse attention can recover attention outputs from limited $\Omega(n^C)$ entries while achieving a decreasing error.

Theorem 3.4. $\delta \in (0, 0.1)$. *For a constant $C \in (0, 1)$, we then denote $f(n) := \Omega(n^C)$, therefore, with a probability at least $1 - \delta$, window size strategy $k = f(n)$ is sufficient to solve $\mathsf{SSAA}(f)$ in Definition 2.5.*

Proof sketch of Theorem 3.4. Please see Theorem E.1. □

Meanwhile, we also confirm sparse attention approximation from $o(\log(n))$ entries is not enough for stability and extensibility since the lower bound on error will grow with increasing input length.

Theorem 3.5. $\delta \in (0, 0.1)$. *For a constant $C \in (0, 1)$, we then denote $f(n) := o(\log(n))$, therefore, with a probability at least $1 - \delta$, window size strategy $k = f(n)$ cannot solve $\mathsf{SSAA}(f)$ in Definition 2.5.*

Proof sketch of Theorem 3.5. Please refer to Theorem E.2 for the formal version and corresponding detailed proofs. □

Therefore, we suggest to use adaptive strategy $k = \alpha \cdot n^C, \alpha > 0, C \in (0, 1)$ for the window size of sparse attention rather than the strategy that fixes the window size for any input. The former is proved more efficient within higher approximation performance. We consider a dataset $\mathcal{D} := \{X^i\}_{i=1}^N$ with dataset size N. For all $i \in [N]$, we use n_i to denote the context length, such that $X^i \in \mathbb{R}^{n_i \times d}$. Hence, the difference of the computational complexities of fixed Top-k strategy and a dynamic (especially $O(n^C)$) strategy could be easily obtained in the Claim below.

Claim 3.6. *We have:*

- **Part 1.** *Choosing the constant window size strategy $k = p$ for some constant integer $p > 0$. The computational complexity of a one-layer p-sparse attention to inference $\mathcal{D} = \{X^i\}_i^N$ is $\Theta(p \sum_{i=1}^N n_i)$.*
- **Part 2.** *Choosing the constant window size strategy $k = \alpha \cdot n^C$ for some constant $\alpha, C > 0$. The computational complexity of a one-layer p-sparse attention to inference $\mathcal{D} = \{X^i\}_i^N$ is $\Theta(\alpha \sum_{i=1}^N n_i^{1+C})$.*

Proof. Recall the computational complexity of Definition 2.4 is $O(nk)$ for input $X \in \mathbb{R}^{n \times d}$ and k window size. We then take the summation for each $X^i \in \mathbb{R}^{n_i \times d}$ in $\mathcal{D} := \{X^i\}_i^N$ to obtain the results of **Part 1** and **Part 2**. □

Proposition 3.7. *For any window size strategy $k = p$ for some constant integer $p > 0$, there exist a context-length adaptive strategy $k = \alpha \cdot n^C$ for some constant $\alpha, C > 0$ that performs lower approximating error.*

Proof. Following Theorem 3.4 and Theorem 3.5, the conclusion of this proposition can be trivially proved by plugging suitable choices of α and C. □

REFERENCES

Addanki, R., Li, C., Song, Z., and Yang, C. One pass streaming algorithm for super long token attention approximation in sublinear space. *arXiv preprint arXiv:2311.14652*, 2023.

Alman, J. and Song, Z. Fast attention requires bounded entries. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

Alman, J. and Song, Z. The fine-grained complexity of gradient computation for training large language models. *arXiv preprint arXiv:2402.04497*, 2024a.

Alman, J. and Song, Z. How to capture higher-order correlations? generalizing matrix softmax attention to kronecker computation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024b. URL `https://openreview.net/forum?id=v0zNCwwkaV`.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Bernstein, S. On a modification of chebyshev's inequality and of the error formula of laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Brand, J. v. d., Song, Z., and Zhou, T. Algorithm and hardness for dynamic attention maintenance in large language models. *arXiv preprint arXiv:2304.02207*, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Cai, H., Lou, Y., Mckenzie, D., and Yin, W. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. *arXiv preprint arXiv:2102.10707*, 2021.

ChatGPT. Optimizing language models for dialogue. *OpenAI Blog*, November 2022. URL `https://openai.com/blog/chatgpt/`.

Chen, B., Liu, Z., Peng, B., Xu, Z., Li, J. L., Dao, T., Song, Z., Shrivastava, A., and Re, C. Mongoose: A learnable lsh framework for efficient neural network training. In *International Conference on Learning Representations*, 2020.

Chernoff, H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pp. 493–507, 1952.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Chu, T., Song, Z., and Yang, C. How to protect copyright data in optimization of large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17871–17879, 2024.

Correia, G. M., Niculae, V., and Martins, A. F. Adaptively sparse transformers. *arXiv preprint arXiv:1909.00015*, 2019.

Deng, Y., Li, Z., Mahadevan, S., and Song, Z. Zero-th order algorithm for softmax attention optimization. *arXiv preprint arXiv:2307.08352*, 2023a.

Deng, Y., Mahadevan, S., and Song, Z. Randomized and deterministic attention sparsification algorithms for over-parameterized feature dimension. *arXiv preprint arXiv:2304.04397*, 2023b.

Deng, Y., Song, Z., Xie, S., and Yang, C. Unmasking transformers: A theoretical approach to data recovery via attention weights. *arXiv preprint arXiv:2310.12462*, 2023c.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Foss, S., Korshunov, D., Zachary, S., et al. *An introduction to heavy-tailed and subexponential distributions*, volume 6. Springer, 2011.

Gao, Y., Song, Z., Wang, W., and Yin, J. A fast optimization view: Reformulating single layer attention in llm based on tensor and svm trick, and solving it in matrix multiplication time. *arXiv preprint arXiv:2309.07418*, 2023a.

Gao, Y., Song, Z., and Xie, S. In-context learning for attention scheme: from single softmax regression to multiple softmax regression via a tensor trick. *arXiv preprint arXiv:2307.02419*, 2023b.

Gao, Y., Song, Z., and Yin, J. Gradientcoin: A peer-to-peer decentralized large language models. *arXiv preprint arXiv:2308.10502*, 2023c.

Haagerup, U. The best constants in the khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

Han, I., Jayaram, R., Karbasi, A., Mirrokni, V., Woodruff, D. P., and Zandieh, A. Hyperattention: Long-context attention in near-linear time. *arXiv preprint arXiv:2310.05869*, 2023.

Hanson, D. L. and Wright, F. T. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

He, W., Wu, Y., and Li, X. Attention mechanism for neural machine translation: A survey. In *2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, volume 5, pp. 1485–1489. IEEE, 2021.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pp. 409–426, 1994.

Kacham, P., Mirrokni, V., and Zhong, P. Polysketchformer: Fast transformers via sketches for polynomial kernels. *arXiv preprint arXiv:2310.01655*, 2023.

Khintchine, A. Über dyadische brüche. *Mathematische Zeitschrift*, 18(1):109–116, 1923.

Kitaev, N., Kaiser, Ł., and Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Lample, G., Sablayrolles, A., Ranzato, M., Denoyer, L., and Jégou, H. Large memory layers with product keys. *Advances in Neural Information Processing Systems*, 32, 2019.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pp. 1302–1338, 2000.

Lee, S., Lee, H., and Shin, D. Proxyformer: Nyström-based linear transformer with trainable proxy tokens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 13418–13426, 2024.

Li, C., Song, Z., Wang, W., and Yang, C. A theoretical insight into attack and defense of gradient leakage in transformer. *arXiv preprint arXiv:2311.13624*, 2023a.

Li, Z., Song, Z., and Zhou, T. Solving regularized exp, cosh and sinh regression problems. *arXiv preprint, 2303.15725*, 2023b.

Liu, H., Li, Z., Hall, D., Liang, P., and Ma, T. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023a.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Liu, Z., Wang, J., Dao, T., Zhou, T., Yuan, B., Song, Z., Shrivastava, A., Zhang, C., Tian, Y., Re, C., et al. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pp. 22137–22176. PMLR, 2023b.

Lu, Y., Dhillon, P., Foster, D. P., and Ungar, L. Faster ridge regression via the subsampled randomized hadamard transform. *Advances in neural information processing systems*, 26, 2013.

Malladi, S., Gao, T., Nichani, E., Damian, A., Lee, J. D., Chen, D., and Arora, S. Fine-tuning language models with just forward passes. *arXiv preprint arXiv:2305.17333*, 2023a.

Malladi, S., Wettig, A., Yu, D., Chen, D., and Arora, S. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pp. 23610–23641. PMLR, 2023b.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Panigrahi, A., Saunshi, N., Zhao, H., and Arora, S. Task-specific skill localization in fine-tuned language models. *arXiv preprint arXiv:2302.06600*, 2023.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. ., 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., D.Manning, C., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Rudelson, M. and Vershynin, R. Hanson-wright inequality and sub-gaussian concentration. 2013.

Song, Z., Wang, W., and Yin, J. A unified scheme of resnet and softmax. *arXiv preprint arXiv:2309.13482*, 2023a.

Song, Z., Yin, J., and Zhang, L. Solving attention kernel regression problem via pre-conditioner. *arXiv preprint arXiv:2308.14304*, 2023b.

Sun, Z., Yang, Y., and Yoo, S. Sparse attention with learning to hash. In *International Conference on Learning Representations*, 2021.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28, 2022.

Tropp, J. A. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., and Singh, V. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14138–14148, 2021.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., et al. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297, 2020.

Zandieh, A., Han, I., Daliri, M., and Karbasi, A. Kdeformer: Accelerating transformers via kernel density estimation. In *International Conference on Machine Learning*, pp. 40605–40623. PMLR, 2023.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Appendix

CONTENTS

A RELATED WORK

Sparse and Efficient Transformer. In the landscape of attention mechanisms, Vaswani et al. introduced the transformative transformer model, revolutionizing NLP with its comprehensive self-attention mechanism Vaswani et al. (2017). Innovations Child et al. (2019); Lample et al. (2019) in sparse attention presented methods to reduce complexity, maintaining essential contextual information while improving computational efficiency. The Reformer Kitaev et al. (2020) utilized Locality Sensitive Hashing to significantly cut down computational demands, enabling the processing of lengthy sequences. Mongoose Chen et al. (2020) adapted sparsity patterns dynamically, optimizing computation without losing robustness. Sun et al. (2021) introduced a learning-to-hash strategy to generate sparse attention patterns, enhancing data-driven efficiency. HyperAttention Han et al. (2023) refined attention approximation, balancing computational savings with accuracy. Longformer Beltagy et al. (2020) extended transformer capabilities to longer texts through a mix of global and local attention mechanisms. The Performe Choromanski et al. (2020) offered a novel approximation of softmax attention, reducing memory usage for long sequences. Big Bird Zaheer et al. (2020) combined global, local, and random attention strategies to surmount traditional transformer limitations regarding sequence length.

Theoretical Approaches to Understanding LLMs. There have been notable advancements in the field of regression models, particularly with the exploration of diverse activation functions, aiding in the comprehension and optimization of these models. The study of over-parameterized neural networks, focusing on exponential and hyperbolic activation functions, has shed light on their convergence traits and computational benefits Brand et al. (2023); Song et al. (2023a); Gao et al. (2023c); Deng et al. (2023b); Gao et al. (2023b); Song et al. (2023b); Zandieh et al. (2023); Alman & Song (2023; 2024b;a); Gao et al. (2023a); Deng et al. (2023c); Li et al. (2023a); Chu et al. (2024). Enhancements in this area include the addition of regularization components and the innovation of algorithms like the convergent approximation Newton method to improve performance Li et al. (2023b). Additionally, employing tensor methods to simplify regression models has facilitated in-depth analyses concerning Lipschitz constants and time complexity Gao et al. (2023b); Deng et al. (2023a). Concurrently, there's a burgeoning interest in optimization algorithms specifically crafted for LLMs, with block gradient estimators being utilized for vast optimization challenges, significantly reducing computational load Cai et al. (2021). Novel methods such as Direct Preference Optimization are revolutionizing the tuning of LLMs by using human preference data, circumventing the need for traditional reward models Rafailov et al. (2023). Progress in second-order optimizers is also notable, offering more leniency in convergence proofs by relaxing the usual Lipschitz Hessian assumptions Liu et al. (2023a). Moreover, a series of studies focus on the intricacies of fine-tuning Malladi et al. (2023a;b); Panigrahi et al. (2023). These theoretical developments collectively push the boundaries of our understanding and optimization of LLMs, introducing new solutions to tackle challenges like the non-strict Hessian Lipschitz conditions.

B PRELIMINARY

B.1 NOTATIONS

In this work, we use the following notations and definitions:

- For integer n, we use $[n]$ to denote the set $\{1, \dots, n\}$.
- We use $\mathbf{1}_n$ to denote all-1 vector in $\mathbb{R}^n$.
- The ℓ_p norm of a vector x is denoted as $\|x\|_p$, for examples, $\|x\|_1 := \sum_{i=1}^n |x_i|$, $\|x\|_2 := (\sum_{i=1}^n x_i^2)^{1/2}$ and $\|x\|_\infty := \max_{i \in [n]} |x_i|$.
- For a vector $x \in \mathbb{R}^n$, $\exp(x) \in \mathbb{R}^n$ denotes a vector where whose i-th entry is $\exp(x_i)$ for all $i \in [n]$.
- For two vectors $x, y \in \mathbb{R}^n$, we denote $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$ for $i \in [n]$.
- Given two vectors $x, y \in \mathbb{R}^n$, we denote $x \circ y$ as a vector whose i-th entry is $x_i y_i$ for all $i \in [n]$.

- For a vector $x \in \mathbb{R}^n$, $\mathrm{diag}(x) \in \mathbb{R}^{n \times n}$ is defined as a diagonal matrix with its diagonal entries given by $\mathrm{diag}(x)_{i,i} = x_i$ for $i = 1, ..., n$, and all off-diagonal entries are 0.
- We use $\mathrm{erf}(x)$ to denote the error function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\mathrm{d}t$, and erf^{-1} is denoted as the inverse function of $\mathrm{erf}(x)$.
- For any matrix $A \in \mathbb{R}^{m \times n}$, we use $A^\top$ to denote its transpose, we use $\|A\|_F$ to denote the Frobenius norm and $\|A\|_\infty$ to denote its infinity norm, i.e., $\|A\|_F := (\sum_{i \in [m]} \sum_{j \in [n]} A_{i,j}^2)^{1/2}$ and $\|A\|_\infty = \max_{i \in [m], j \in [n]} |A_{i,j}|$.
- For $\mu, \sigma \in \mathbb{R}$, we use $\mathcal{N}(\mu, \sigma^2)$ to denote Gaussian distribution with expectation of μ and variance of σ^2.
- For a mean vector $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, we use $\mathcal{N}(\mu, \Sigma^2)$ to denote the vector Gaussian distribution.
- We use $\mathbb{E}[\cdot]$ to denote the expectation and $\mathrm{Var}[\cdot]$ to denote the variance.
- We use $\Gamma(x)$ to denote the gamma function where $\Gamma(x) = \int_0^\infty t^{z-1} \exp(-t)\mathrm{d}t$.
- For an integer $k > 0$, we use χ_k^2 to denote the Chi-squared distribution with k degrees of freedom.
- Usually, we use $C \geq 1$ to denote a sufficient large constant.

B.2 Basic Fact for Softmax

Fact B.1. *For a vector $x \in \mathbb{R}^d$ and a scalar $b \in \mathbb{R}$, we have:*

$$\mathsf{softmax}(x) = \mathsf{softmax}(x + b \cdot \mathbf{1}_d)$$

B.3 Basic Facts for Calculation

Fact B.2. *For $a, b \geq 1$ and there exist a constant $C \geq 0$ such that*

$$\sqrt{a} + \sqrt{b} \leq C\sqrt{a+b}$$

Fact B.3. *For a sufficient large $x \in \mathbb{R}$ ($x \geq 55$), we have*

$$\exp(\sqrt{\log(x)}) \leq \sqrt{x}$$

B.4 Probability Tools

Here, we state a probability toolkit in the following, including several helpful lemmas we'd like to use. Firstly, we provide the lemma about Chernoff bound in Chernoff (1952) below.

Lemma B.4 (Chernoff bound, Chernoff (1952))**.** *Let $X = \sum_{i=1}^n X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$. Then*

- $\Pr[X \geq (1+\delta)\mu] \leq \exp(-\delta^2\mu/3)$, $\forall \delta > 0$;
- $\Pr[X \leq (1-\delta)\mu] \leq \exp(-\delta^2\mu/1)$, $\forall 0 < \delta < 1$.

Next, we offer the lemma about Hoeffding bound as in Hoeffding (1994).

Lemma B.5 (Hoeffding bound, Hoeffding (1994))**.** *Let $X_1, \cdots, X_n$ denote n independent bounded variables in $[a_i, b_i]$ for $a_i, b_i \in \mathbb{R}$. Let $X := \sum_{i=1}^n X_i$, then we have*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2\exp(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2})$$

We show the lemma of Bernstein inequality as Bernstein (1924).

Lemma B.6 (Bernstein inequality, Bernstein (1924))**.** *Let $X_1, \cdots, X_n$ denote n independent zero-mean random variables. Suppose $|X_i| \leq M$ almost surely for all i. Then, for all positive t,*

$$\Pr[\sum_{i=1}^n X_i \geq t] \leq \exp(-\frac{t^2/2}{\sum_{j=1}^n \mathbb{E}[X_j^2] + Mt/3})$$

Then, we give the Khintchine's inequality in Khintchine (1923); Haagerup (1981) as follows:

Lemma B.7 (Khintchine's inequality, Khintchine (1923); Haagerup (1981))**.** *Let $\sigma_1, \cdots, \sigma_n$ be i.i.d sign random variables, and let $z_1 \cdots, z_n$ be real numbers. Then there are constants $C > 0$ so that for all $t > 0$*

$$\Pr[|\sum_{i=1}^{n} z_i \sigma_i| \geq t\|z\|_2] \leq \exp(-Ct^2)$$

We give Hason-wright inequality from Hanson & Wright (1971); Rudelson & Vershynin (2013) below.

Lemma B.8 (Hason-wright inequality, Hanson & Wright (1971); Rudelson & Vershynin (2013))**.** *Let $x \in \mathbb{R}^n$ denote a random vector with independent entries x_i with $\mathbb{E}[x_i] = 0$ and $|x_i| \leq K$ Let A be an $n \times n$ matrix. Then, for every $t \geq 0$*

$$\Pr[|x^\top Ax - \mathbb{E}[x^\top Ax]| > t] \leq 2\exp(-c\min\{t^2/(K^4\|A\|_F^2), t/(K^2\|A\|)\})$$

We state Lemma 1 on page 1325 of Laurent and Massart Laurent & Massart (2000).

Lemma B.9 (Lemma 1 on page 1325 of Laurent and Massart, Laurent & Massart (2000))**.** *Let $X \sim \mathcal{X}_k^2$ be a chi-squared distributed random variable with k degrees of freedom. Each one has zero mean and σ^2 variance. Then*

$$\Pr[X - k\sigma^2 \geq (2\sqrt{kt} + 2t)\sigma^2] \leq \exp(-t)$$
$$\Pr[X - k\sigma^2 \geq 2\sqrt{kt}\sigma^2] \leq \exp(-t)$$

Here, we provide a tail bound for sub-exponential distribution Foss et al. (2011).

Lemma B.10 (Tail bound for sub-exponential distribution, Foss et al. (2011))**.** *We say $X \in \mathrm{SE}(\sigma^2, \alpha)$ with parameters $\sigma > 0$, $\alpha > 0$, if*

$$\mathbb{E}[e^{\lambda X}] \leq \exp(\lambda^2\sigma^2/2), \forall|\lambda| < 1/\alpha.$$

Let $X \in \mathrm{SE}(\sigma^2, \alpha)$ and $\mathbb{E}[X] = \mu$, then:

$$\Pr[|X - \mu| \geq t] \leq \exp(-0.5\min\{t^2/\sigma^2, t/\alpha\})$$

In the following, we show the helpful lemma of matrix Chernoff bound as in Tropp (2011); Lu et al. (2013).

Lemma B.11 (Matrix Chernoff bound, Tropp (2011); Lu et al. (2013))**.** *Let $\mathcal{X}$ be a finite set of positive-semidefinite matrices with dimension $d \times d$, and suppose that*

$$\max_{X \in \mathcal{X}} \lambda_{\max}(X) \leq B.$$

Sample $\{X_1, \cdots, X_n\}$ uniformly at random from $\mathcal{X}$ without replacement. We define $\mu_{\min}$ and $\mu_{\max}$ as follows:

$$\mu_{\min} := n \cdot \lambda_{\min}(\mathop{\mathbb{E}}_{X \in \mathcal{X}}(X))$$
$$\mu_{\max} := n \cdot \lambda_{\max}(\mathop{\mathbb{E}}_{X \in \mathcal{X}}(X)).$$

Then

$$\Pr[\lambda_{\min}(\sum_{i=1}^{n} X_i) \leq (1-\delta)\mu_{\min}] \leq d \cdot \exp(-\delta^2\mu_{\min}/B) \textit{ for } \delta \in (0, 1],$$

$$\Pr[\lambda_{\max}(\sum_{i=1}^{n} X_i) \geq (1+\delta)\mu_{\max}] \leq d \cdot \exp(-\delta^2\mu_{\max}/(4B)) \textit{ for } \delta \geq 0.$$

C Problem Definitions

C.1 Input Assumption

Definition C.1. *We consider for any input matrix to an attention network $X \in \mathbb{R}^{n \times d}$ where integer n denotes the input length and d denotes the dimension. We assume:*

- **Independent Entries.** *For any two entries X_{i_1,j_1} and X_{i_2,j_2} in matrix X, $\forall i_1, i_2 \in [n]$ and $j_1, j_2 \in [d]$, they are independent.*
- **Bounded Entries.** *For failure probability $\delta \in (0, 0.1)$. With a probability $1 - \delta$, the entry $X_{i,j}$ in matrix X, $\forall i \in [n]$ and $j \in [d]$, we have $|X_{i,j}| \le B$ for some positive constant $B > 0$.*

C.2 Attention Computation

Definition C.2 (Attention computation)**.** *If the following conditions hold:*

- *Let $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ be denoted as Query, Key and Value projection matrices of attention.*
- *Given an input $X \in \mathbb{R}^{n \times d}$ that holds properties in Definition C.1.*
- *Define Query, Key and Value states matrices $Q := XW_Q, K := XW_K, V := XW_V \in \mathbb{R}^{n \times d}$.*
- $A := \exp(QK^\top / \sqrt{d}) \in \mathbb{R}^{n \times n}$.
- $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n \times n}$.

Then we have attention computation $\mathsf{Attn}(Q, K, V) \in \mathbb{R}^{n \times d}$ as follows:

$$\mathsf{Attn}(Q, K, V) := D^{-1}AV$$

C.3 ϵ-Approximated k-Sparse Softmax Vector

In order to describe the sparsity of the softmax, we define the following notation.

Definition C.3. *For a vector $u \in \mathbb{R}^n$ and $\epsilon \ge 0$, we define sparse set $\mathcal{S}_\epsilon(u)$ as follows:*

$$\mathcal{S}_\epsilon(u) := \left\{ i \in [n] \;\middle|\; |u_i| \le \epsilon \right\}$$

Definition C.4 ((ϵ, k)-sparsity)**.** *For a vector $u \in \mathbb{R}^n$, we say u is (ϵ, k)-sparse if for a constant $\epsilon \in (0, 1)$, it holds that*

$$|\mathcal{S}_\epsilon(u)| \ge n - k.$$

C.4 Sparse Attention

Definition C.5. *For a vector $u \in \mathbb{R}^n$. Given a sparsity integer k. Let $\mathcal{S}_\epsilon$ be defined as Definition C.3 for some error $\epsilon > 0$. We define the top-k set $\mathcal{T}_k(u) := \{i \in [n] \mid \mathcal{S}_{u_i}(u) \ge n - k\}$.*

Definition C.6. *If the following conditions hold:*

- *For a vector $u \in \mathbb{R}^n$.*
- *Given a sparsity integer k.*
- *Let $\mathcal{S}_\epsilon$ be defined as Definition C.3 for some error $\epsilon > 0$.*
- *Denote a top-k set $\mathcal{T}_k(u) := \{i \in [n] \mid \mathcal{S}_{u_i}(u) \ge n - k\}$ as Definition C.5.*

Then we define

$$\mathsf{topk}(u) := [u_i \cdot \mathbf{1}_{i \in \mathcal{T}_k(u)}]_{i \in [n]} \in \mathbb{R}^n$$

We define the sparse attention computation as follows:

Definition C.7. *If the following conditions hold:*

- *Let* $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ *be denoted as Query, Key and Value projection matrices of attention.*
- *Given an input* $X \in \mathbb{R}^{n\times d}$ *that holds properties in Definition C.1.*
- *Define Query, Key and Value states matrices* $Q := XZW_Q, K := XW_K, V := XW_V \in \mathbb{R}^{n\times d}$.
- $A := \exp(QK^\top/\sqrt{d}) \in \mathbb{R}^{n\times n}$.
- *Let* topk *be defined as Definition C.6.*
- *Define* $A_{\mathrm{spar}} := [\mathsf{topk}(A_1), \quad \mathsf{topk}(A_2), \quad \cdots, \quad \mathsf{topk}(A_n)]^\top \in \mathbb{R}^{n\times n}$.
- $D_{\mathrm{spar}} := \mathrm{diag}(A_{\mathrm{spar}}\mathbf{1}_n) \in \mathbb{R}^{n\times n}$.
- $\delta \in (0, 0.1)$.
- *Let* $R \geq 0$ *be defined as Definition C.9.*

The sparse attention computation $\mathsf{SparseAttn}(Q, K, V) \in \mathbb{R}^{n\times d}$ *is given by:*

$$\mathsf{SparseAttn}(Q, K, V) := D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} V \in \mathbb{R}^{n\times d}$$

C.5 Helpful Definitions

We introduce the following algebraic lemmas to be used later.

Definition C.8. *Let Query and Key projection matrices* $W_Q, W_K \in \mathbb{R}^{d\times d}$ *be defined as Definition C.2. We define*

$$W := W_Q W_K^\top / \sqrt{d}.$$

Definition C.9. *If the following conditions hold:*

- *Let* $W \in \mathbb{R}^{d\times d}$ *be define as Definition C.8.*

Then for any $i \in [n]$*, we define:*

$$R := B^2 \|W\|_F.$$

D Attention Sparsity

D.1 Main Result 1: Attention Sparsity with Upper Bound on Error

Theorem D.1. *If the following conditions hold:*

- *Let* $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ *be denoted as Query, Key and Value projection matrices of attention.*
- *Given an input* $X \in \mathbb{R}^{n\times d}$ *that holds properties in Definition C.1.*
- *Define Query, Key and Value states matrices* $Q := XW_Q, K := XW_K, V := XW_V \in \mathbb{R}^{n\times d}$.
- $A := \exp(QK^\top/\sqrt{d}) \in \mathbb{R}^{n\times n}$.
- $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n\times n}$.
- *Denote* $\beta_i := Bd \max_{j_1\in[d]} |\mathbb{E}[\sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i,j_2}]|$.

- *Define* $\Gamma := [\beta_1 \cdot \mathbf{1}_n, \beta_2 \cdot \mathbf{1}_n, \cdots, \beta_n \cdot \mathbf{1}_n]^\top \in \mathbb{R}^{n \times n}$
- $\widetilde{A} := \exp(QK^\top/\sqrt{d} + \Gamma) \in \mathbb{R}^{n \times n}$.
- $\widetilde{D} := \mathrm{diag}(\widetilde{A}\mathbf{1}_n) \in \mathbb{R}^{n \times n}$.
- $\delta \in (0, 0.1)$.
- *Let* $R \geq 0$ *be defined as Definition C.9.*
- *Given sparsity integer* $k \leq n$.
- *Denote* $T := \exp(\sqrt{\log(n(n-k)d/\delta)})$.
- *Let* $\mathcal{S}_\epsilon$ *be defined as Definition C.3.*

If we choose $\epsilon \geq \frac{T^{O(R)}}{n}$*), then with a probability at least* $1 - \delta$*, we have*

$$\left|\mathcal{S}_\epsilon(D^{-1}_{i_1,i_1} A_{i_1})\right| = \left|\mathcal{S}_\epsilon(\widetilde{D}^{-1}_{i_1,i_1} \widetilde{A}_{i_1})\right| \geq n - k$$

Proof. **Remark.** We re-denote $\mathcal{S}_\epsilon = \mathcal{S}_\epsilon(D^{-1}_{i_1,i_1} A_{i_1,*})$ in the statement, and $i_2 \in \mathcal{S}_\epsilon$.

Following Part 1 of Lemma D.3, with a probability at least $1 - \delta_1$, we have

$$\widetilde{A}_{i_1,i_2} \leq \exp(O(R) \cdot \sqrt{\log((n-k)d/\delta_1)}) \tag{1}$$

Following Part 3 of Lemma D.3, with a probability at least $1 - \delta_2$, we have

$$\widetilde{D}^{-1}_{i_1,i_1} \leq \exp(O(R) \cdot \sqrt{\log(nd/\delta_2)})/n \tag{2}$$

Now we combine Eq. (1) and Eq. (2), with a probability at least $1 - \delta_1 - \delta_2$, we have

$$\begin{aligned}
D^{-1}_{i_1,i_1} A_{i_1,i_2} &= \widetilde{D}^{-1}_{i_1,i_1} \widetilde{A}_{i_1,i_2} \\
&\leq \exp(O(R) \cdot \sqrt{\log((n-k)d/\delta_1)} + O(R) \cdot \sqrt{\log(nd/\delta_2)})/n \\
&\leq \exp(O(R) \cdot \sqrt{\log((n-k)d/\delta_1) + \log(nd/\delta_2)})/n \\
&\leq \exp(O(R) \cdot \sqrt{\log(n(n-k)d^2/(\delta_1\delta_2))})/n \\
&\leq \exp(O(R) \cdot \sqrt{\log(n(n-k)d/\delta)})/n \\
&\leq \exp(\sqrt{\log(n(n-k)/\delta)})^{O(R)}/n \\
&\leq T^{O(R)} \cdot n^{-1}
\end{aligned}$$

where the first step follows from Fact B.1, the second step follows from Eq. (1) and Eq. (2), the third step follows from Fact B.2, the fourth step follows from simple algebras, the fifth step follows from choosing $\delta_1 = \delta_2 = \delta/2$, the sixth step follows from simple algebras, the last step follows from the definition of T. $\square$

D.2 Main Result 2: Attention Collapse

Theorem D.2. *If the following conditions hold:*

- *Let* $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ *be denoted as Query, Key and Value projection matrices of attention.*
- *Given an input* $X \in \mathbb{R}^{n \times d}$ *that holds properties in Definition C.1.*
- *Define Query, Key and Value states matrices* $Q := XW_Q, K := XW_K, V := XW_V \in \mathbb{R}^{n \times d}$.
- $A := \exp(QK^\top/\sqrt{d}) \in \mathbb{R}^{n \times n}$.

- $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n\times n}$.
- *Denote* $\beta_i := Bd\max_{j_1\in[d]} |\mathbb{E}[\sum_{j_2=1}^{d} W_{j_1,j_2}\cdot X_{i,j_2}]|$.
- *Define* $\Gamma := [\beta_1\cdot\mathbf{1}_n, \beta_2\cdot\mathbf{1}_n, \cdots, \beta_n\cdot\mathbf{1}_n]^\top \in \mathbb{R}^{n\times n}$
- $\widetilde{A} := \exp(QK^\top/\sqrt{d}+\Gamma) \in \mathbb{R}^{n\times n}$.
- $\widetilde{D} := \mathrm{diag}(\widetilde{A}\mathbf{1}_n) \in \mathbb{R}^{n\times n}$.
- $\delta \in (0, 0.1)$.
- *Let* $R \geq 0$ *be defined as Definition C.9.*
- *Assuming* $R = o(\sqrt{\log(n)})$.
- *Given sparsity integer* $k \leq n$.
- *Denote* $T := \exp(\sqrt{\log(n(n-k)d/\delta)})$.
- *Let* $\mathcal{S}_\epsilon$ *be defined as Definition C.3.*

For any $\epsilon > 0$*, with probability at least* $1-\delta$*, we have:*

$$\lim_{n\to+\infty} |\mathcal{S}_\epsilon(D_{i,i}^{-1}A_i)| = n-1$$

Proof. In order to choose k that meets the ϵ-approximated sparsity, we have:

$$\epsilon \geq \exp\Big(O(R)\cdot\sqrt{\log(n\cdot(n-k)d/\delta)}\Big)$$

where this step follows from Theorem D.1.

We obtain:

$$k \leq n - \exp\Big(O(\frac{\log^2(\epsilon\cdot n)}{R^2})\Big)\cdot\frac{\delta}{nd} \tag{3}$$

Hence, we have:

$$\begin{aligned}
\lim_{n\to+\infty} |\mathcal{S}_\epsilon(D_{i,i}^{-1}A_i)| \geq & \lim_{n\to+\infty}(n-k) \\
\geq & \lim_{n\to+\infty}\exp\Big(O(\frac{\log^2(\epsilon\cdot n)}{R^2})\Big)\cdot\frac{\delta}{nd} \\
= & \, n-1
\end{aligned}$$

where the first step follows from Definition C.3, the second step follows from Eq.(3), the last step follows from $R = o(\sqrt{\log(n)})$ and $\langle D_{i,i}^{-1}A_i, \mathbf{1}_n\rangle = 1$, then

$$\max|\mathcal{S}_\epsilon(D_{i,i}^{-1}A_i)| = n-1.$$

□

D.3 Bounding D^{-1}

Lemma D.3. *If the following conditions hold:*

- *Let* $W_Q, W_K, W_V \in \mathbb{R}^{d\times d}$ *be denoted as Query, Key and Value projection matrices of attention.*
- *Given an input* $X \in \mathbb{R}^{n\times d}$ *that holds properties in Definition C.1.*
- *Define Query, Key and Value states matrices* $Q := XW_Q, K := XW_K, V := XW_V \in \mathbb{R}^{n\times d}$.

- *Denote* $\beta_i := Bd \max_{j_1 \in [d]} |\mathbb{E}[\sum_{j_2=1}^{d} W_{j_1,j_2} \cdot X_{i,j_2}]|$.
- *Define* $\Gamma := [\beta_1 \cdot \mathbf{1}_n, \beta_2 \cdot \mathbf{1}_n, \cdots, \beta_n \cdot \mathbf{1}_n]^\top \in \mathbb{R}^{n \times n}$
- $\widetilde{A} := \exp(QK^\top/\sqrt{d} + \Gamma) \in \mathbb{R}^{n \times n}$.
- $\widetilde{D} := \mathrm{diag}(\widetilde{A}\mathbf{1}_n) \in \mathbb{R}^{n \times n}$.
- $\delta \in (0, 0.1)$.
- *Let* $R \geq 0$ *be defined as Definition C.9.*

Then with a probability at least $1 - \delta$*, we have*

- *Part 1. For* $i_1, i_2 \in [n]$
$$\exp(-O(R) \cdot \sqrt{\log(d/\delta)}) \leq \widetilde{A}_{i_1,i_2} \leq \exp(O(R) \cdot \sqrt{\log(d/\delta)})$$
- *Part 2. For* $i_1 \in [n]$
$$n \cdot \exp(-O(R) \cdot \sqrt{\log(nd/\delta)}) \leq \widetilde{D}_{i_1,i_1} \leq n \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)})$$
- *Part 2. For* $i_1 \in [n]$
$$\widetilde{D}^{-1}_{i_1,i_1} \leq \exp(O(R) \cdot \sqrt{\log(nd/\delta)})/n$$

Proof. **Proof of Part 1.** We have

$$\begin{aligned} |\widetilde{A}_{i_1,i_2}| =&\ \exp((QK^\top)_{i_1,i_2}) \\ \leq&\ \exp(O(R) \cdot \sqrt{\log(d/\delta)}) \end{aligned}$$

where the first step follows from the definition of A, the second step follows from Lemma D.4.

Proof of Part 2. This proof follows from the union bound of Part 1 of this Lemma and the Definition of D.

Proof of Part 3. This proof follows from the lower bound on D_{i_1,i_1} and simple algebras. ☐

D.4 Concentration on $QK^\top$

Lemma D.4. *If the following conditions hold:*

- *Let* $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ *be denoted as Query, Key and Value projection matrices of attention.*
- *Given an input* $X \in \mathbb{R}^{n \times d}$ *that holds properties in Definition C.1.*
- *Define Query, Key and Value states matrices* $Q := XW_Q, K := XW_K, V := XW_V \in \mathbb{R}^{n \times d}$.
- *Let* $W \in \mathbb{R}^{d \times d}$ *be define as Definition C.8.*
- *Denote* $C \geq 1$ *as a sufficient large constant.*
- $\delta \in (0, 0.1)$.
- *Let* $R \geq 0$ *be defined as Definition C.9.*
- *For* $i_1, i_2 \in [n]$

Then with a probability at least $1 - \delta$*, we have*

$$|(QK^\top)_{i_1,i_2}| \leq O(R) \cdot \sqrt{\log(d/\delta)} + \beta_{i_2}$$

Proof. We have:

$$\begin{aligned}
|(QK^\top)_{i_1,i_2}| &= |(XW_QW_K^\top X^\top/\sqrt{d})_{i_1,i_2}| \\
&= |(XWX^\top)_{i_1,i_2}| \\
&= |X_{i_1}^\top W X_{i_2}| \\
&= |\sum_{j_1=1}^d \sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_1,j_1} X_{i_2,j_2}| \\
&\leq B|\sum_{j_1=1}^d \sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_2,j_2}|
\end{aligned}$$

where the first step follows from $Q := XW_Q, K := XW_K$, the second step follows from Definition C.9, the third and fourth steps follow from simple algebras, the fifth step follows from Definition C.1.

We then apply Hoeffding inequality (Lemma B.5) to each $W_{j_1,j_2} \cdot X_{i_2,j_2}$. Hence, with a probability at least $1 - \delta$, we have:

$$|\sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_2,j_2} - \mathbb{E}[\sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_2,j_2}]| \leq O(B) \cdot \|W_{j_1}\|_2 \cdot \sqrt{\log(d/\delta)}$$

since $|X_{i_2,j_2}| \leq B$ for any $j_2 \in [d]$.

By triangle inequality, we have:

$$|\sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_2,j_2}| \leq O(B) \cdot \|W_{j_1}\|_2 \cdot \sqrt{\log(d/\delta)} + |\mathbb{E}[\sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_2,j_2}]| \tag{4}$$

We obtain:

$$\begin{aligned}
B|\sum_{j_1=1}^d \sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_2,j_2}| &\leq O(B^2) \cdot \|W\|_F \cdot \sqrt{\log(d/\delta)} + Bd \max_{j_1 \in [d]} |\mathbb{E}[\sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_2,j_2}]| \\
&= O(R) \cdot \sqrt{\log(d/\delta)} + \beta_{i_2}
\end{aligned}$$

where the first step follows from Eq 4 and Fact B.2 ($\|W\|_F = \sum_{j_1=1}^d \|W_{j_1}\|_2^2$), the second step follows from Definition C.9 and define

$$\beta_{i_2} := Bd \max_{j_1 \in [d]} |\mathbb{E}[\sum_{j_2=1}^d W_{j_1,j_2} \cdot X_{i_2,j_2}]|$$

$\square$

Remark D.5. *The formal results of Lemma D.4 in the appendix have slight differences with the informal forms, in which we omit the additional terms of each upper bound since such terms are trivially some constants. Fact B.1 shows any constant bias term added before the softmax function will not change the output. We thus simplify the equations for tighter boundaries and more convenient notation.*

E Sparse Attention Approximation

E.1 Main Result 3: Upper Bound on Error

Theorem E.1. *If the following conditions hold:*

- *Let Query, Key and Value states matrices $Q, K, V \in \mathbb{R}^{n \times d}$ be defined as Definition C.2.*
- $A := \exp(QK^\top/\sqrt{d}) \in \mathbb{R}^{n \times n}$.

- $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n \times n}$
- *Let $\mathcal{T}_k$ be defined as Definition C.5.*
- *Let* topk *be defined as Definition C.6.*
- *Let $\mathcal{S}_\epsilon$ be defined as Definition C.3, we omit $\mathcal{S}_\epsilon(D_{i,i}^{-1} A_{i,*})$ for $i \in [n]$ to $\mathcal{S}_{\epsilon,i}$.*
- *Define* $A_{\mathrm{spar}} := [\mathsf{topk}(A_{1,*}) \quad \mathsf{topk}(A_{2,*}) \quad \cdots \quad \mathsf{topk}(A_{n,*})]^\top$.
- $D_{\mathrm{spar}} := \mathrm{diag}(A_{\mathrm{spar}}\mathbf{1}_n) \in \mathbb{R}^{n \times n}$.
- $\delta \in (0, 0.1)$.
- *Let R be defined as Definition C.9.*

Then with a probability at least $1 - \delta$, we have

- **Part 1.** *Choosing $k = \Omega(n^C)$ for $C \in (0,1)$, we have $C_{\mathrm{error}} \in (0, C)$:*
$$\|D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} V - D^{-1} A V\|_\infty \leq o(n^{-C_{\mathrm{error}}})$$
- **Part 2.** *Choosing $k = o(\log(n))$ for $C \in (0,1)$, we have $C_{\mathrm{error}} \in (0, C)$:*
$$\|D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} V - D^{-1} A V\|_\infty \leq \Omega(n^{C_{\mathrm{error}}})$$

Proof. We have

$$\begin{aligned}
\|D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} - D^{-1} A\|_\infty &= \|D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} - D^{-1} A_{\mathrm{spar}} + D^{-1} A_{\mathrm{spar}} - D^{-1} A\|_\infty \\
&\leq \|D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} - D^{-1} A_{\mathrm{spar}}\|_\infty + \|D^{-1} A_{\mathrm{spar}} - D^{-1} A\|_\infty \\
&\leq (\frac{n-k}{nk} + \frac{1}{n}) \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \\
&\leq \frac{1}{k} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \qquad (5)
\end{aligned}$$

where the first step follows from simple algebras, the second step follows from triangle inequality, the third step follows from Part 1 and Part 4 of Lemma E.3, the fourth step follows from simple algebras.

Part 1. Choosing $k = \Omega(n^C)$ for $C \in (0,1)$, we have:

$$\begin{aligned}
\|D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} - D^{-1} A\|_\infty &\leq \frac{1}{k} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \\
&\leq o(n^{-C_{\mathrm{error}}})
\end{aligned}$$

where the first step follows from Eq. (5), the second step follows from $0 < C_{\mathrm{error}} < C$.

Part 2. Choosing $k = o(\log(n))$ for $C \in (0,1)$, we have:

$$\begin{aligned}
\|D_{\mathrm{spar}}^{-1} A_{\mathrm{spar}} - D^{-1} A\|_\infty &\leq \frac{1}{k} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \\
&\leq \Omega(n^{C_{\mathrm{error}}})
\end{aligned}$$

where the first step follows from Eq. (5), the second step follows from $0 < C_{\mathrm{error}}$. ☐

E.2 Lower Bound on Error

Theorem E.2. *If the following conditions hold:*

- *Let Query, Key and Value states matrices $Q, K, V \in \mathbb{R}^{n \times d}$ be defined as Definition C.2.*
- $A := \exp(QK^\top / \sqrt{d}) \in \mathbb{R}^{n \times n}$.
- $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n \times n}$

- *Let $\mathcal{T}_k$ be defined as Definition C.5.*
- *Let* topk *be defined as Definition C.6.*
- *Let $\mathcal{S}_\epsilon$ be defined as Definition C.3, we omit $\mathcal{S}_\epsilon(D_{i,i}^{-1}A_{i,*})$ for $i \in [n]$ to $\mathcal{S}_{\epsilon,i}$.*
- *Define* $A_{\mathrm{spar}} := [\mathsf{topk}(A_{1,*}) \quad \mathsf{topk}(A_{2,*}) \quad \cdots \quad \mathsf{topk}(A_{n,*})]^\top$.
- $D_{\mathrm{spar}} := \mathrm{diag}(A_{\mathrm{spar}}\mathbf{1}_n) \in \mathbb{R}^{n\times n}$.
- $\delta \in (0, 0.1)$.
- *Let R be defined as Definition C.9.*
- *Choosing* $k = o(\log(n))$

Then with a probability at least $1-\delta$, we have

$$\|D_{\mathrm{spar}}^{-1}A_{\mathrm{spar}} - D^{-1}A\|_F \geq O(1)$$

Proof. We have:

$$\begin{aligned}
\|D_{\mathrm{spar}}^{-1}A_{\mathrm{spar}} - D^{-1}A\|_F^2 = & \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}(D_{\mathrm{spar},i_1,i_1}^{-1}A_{\mathrm{spar},i_1,i_2} - D_{i_1,i_1}^{-1}A_{i_1,i_2})^2 \\
\geq & \sum_{i_1=1}^{n}\sum_{i_2\in\mathcal{T}_k}(\frac{1}{k}\exp(-O(R)\cdot\sqrt{\log(\frac{nd}{\delta})}) - \frac{1}{n}\exp(O(R)\cdot\sqrt{\log(\frac{nd}{\delta})}))^2 \\
+ & \sum_{i_1=1}^{n}\sum_{i_2\in[n]/\mathcal{T}_k}(\frac{1}{n}\exp(-O(R)\cdot\sqrt{\log(\frac{nd}{\delta})}))^2 \\
\geq & \sum_{i_1=1}^{n}\sum_{i_2\in\mathcal{T}_k}(O(\frac{1}{\sqrt{n\cdot o(\log(n))}}))^2 \\
\geq & O(1)
\end{aligned}$$

where the first step follows from the definition of Frobenius norm ℓ_F, the second step follows from plugging $k = o(\sqrt{\log(n)})$ and we have:

$$\frac{\mathrm{d}}{\mathrm{d}n}\Big(\frac{1}{k}\exp(-O(R)\cdot\sqrt{\log(\frac{nd}{\delta})}) - \frac{1}{n}\exp(O(R)\cdot\sqrt{\log(\frac{nd}{\delta})})\Big) \geq \frac{\mathrm{d}}{\mathrm{d}n}\frac{1}{\sqrt{n\cdot o(\log(n))}},$$

and the last step follows from simple algebras. $\square$

E.3 APPROXIMATING SOFTMAX FUNCTION

Lemma E.3. *If the following conditions hold:*

- *Let Query, Key and Value states matrices $Q, K, V \in \mathbb{R}^{n\times d}$ be defined as Definition C.2.*
- $A := \exp(QK^\top/\sqrt{d}) \in \mathbb{R}^{n\times n}$.
- $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n\times n}$
- *Let $\mathcal{T}_k$ be defined as Definition C.5.*
- *Let* topk *be defined as Definition C.6.*
- *Let $\mathcal{S}_\epsilon$ be defined as Definition C.3, we omit $\mathcal{S}_\epsilon(D_{i,i}^{-1}A_{i,*})$ for $i \in [n]$ to $\mathcal{S}_{\epsilon,i}$.*
- *Define* $A_{\mathrm{spar}} := [\mathsf{topk}(A_{1,*}) \quad \mathsf{topk}(A_{2,*}) \quad \cdots \quad \mathsf{topk}(A_{n,*})]^\top$.
- $D_{\mathrm{spar}} := \mathrm{diag}(A_{\mathrm{spar}}\mathbf{1}_n) \in \mathbb{R}^{n\times n}$.

- $\delta \in (0, 0.1)$.
- *Let $R \geq 0$ be defined as Definition C.9.*

Then with a probability at least $1 - \delta$, we have

- *Part 1.*
$$\|D^{-1} A_{\rm spar} - D^{-1} A\|_\infty \leq \frac{1}{n} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)})$$
- *Part 2.*
$$\|D_{\rm spar} - D\|_\infty \leq (n-k) \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)})$$
- *Part 3.*
$$\|D_{\rm spar}^{-1} - D^{-1}\|_\infty \leq \frac{n-k}{nk} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)})$$
- *Part 4.*
$$\|D_{\rm spar}^{-1} A_{\rm spar} - D^{-1} A_{\rm spar}\|_\infty \leq \frac{n-k}{nk} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)})$$

Proof. Before we begin the proof, we construct a toolkit as follows:

For $x_1 \leq \exp(\sqrt{\log(a/\delta_1)})$ and $x_2 \leq \exp(\sqrt{\log(a/\delta_2)})$, we have

$$\begin{aligned} x_1 x_2 \leq &~ \exp(\sqrt{\log(a/\delta_1)}) \cdot \exp(\sqrt{\log(b/\delta_2)}) \\ \leq &~ \exp(\sqrt{\log(a/\delta_1)} + \sqrt{\log(b/\delta_2)}) \\ \leq &~ \exp(C\sqrt{\log(a/\delta_1) + \log(b/\delta_2)}) \\ \leq &~ \exp(C\sqrt{\log(ab/\delta)}) \end{aligned} \tag{6}$$

where these steps follow from simple algebras, Fact B.3 and choose $\delta_1 = \delta_2 = \delta/2$.

Proof of Part 1. This proof follows from Theorem D.1 and $n - k \leq n$.

Proof of Part 2. We have

$$\begin{aligned} \|D_{\rm spar} - D\|_\infty = &~ \|A_{\rm spar} \mathbf{1}_n - A \mathbf{1}_n\|_\infty \\ = &~ \|D \circ (D^{-1} A_{\rm spar} \mathbf{1}_n - D^{-1} A \mathbf{1}_n)\|_\infty \\ \leq &~ \|D\|_\infty \cdot \|D^{-1} A_{\rm spar} \mathbf{1}_n - D^{-1} A \mathbf{1}_n\|_\infty \\ \leq &~ \|D\|_\infty \cdot \frac{n-k}{n} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \\ \leq &~ (n-k) \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \\ \leq &~ (n-k) \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \end{aligned}$$

where the first step follows from the definitions of $D_{\rm spar}$ and D, the second step follows from simple algebras, the third step follows from Cauchy-Schwarz inequality, the fourth step follows from Part 1 of this Lemma and the definition of $A_{\rm spar}$, the fifth step follows from Part 4 of Lemma E.4, the sixth step follows from Eq. (6).

Proof of Part 3. We have

$$\begin{aligned} \|D_{\rm spar}^{-1} - D^{-1}\|_\infty = &~ \|D_{\rm spar}^{-1}\|_\infty \cdot \|D^{-1}\|_\infty \cdot \|D_{\rm spar} - D\|_\infty \\ \leq &~ \|D_{\rm spar}^{-1}\|_\infty \cdot \|D^{-1}\|_\infty \cdot (n-k) \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \\ \leq &~ \frac{n-k}{nk} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \\ \leq &~ \frac{n-k}{nk} \cdot \exp(O(R) \cdot \sqrt{\log(nd/\delta)}) \end{aligned}$$

where the first step follows from simple algebras, the second step follows from Part 2 of this Lemma, the third step follows from Part 5 and Part 6 of Lemma E.4, the last step follows from Eq. (6).

Proof of Part 4. This proof follows from combining Part 2 of Lemma E.4 and Part 3 of this Lemma. $\square$

E.4 HELPFUL BOUND TOOLKIT

Lemma E.4. *If the following conditions hold:*

- *Let Query, Key and Value states matrices $Q, K, V \in \mathbb{R}^{n\times d}$ be defined as Definition C.2.*
- $A := \exp(QK^\top/\sqrt{d}) \in \mathbb{R}^{n\times n}$.
- $D := \mathrm{diag}(A\mathbf{1}_n) \in \mathbb{R}^{n\times n}$
- *Let $\mathcal{T}_k$ be defined as Definition C.5.*
- *Let* topk *be defined as Definition C.6.*
- *Let $\mathcal{S}_\epsilon$ be defined as Definition C.3, we omit $\mathcal{S}_\epsilon(D_{i,i}^{-1}A_{i,*})$ for $i \in [n]$ to $\mathcal{S}_{\epsilon,i}$.*
- *Define* $A_{\mathrm{spar}} := [\mathsf{topk}(A_{1,*}) \quad \mathsf{topk}(A_{2,*}) \quad \cdots \quad \mathsf{topk}(A_{n,*})]^\top$.
- $D_{\mathrm{spar}} := \mathrm{diag}(A_{\mathrm{spar}}\mathbf{1}_n) \in \mathbb{R}^{n\times n}$.
- $\delta \in (0, 0.1)$.
- *Let $R \geq 0$ be defined as Definition C.9.*

Then with a probability at least $1-\delta$, we have

- *Part 1.* $\exp(-O(R)\cdot\sqrt{\log(\frac{n-k}{\delta}d)}) \leq A_{i_1,i_2} \leq \exp(O(R)\cdot\sqrt{\log(\frac{n-k}{\delta}d)})$, $\forall i_1 \in [n], i_2 \in \mathcal{S}_\epsilon$
- *Part 2.* $\exp(-O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)}) \leq A_{i_1,i_2} \leq \exp(O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)})$, $\forall i_1, i_2 \in [n]$
- *Part 3.* $k\exp(-O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)}) \leq \sum_{i_2\in\mathcal{T}_k} A_{i_1,i_2} \leq k\exp(O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)})$, $\forall i_1 \in [n]$
- *Part 4.* $n\exp(-O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)}) \leq D_{i_1,i_1} \leq n\exp(O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)})$, $\forall i_1 \in [n]$
- *Part 5.* $\frac{1}{k}\exp(-O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)}) \leq \Big(\sum_{i_2\in\mathcal{T}_k} A_{i_1,i_2}\Big)^{-1} \leq \frac{1}{k}\exp(O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)})$, $\forall i_1 \in [n]$
- *Part 6.* $\frac{1}{n}\exp(-O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)}) \leq D_{i_1,i_1}^{-1} \leq \frac{1}{n}\exp(O(R)\cdot\sqrt{\log(\frac{n}{\delta}d)})$, $\forall i_1 \in [n]$

Proof. **Proof of Part 1.**

This proof follows from Eq. (1).

Proof of Part 2.

This proof follows from Part 1 and Part 2 of Lemma D.3.

Proof of Part 3.

This proof follows from Part 1 of this Lemma.

Proof of Part 4.

This proof follows from Part 2 of Lemma D.3.

Proof of Part 5.

This proof follows from Part 3 of this Lemma.

Proof of Part 6. This proof follows from Part 4 of this Lemma. □