A Spectral Perspective on Generalization in Transformers

Paul Lintilhac
Thayer School of Engineering
Dartmouth College
paul.s.lintilhac.th@dartmouth.edu
Sair Shaikh
Dartmouth College
{sair.shaikh.26}@dartmouth.edu
Michael Hahn
Saarland University
{mhahn}@lst.uni-saarland.de

Abstract

We study transformers' generalization behavior on boolean domains from the perspective of the Fourier Spectra of their target functions. In contrast to prior work [12, 26], which derived generalization bounds from Rademacher complexity, we investigate the feasibility of obtaining generalization bounds via PAC-Bayes theory. We show that sparse spectra concentrated on low-degree components enable low-sharpness constructions with good generalization properties. Our idea is to show the existence of flat minima implementing any boolean function of sparsity no greater than the context length, and then apply a PAC-Bayes bound to an idealized low-sharpness learner, resulting in a non-vacuous generalization bound. We evaluate predictions empirically and conduct a mechanistic interpretability study to support the realism of our theoretical construction in real transformers.

While various authors have studied the expressive capacity of transformers to express different classes of functions [e.g. 14, 6, 20, 25, 23, 19, 20, 24], their learning biases still are incompletely understood. Edelman et al. [12], Trauger & Tewari [26] showed that transformers with fixed positional encodings have a relatively small statistical complexity. Other work showed a low-sensitivity and low-degree biases in Transformers [7, 16, 1].

The present paper provides a generalization result for transformers learning low-degree functions, under an idealized low-sharpness learner. In contrast to prior work on capacity-based bounds grounded in parameter norms [12, 26], we explore the feasibility of using PAC-Bayes methods grounded in the existence of flat minima. Any function on a Boolean domain, $f : \{0,1\}^T \to \mathbb{R}$, possesses a unique representation in terms of parity functions:

$$f(x) = \sum_{A \subseteq [1,T]} \lambda_A \cdot \chi_A(x) \tag{1}$$

where $\chi_A(x) = (\sum_{i \in A} x_i) \bmod 2$. The *degree* of f is $D_f := \max\{|A| : \lambda_A \neq 0\}$. Re-expressing the above sum (1) as $f(x) = \sum_{t=1}^{\omega} c_t \chi_{S_t}$, ω is the number of nonzero coefficients (the spectral *sparsity*), $c_t \in \mathbb{R}$, $S_t \subseteq [T]$ are the indices of the t^{th} Fourier component, and $|S_t| \leq D_f$. For simplicity, we take all Fourier components to have the same degree, i.e. $|S_t| = D_f \, \forall t \in [\omega]$. A function $f : \{0,1\}^T \to \mathbb{R}$ of low degree, and having $\omega \leq T$ nonzero Fourier components, can be learned by a 1.5-layer (2 attention+1 MLP sublayers), 1-head transformer. Our bound applies to this function class.

Theoretical Results Empirical evidence indicates that stochastic gradient descent (SGD) tends to find solutions with maximal parameter flatness on the training dataset, especially with smaller batch sizes [17]. Sharpness-aware minimization has now been accepted as a powerful tool for regularization [22], while PAC-Bayes theory has used sharpness as a key term in some of the first tight generalization bounds for deep learning [21]. Abstracting away from training dynamics, we study a general learner that, through implicit bias or explicit regularization, finds an interpolator that minimizes parameter norms and the sharpness of the loss:

Definition 1 (Idealized Low-Sharpness Learner)**.** *Given a space of possible parameters Θ for transformers, and a finite training set for a function $f : \{0,1\}^T \to \mathbb{R}$, find parameters Θ minimizing*

$\hat{L}(f_{\mathbf{\Theta}}) + \alpha\|\Theta\| + \beta \operatorname{Tr}\left(\nabla^2\left[\hat{L}(f_\Theta)\right]\right)$, *where* $\hat{L}(f_{\mathbf{\Theta}})$ *is the training loss, and* α, β *are parameters that control the aversion to high parameter norm and sharpness, respectively.*

Our main theorem is a generalization bound for this learner.

Theorem 2. *Let* f *be a target function, with degree* D_f *and sparsity* $\omega \leq T$*. There are* $\alpha, \beta \geq 0$ *such that the following holds. Let* $\hat{\mathbf{\Theta}}$ *be the solution returned by the general learning procedure on the training set of size* m*. Let* $\sigma^2 > 0$*. Let* $L_f(\hat{\Theta})$ *be the global loss, and assume that our learner has achieved* 0 *training loss,* $\hat{L}(f_{\hat{\mathbf{\Theta}}}) = 0$*. Suppose further that for some constant* $\Sigma > 0$*, our loss function satisfies:* $\mathbb{E}_X\left[e^{t[L(X,f_{\hat{\Theta}}) - \mathbb{E}_X[L(X,f_{\hat{\Theta}})]]}\right] \leq e^{\frac{\Sigma^2 t^2}{2}}$*. For each* $\lambda > 0$*, with probability* $\geq 1 - \delta$*,*

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[L(f_{\hat{\Theta}+\epsilon})] \leq \underbrace{\sigma^2\Big(O(\omega D_f^3) + \frac{1}{2}o\Big(\sigma D_f^5\omega^2\big(\log^{\frac{7}{2}}(T) + D_f\log^{\frac{5}{2}}(T)\big)\Big)\Big)}_{\textit{Perturbed Sharpness Term}}$$

$$+ \underbrace{2\sqrt{\frac{\Sigma^2}{2m}\Big(\frac{O\Big(D_f^3 + \log(T)^2(\omega D_f + T - \omega)\Big)}{2\sigma^2} + ln\frac{1}{\delta}\Big)}}_{\textit{Parameter Norm Term}}$$

Our bound is non-vacuous: as long as m grows asymptotically fast enough, the norm term will be $o(1)$, and we are free to choose σ small to make everything $o(1)$, yielding an overall non-vacuous bound. The proof proceeds via PAC-Bayes theory. We first construct a transformer with small parameter norm and low sharpness. PAC-Bayes theory implies that learned minima with such properties will generalize well. As $\hat{\Theta}$ is a low-sharpness interpolator of the target function, the (unknown) norm and sharpness of our interpolator $\hat{\Theta}$ can be bounded by the (known) norm and the sharpness of the exact transformer Θ, leading to our bound. See App. C for full exposition, and D.2 for an empirical validation of our "low sharpness interpolator" assumption. Experiments with small transformer models showed behavior similar to the construction (App. D.3). Our result also applies to Chain-of-Thought (CoT), providing another perspective on its effectiveness: For PARITY, there is a CoT using a function of degree 2; we get a much better generalization bound for the CoT than for a transformer performing it without CoT. See App. E for more.

Implications Recent work shows empirical and theoretical links between degree and generalization of transformers [7, 1, 15]. Our result uses these properties in a generalization bound, supporting the degree as a complexity measure for transformers. Our work also introduces the Fourier sparsity as a secondary factor in generalization. To our knowledge, this is the first generalization bound that relies on explicitly upper bounding the gradients and hessian of a given class of transformers.

In H.7, we discuss the cases in which our approach could yield a tighter guarantee than a purely norm-based approach [e.g. 12], noting that the key limitation to our bound is an outsized perturbation term, which can make our bound less efficient than the norm-based bounds. We suggest possible ways to address this challenge in the Limitations section F.3.

We believe our overall approach might offer a promising path to explaining generalization for functions for which there is a known canonical transformer construction. For other kinds of algorithmic problems, one could use a canonical transformer construction that solves the problem (using e.g. the RASP program for the task [27]), and then analyze its norms and gradients. This might lead to more function-specific bounds than what is possible using the existing covering-number-based bounds [12, 26]. We leave exploration of this idea to future work.

Discussion We have shown that transformers represent Boolean functions with low degree and low sparsity in flat minima, and deduced a generalization bound via PAC-Bayes methods. This result expands the understanding of the learning abilities and biases of the transformer architectures. Our bound could be instantiated in practice without a priori knowing the target boolean function that has been learned by the transformer. This can be done by sampling inputs and outputs and directly estimating the Fourier decomposition of the function using Monte Carlo methods or Random Fourier Features. These methods may still be considered expensive, but it is worth noting that both of these properties (Fourier degree and sparsity) have known efficient property testing algorithms [3, 13], and therefore can be upper bounded with high probability with relatively few input-output samples.

References

[1] Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31–60. PMLR, 2023. URL https://proceedings.mlr.press/v202/abbe23a.html.

[2] Emmanuel Abbé, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? the globality barrier and inductive scratchpad. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FoGwiFXzuN.

[3] Noga Alon, Tali Kaufman, Michael Krivelevich, Simon Litsyn, and Dana Ron. *Testing low-degree polynomials over GF(2)*, pp. 188–199. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag, Germany, 2003. ISBN 3540407707. doi: 10.1007/978-3-540-45198-3_17. Copyright: Copyright 2020 Elsevier B.V., All rights reserved.

[4] Pierre Alquier. A user-friendly introduction to pac-bayes bounds. 2023.

[5] Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35: 38546–38556, 2022.

[6] Satwik Bhattamishra, Kabir Ahuja, and Navin Goyal. On the ability and limitations of transformers to recognize formal languages. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7096–7116. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.576. URL https://doi.org/10.18653/v1/2020.emnlp-main.576.

[7] Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. Simplicity bias in transformers and their ability to learn sparse boolean functions. *arXiv preprint arXiv:2211.12316*, 2022.

[8] Vivien Cabannes, Charles Arnal, Wassim Bouaziz, Xingyu Alice Yang, Francois Charton, and Julia Kempe. Iteration head: A mechanistic study of chain-of-thought. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=QBCxWpOt5w.

[9] David Chiang and Peter Cholak. Overcoming a theoretical limitation of self-attention. *arXiv preprint arXiv:2202.12172*, 2022.

[10] Puneesh Deora, Rouzbeh Ghaderi, Hossein Taheri, and Christos Thrampoulidis. On the optimization and generalization of multi-head attention. *Transactions on Machine Learning Research*, 2023.

[11] Irit Dinur*, Yuval Filmus†, and Prahladh Harsha‡. Low degree almost boolean functions are sparse juntas, 2021.

[12] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

[13] Parikshit Gopalan, Ryan O'Donnell, Rocco A. Servedio, Amir Shpilka, and Karl Wimmer. Testing fourier dimensionality and sparsity. *SIAM Journal on Computing*, 40(4):1075–1100, 2011. doi: 10.1137/100785429. URL https://doi.org/10.1137/100785429.

[14] Michael Hahn. Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171, 2020.

[15] Michael Hahn and Mark Rofin. Why are sensitive functions hard for transformers? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14973–15008, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.800. URL https://aclanthology.org/2024.acl-long.800/.

[16] Michael Hahn and Mark Rofin. Why are sensitive functions hard for transformers? In *Proceedings of the 2024 Annual Conference of the Association for Computational Linguistics (ACL 2024)*, 2024. arXiv Preprint 2402.09963.

[17] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016. URL http://arxiv.org/abs/1609.04836.

[18] Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *NeurIPS 2024 Workshop on Mathematics of Modern Machine Learning*, 2024. URL https://openreview.net/forum?id=E7HwPhfX1B.

[19] Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnahttps://arxiv.org/help/api/indexmurthy, and Cyril Zhang. Transformers learn shortcuts to automata, 2023. URL https://arxiv.org/abs/2210.10749.

[20] William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023.

[21] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[22] Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8rvqpiTTFv.

[23] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Representational strengths and limitations of transformers. *CoRR*, abs/2306.02896, 2023. doi: 10.48550/ARXIV.2306.02896. URL https://doi.org/10.48550/arXiv.2306.02896.

[24] Clayton Sanford, Daniel Hsu, and Matus Telgarsky. Transformers, parallel computation, and logarithmic depth. In *Forty-first International Conference on Machine Learning*, 2024.

[25] Lena Strobl, William Merrill, Gail Weiss, David Chiang, and Dana Angluin. Transformers as recognizers of formal languages: A survey on expressivity. *CoRR*, abs/2311.00208, 2023. doi: 10.48550/ARXIV.2311.00208. URL https://doi.org/10.48550/arXiv.2311.00208.

[26] Jacob Trauger and Ambuj Tewari. Sequence length independent norm-based generalization bounds for transformers. In *International Conference on Artificial Intelligence and Statistics*, pp. 1405–1413. PMLR, 2024.

[27] Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *International Conference on Machine Learning*, pp. 11080–11090. PMLR, 2021.

A Appendix

Contents

B Detailed Exposition of Setup

We can re-express the above sum (1) as

$$f(x) = \sum_{t=1}^{\omega} c_t \chi_{S_t} \tag{2}$$

where ω is the number of nonzero coefficients (the spectral *sparsity*), $c_t \in \mathbb{R}$, $S_t \subseteq [T]$ are the indices of the t^{th} fourier component, and $|S_t| \leq D_f$.

For simplicity, we assume that all Fourier components have the same degree, i.e. $|S_t| = D_f \, \forall t \in [\omega]$. We claim that a transformer representing a function $f : \{0,1\}^T \to \mathbb{R}$ of low degree, and having $\omega \leq T$ nonzero Fourier components, can be learned by a 1.5-layer (2 attention+1 MLP sublayers), 1-head transformer. Let $x_t \in \mathbb{R}^{d+1}$ be the embedding at the t^{th} token position. The total hidden dimension includes d purely positional dimensions as well as one extra dimension holding the bit value. The input matrix $X \in \mathbb{R}^{(T+1)\times(d+1)}$ then has the vector x_t as its t^{th} row, with a special CLS token at the end.

The attention score between locations i and j in the first layer is given by

$$a_{i,j} = x_i^T W^{(1)} x_j, i, j \in [T],$$

where $W^{(1)} \in \mathbb{R}^{(d+1)\times(d+1)}$ is the combined key and query projection matrices. The softmax score is given by

$$\hat{a}_{i,j} = \frac{\exp(a_{i,j})}{\sum_s \exp(a_{i,s})}$$

The output of the attention layer is then given by

$$b_t = \sum_{j=1}^{T} \hat{a}_{t,j} (V^{(1)})^T x_j,$$

where $V^{(1)} \in \mathbb{R}^{(d+1)\times(d+1)}$ is our value projection. The activations after the attention layers are the result of applying the two-layer MLP with a ReLU with a skip-connection:

$$g_t = b_t + f^{MLP}\big(b_t\big) = b_t + F^T(M^T b_t + \Gamma)_+$$

Note that our construction does not include the layer norm. We then use a second attention layer parametrized by $W^{(2)} \in \mathbb{R}^{(d+1)\times(d+1)}, V^{(2)} \in \mathbb{R}^{d\times 1}$, such that the output is read from the final activation of the CLS token, i.e.

$$\mathcal{T}(X, W^{(1)}, V^{(1)}, M, F, \Gamma, W^{(2)}, V^{(2)}) = (V^{(2)})^T G^T \phi\Big(G(W^{(2)})^T g_{T+1}\Big),$$

where $G \in \mathbb{R}^{(T+1)\times(d+1)}$ is a matrix with rows equal to $g_t, \forall t \in [T+1]$.

C Detailed Exposition of Theoretical Results

Our theoretical results are centered around the relationship between generalization and sharpness. Empirical evidence that stochastic gradient descent (SGD) tends to find solutions with maximal parameter flatness on the training dataset, especially with smaller batch sizes [17]. Sharpness-aware

minimization has now been accepted as a powerful tool for regularization [22], while PAC-Bayes theory has used sharpness as a key term in some of the first tight generalization bounds for deep learning [21].

We study the general setup of a learner minimizing loss, sharpness, and norm. Abstracting away from training dynamics, we study a general learner that, through implicit bias or explicit regularization, finds an interpolator that minimizes parameter norms and the sharpness of the loss:

Definition 3 (Idealized Low-Sharpness Learner)**.** *Given a space of possible parameters Θ for transformers, and a finite training set for a function $f : \{0,1\}^T \to \mathbb{R}$, find parameters Θ minimizing*

$$\hat{L}(f_{\mathbf{\Theta}}) + \alpha\|\Theta\| + \beta Tr\Big(\nabla^2\big[\hat{L}(f_\Theta)\big]\Big), \tag{3}$$

where $\hat{L}(f_{\mathbf{\Theta}})$ is the training loss, and α, β are parameters that control the aversion to high parameter norm and sharpness, respectively.

Our main theorem will be a generalization bound for this learner.

Theorem 4. *Let f be a target function, with degree D_f and sparsity $\omega \leq T$. There are $\alpha, \beta \geq 0$ such that the following holds. Let $\hat{\mathbf{\Theta}}$ be the solution returned by the general learning procedure on the training set of size m. Let $\sigma^2 > 0$. Let $L_f(\Theta)$ be the global loss, and assume that our learner has achieved 0 training loss, $\hat{L}(f_{\mathbf{\Theta}}) = 0$ Suppose further that for some constant $\Sigma > 0$, our loss function satisfies:*

$$\mathbb{E}_X\left[e^{t[L(X,f_{\hat{\Theta}}) - \mathbb{E}_X[L(X,f_{\hat{\Theta}})]]}\right] \leq e^{\frac{\Sigma^2 t^2}{2}}.$$

For each $\lambda > 0$, with probability at least $1-\delta$, the following bound holds:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[L(f_{\Theta+\epsilon})] \leq \frac{\sigma^2}{2}Tr\Big(\nabla^2\big[\hat{L}(f_{\Theta+\zeta})\big]\Big) + 2\sqrt{\frac{\Sigma^2}{2m}\Big(\frac{\|\hat{\Theta}\|^2}{2\sigma^2} + ln\frac{1}{\delta}\Big)}$$

$$\leq \underbrace{\sigma^2\Big(G_u(\omega, D_f) + \frac{1}{2}P(\sigma,\omega,D_f,T)\Big)}_{\textit{Perturbed Sharpness Term}} + \underbrace{2\sqrt{\frac{\Sigma^2}{2m}\Big(\frac{L(\omega,D_f,T)}{2\sigma^2} + ln\frac{1}{\delta}\Big)}}_{\textit{Parameter Norm Term}}$$

where

$$G_u(\omega, D_f) \in O(\omega D_f^3),$$

$$P(\sigma,\omega,D_f,T) \in o\Big(\sigma D_f^5\omega^2\big(log^{\frac{7}{2}}(T) + D_f log^{\frac{5}{2}}(T)\big)\Big)$$

$$L(\omega,D_f,T) \in O\Big(D_f^3 + log(T)^2(\omega D_f + T - \omega)\Big)$$

Discussion To show that the bound is non-vacuous, we note the following. The perturbed sharpness term above does not have any polynomial dependence on T, one of the key challenges overcome in our proof. The only polynomial dependency on T in the above bound is the T dependency in the parameter norm bound, $L(\omega, D_f, T) \in O(D_f^3 + log(T)^2(\omega D_f + T - \omega) \in o(T^3 log(T)^2)$. As long as m grows asymptotically faster than this, the norm term will be $o(1)$, and we are free to choose σ to be small enough to make the perturbed sharpness term also in $o(1)$, yielding an overall non-vacuous bound. Figure 1 shows a concrete instantiation of our bound that is non-vacuous. The generalization gap predicted in this plot used a sample complexity of $m = 10^{15}$. While this is large, we note that the number of possible strings of length 50 is 2^{50}. Since $10^{15} = 2^{49} \leq 2^{50}$, our bound is also non-trivial for input strings of length at least $T \geq 50$, in the sense that it improves on the sample complexity needed for perfect function memorization. We note that the polynomial bound on sample complexity results whenever $\omega \leq T$.[1] This is a much larger class than the class of k-sparse functions.

The majority of the sample complexity in our bound comes from the presence of the perturbation term, $P(\sigma)$, which carries higher degrees of the variables D_f, ω, d and much larger constants than the term representing the unperturbed Hessian, G_u. The presence of this term forces σ to be smaller, and m to be larger, than our bound otherwise would. We note that the G_u term is in a sense more fundamental than $P(\sigma)$, because it represents the curvature of our exact construction, which can be

[1]We believe that an extension of our construction would provide this for any ω polynomial in T.

bounded fairly tightly. On the other hand, our bounds on $P(\sigma)$ involve several imperfect techniques, which are discussed in the limitations section. This indicates that our overall approach could lead to much tighter bounds with far smaller sample complexity, if the analysis of perturbations were tightened. We refer the reader to H.6for more details on the outsized impact of $P(\sigma)$ in our bound.

Proof Strategy The proof of this result proceeds via PAC-Bayes theory. Our proof proceeds in two stages. The first part is to show the existence of a construction with a small parameter volume and low sharpness. PAC-Bayes theory then tells us that learned minima with such properties will possess good generalization properties. The size of the perturbation σ controls the trade-off between the parameter norms and the sharpness of the loss landscape. When combined with our assumption that $\hat{\Theta}$ is a low-sharpness interpolator of our exact construction, Θ, the (unknown) norm and sharpness of our interpolator $\hat{\Theta}$ can be replaced by the (known) norm and the sharpness of the exact transformer Θ, for which the norm and sharpness have been bounded explicitly.

C.1 Step 1: Existence of a "Good" Construction

We specify a simple transformer construction for Boolean functions, and show that – for bounded degree and sparsity – it has bounded norm and sharpness. Our first theoretical results show that the trace of the loss Hessian for our construction is bounded by a function that increases polynomially in both the maximum degree, D_f, and the Fourier sparsity, ω. The basic intuition behind this result is that when our construction implements a higher-degree function, the MLP must interpolate a function of higher frequency but equal amplitude, and will therefore carry a larger derivative with respect to changes in the parameters. As a first step, we bound the trace of the loss hessian in terms of the norms of the gradient.

Theorem 5. *Let f be a target boolean function of sparsity $\omega \leq T$ and maximum degree D_f, and let $\mathcal{T}(X, \Theta)$ be a transformer of context length T implementing it exactly according to our construction. Let $L_f(X, \Theta) = (T(\Theta, X) - f(X))^2$ be the unbounded, quadratic loss for our transformer learning the function f evaluated at input X. Define $L(f_{\mathbf{\Theta}}) := \mathbb{E}\big[L_f(X, \Theta)\big]$ to be the global loss averaged over all possible bit strings of length T. Then under the quadratic loss, the trace of the loss Hessian is bounded by:*

$$Tr\Big(\nabla^2 L(f_{\mathbf{\Theta}})\Big) = 2||\nabla \mathcal{T}(X, \Theta)||^2 \leq 2G_u(\omega, D_f) \in O(\omega D_f^3)$$

We defer a full proof to 11 in the Appendix. Here, we provide an outline of the construction. We start with a simple transformer construction that approximates functions on Boolean domains, based on their Fourier-Walsh transforms. The activations at the output of the first (attention + MLP) layer contain the values of each non-zero Fourier component χ_{S_t} in one of the coordinates. From there, our second attention layer will take a linear combination of these components with weights c_t.

In order to calculate the value of each Fourier component χ_{S_t} at each position in the first layer, we will make use of a purely position-aware attention mechanism with $O(log(T))$ scaling. This trick allows us to overcome a theoretical limitation of transformers resulting from the softmax giving significant weight to even inactive positions, which becomes more pronounced for large T [14, 9]. For each position this puts weights of approximately $\frac{1}{D_f}$ on all positions $j \in S_t$, and 0 elsewhere. $V^{(1)} \in \mathbb{R}^{(d+1)\times(d+1)}$ is a projection matrix that multiplies the attention weights with the bit values and stores the normalized component sums ($\frac{1}{D_f}\sum_{j\in S_t} x_j$) in the final, $d + 1^{th}$ dimension.

After computing this normalized prefix sum and storing it in each of the activations after the first attention layer, we then apply the MLP to approximate the "mod 2" function. To do so minimally, we employ the first layer MLP matrix, $M \in \mathbb{R}^{d\times 4(D_f+1)}$, which together with our bias term $\Gamma \in \mathbb{R}^{d\times 4(D_f+1)}$ acts as a a set of indicators for each unique value in our grid of possible prefix sums. Then our second-layer MLP matrix $F \in \mathbb{R}^{4(D_f+1)\times d}$ linearly combines these indicators with the correct memorized function values. [2]. Finally, we leverage a second attention layer whose sole purpose is to linearly combine the parity of each component stored in the first ω positions of the final activation, using the Fourier-Walsh weights c_t.

[2]This assumption about the MLP function being a periodic trapezoidal wave to interpolate the Mod 2 function, as opposed to for example a triangular wave, is key to our ability to limit the probability of any neuron in the MLP flipping from being active to inactive, thereby simplifying the analysis

The following theorem approximates the error of the perturbed transformer on a test point, assuming the model has achieved 0 error on the training set.

Our next result upper bounds the trace term on the right through a series of steps. First, we bound the trace of the hessian of the loss with two major terms: the unperturbed hessian and the perturbed Hessian. As shown in 5, the unperturbed Hessian is controlled by the norm of the gradient. The perturbed Hessian term itself consists of two major terms: the norm of the gradient of the transformer evaluated at the perturbed weights (bounded in 28), and a term involving a product of the loss of the perturbed transformer and the trace of the hessian of the perturbed transformer. Bounds on these are derived in 22 and 29, respectively. Next, we upper bound the trace of the perturbed transformer's Hessian using a bound on its maximum eigenvalue, which itself is upper bounded in 29. While this incurs a factor of $|\Theta|$, which may seem large, by modifying our basic construction to use random projections we can reduce the dimensions needed for our construction to $O(\log(T))$, thereby reducing the total parameter count to $O(D_f \log(T) + log(T)^2)$.

Theorem 6. *Given our perturbed transformer $\mathcal{T}(X, \Theta + \zeta)$, we can upper bound the trace of the hessian as follows:*

$$Tr\Big(\nabla^2\big[L(f_{\Theta+\zeta})\big]\Big) \leq 2G_u(\omega, D_f) + P(\sigma)$$

where

$$P(\sigma) = 2G_p(\sigma, \omega, D_f, T)\Big(2G_u(\omega, D_f) + G_p(\sigma, \omega, D_f, T)\Big)$$

$$+T_p(\sigma, \omega, D_f, T)\Big(H_u(\omega, D_f, T) + H_p(\sigma, \omega, D_f, T)\Big)|\Theta|$$

The proof is given in Theorem 13. We stress that the need to evaluate the hessian of the loss of the transformer at a perturbed point arises from the fact that we are making a second-order Taylor approximation to the sharpness. While we could have simply written the trace of our exact construction plus an error term that is $O(\sigma^2)$, this would have lacked rigor, as it would potentially bury hidden factors in ω, D_f, T that could be significant. Indeed, our explicit calculation of these perturbation remainder terms shows that without a higher-order approximation, the remainder term is significant unless σ is extremely small compared to ω, D_f, and T.

C.2 Step 2: PAC-Bayes

Our construction provides an interpolator Θ with 0 training loss for some particular values of $\|\Theta\|, Tr\Big(\nabla^2\big[\hat{L}(f_\Theta)\big]\Big)$. For some setting of the regularization parameters α, β, one can thus find a interpolator $\hat{\Theta}$ for which parameter norms and sharpness will both be upper-bounded by those of Θ: $\hat{L}(f_{\boldsymbol{\Theta}}) = 0, Tr\Big(\nabla^2\big[\hat{L}(f_{\hat{\boldsymbol{\Theta}}})\big]\Big) \leq Tr\Big(\nabla^2\big[\hat{L}(f_{\boldsymbol{\Theta}})\big]\Big), ||\hat{\Theta}|| \leq \|\Theta\|$ Thus, our construction provides information about the outcome of this general learner even if it finds an entirely different construction. To show that $\hat{\Theta}$ has good generalization properties, we can leverage the machinery of PAC-Bayes theory. PAC-Bayes theory considers the weights learned during training to be a data-dependent distribution (referred to as the "posterior", denoted by $\mathcal{Q}$), which is compared against the initial distribution of weights (referred to as the "prior", denoted by $\mathcal{P}$) using the KL divergence to bound the generalization gap. Our specific PAC-Bayes bound builds upon a variant due to [4]. Our approach is to use the assumed low-sharpness interpolator of our construction as the center of the posterior distribution. This version of the PAC-Bayes bound can be expressed in such a way that it includes a term representing the parameter sharpness on the training set. We further adapt our bound to the case of a quadratic loss, using the assumption that the test losses are sub-gaussian. [3] For a detailed derivation of our Oracle PAC-Bayes bound, see H.6 in the appendix.

[3]We note that our PAC-Bayes bound is of the type where the bound on the sharpness is not data-dependent, but rather dependent on other assumptions about the complexity of the target function and the learning algorithm. These PAC-Bayes bounds are referred to by e.g. [4] as "Oracle PAC-Bayes Bounds"

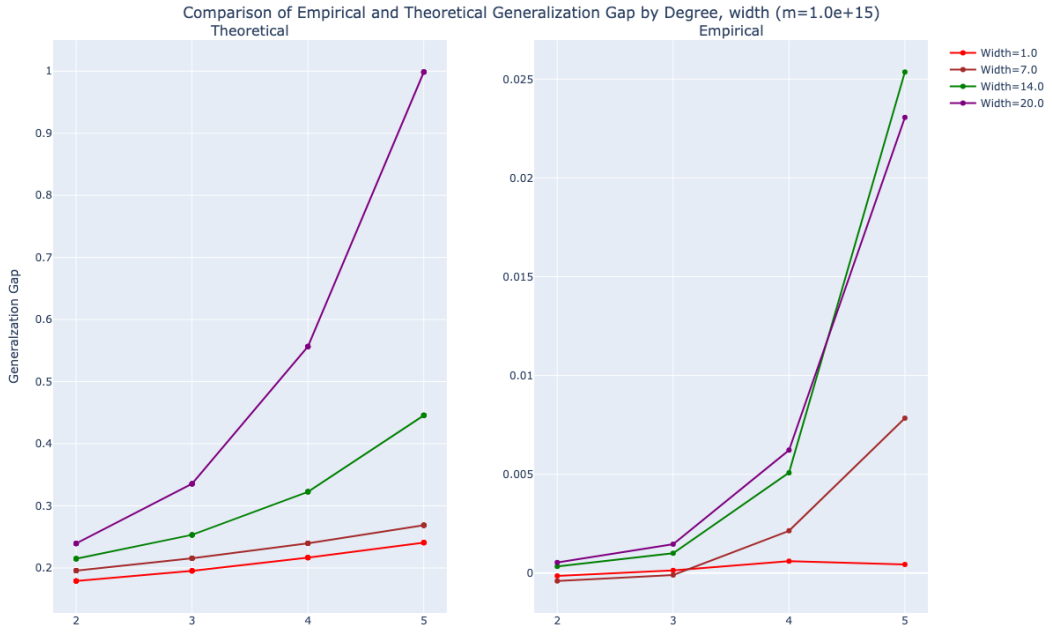

Figure 1: (left) A plot showing our generalization bound for a fixed $\sigma = 10^{-8}, m = 10^{15}$. (right) A plot showing the average empirical generalization gap over a sample of 5 functions from each Degree, Sparsity class. We note that, as predicted by our generalization bound, the gap increases super-linearly with degree, and the rate of increase increases with larger sparsity. The size of the training set here is 8192.

D Experiments

D.1 Generalization Gap

As a first step in instantiating our bound, we algorithmically picked our values for Σ, m according to the procedure described in G. With the bound of 4 instantiated, we then had to pick the value to be used for σ. As described in H.6, we chose σ based on the minimum of the optimal σs in our set of complexity classes.

With our bound instantiated, we obtained the predictions that an increase in both of the complexity parameters D_f, ω should increase the generalization gap super-linearly, and that since they are multiplied together in our bound, the increasing one of the complexity parameters should increase the rate of increase for the other. We then validated these predictions empirically.

For our main experiments for calculating the empirical generalization gap, we trained a simple transformer with 2 layers and 1 head, and without layernorm as in our construction. Due to the difficulty in generating boolean functions of arbitrary spectral sparsity and degree, we opted to relax this requirement to functions on a boolean domain. Note that our above theoretical analysis is valid for any function with a Fourier-Walsh spectrum with all-positive coefficients and sparsity no greater than the sequence length, T.

In order to simplify both our analysis and our experiments, we restricted the class of functions to those with a constant degree for all components. In order to generate such random functions, we generate the coefficients according to a standard normal distribution, and then normalize them so that they were centered and with variance 1. For the Fourier components of degree D_f, we select a random subset of size D_f from our T input dimensions by selecting the initial ω elements of a permutation of all D_f sized combinations of $[1...T]$.

D.2 Validating Low-Sharpness Interpolator Assumption

In order to validate our approach, we hard-coded a transformer according to our exact construction, and compared its sharpness (as measured by the trace of the loss Hessian) to that of our learned solutions for each degree and sparsity. We find that in general, both the sharpness and Frobenius norm of the hard-coded transformer were significantly higher than that of the learned constructions, empirically validating our low-sharpness interpolator assumption. This is a key assumption in our methodology.

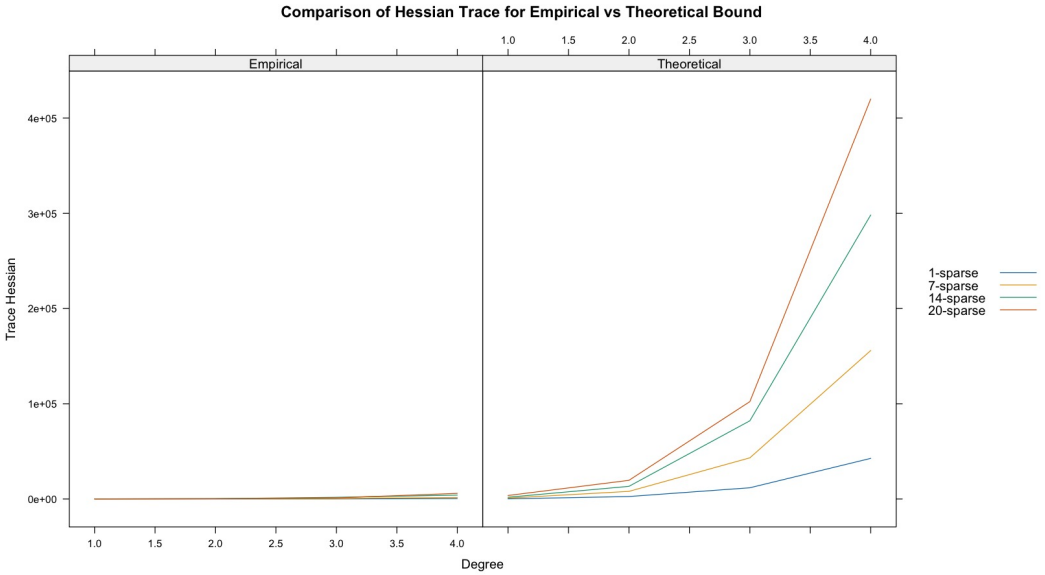

Figure 2: Right: the sharpness (as measured by the trace of the loss Hessian) of our exact construction for each degree and sparsity. Left: the sharpness of the learned solutions for each degree and sparsity. Note that the sharpness of our construction indeed upper bounds the sharpness of the learned solutions at each degree and sparsity, by roughly two orders of magnitude.

D.3 Mechanistic Interpretability

While our bound does not directly rely on the assumption that anything resembling our construction is actually learned in practice, it would nonetheless provide additional supporting evidence for the relevancy of our constructive approach. To demonstrate this, we conducted experiments with small transformer models. These models were designed with fixed embedding sizes of $T + 2$, where the first T channels correspond to hard-coded one-hot positional embeddings, and the remaining channel represents one-hot encoded bit values. Additionally, we constrained the multi-layer perceptron (MLP) hidden dimension to 32 to ensure a compact model size matching our construction. In order to be able to better visualize the behavior of the MLP, our mechinterp experiments projected down to a single dimension after the attention sub-layer, so that the MLP matrices are all $4(D_f + 1) \times 1$ or $1 \times 4(D_f + 1)$ vectors. Since this would be incompatible with a second attention layer, we replace the second layer with a simple $(T + 1) \times 1$ projection that performs the final linear combination, rather than using a purely position-aware attention mechanism (the expectation is that these parameters will store the values of the Fourier-Walsh coefficients c_t). The model contained just over 1000 parameters total. We visualized the attention weight matrix $\bar{W}^{(1)}$ and observed that in most cases, $\bar{W}^{(1)}$ contained rows corresponding to the Fourier components of its learned function, i.e. depicting bright spots in columns corresponding to the bits in a Fourier component. For example, in Figure 4, one can identify rows corresponding clearly to the components $x_9x_{14}x_{17}$ and $x_3x_8x_{11}$, while the third component, with the smallest coefficient, is less identifiable.

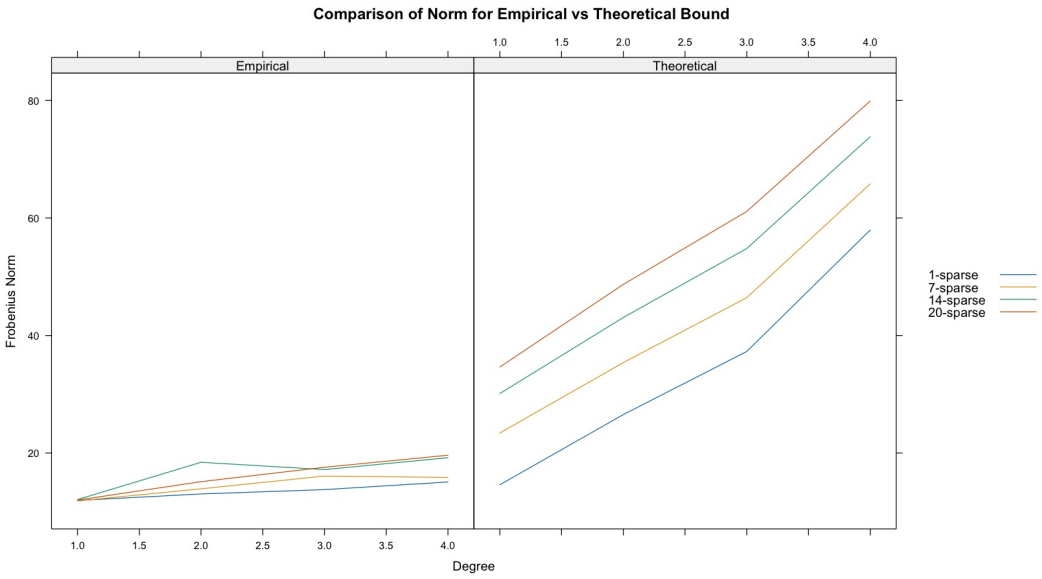

Figure 3: Right: the Frobenius norm of our exact constructions for each degree and sparsity. Left: the Frobenius norm of the learned solutions for each degree and sparsity. Note that the norm of our construction indeed upper bounds the norm of the learned solutions for each degree and sparsity, by around two orders of magnitude.

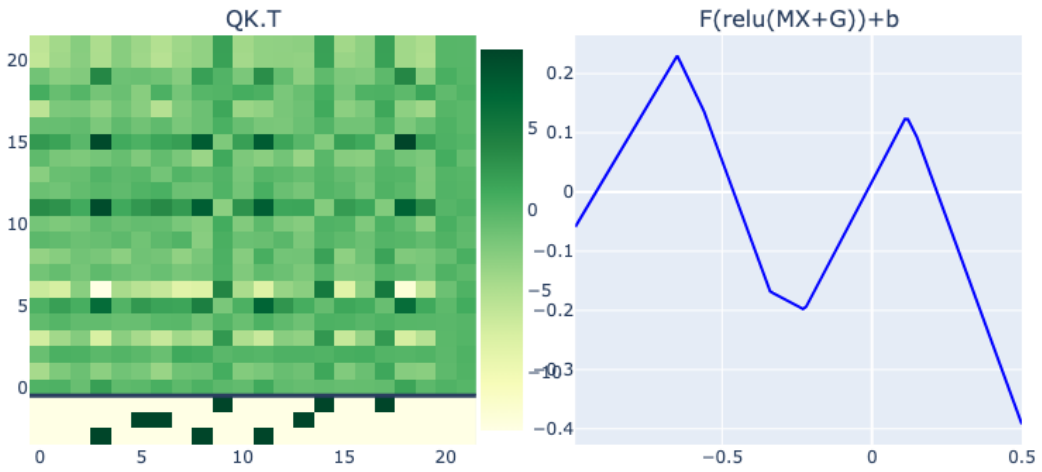

Figure 4: Left: A plot showing the combined attention matrix W for a function learned with an architecture matching our construction. Right: The learned 1-dimensional MLP function, which appears to exhibit cyclical behavior.

E Result for Chain of Thought

We now apply our theoretical framework to understand the benefit of chain-of-thought for computing high-degree functions. It is well-documented that computing intermediate steps substantially boosts transformers' ability to learn to compute high-degree functions such as Parity [e.g. 5, 15, 1]. As an explanation for this benefit, we now use our techniques to show a favorable learning bound for Parity in the presence of chain-of-thought:

Theorem 7. *Let $T \in \mathbb{N}$. Let $f_1, \ldots, f_T$ where $f_i(x_1 \ldots x_T f_1 \ldots f_{i-1}) = \bigoplus_{j \leq i} x_j$. Let $\widehat{\Theta_{CoT}}(\alpha, \beta)$ be the solution returned by the general learning procedure on the training set of size m jointly consisting of these functions using regularization parameters α, β, and let $\delta_{CoT}(\alpha, \beta)$ be the expected*

final error for a chain of T auto-regressive steps using the transformer parametrized by $\widehat{\Theta_{CoT}}$ to solve the Parity Task of length T. Then there exist regularization parameters $\alpha_{CoT}, \beta_{CoT}$ such that

$$\delta_{CoT}(\alpha_{CoT}, \beta_{CoT}) \leq Te^{-\frac{m}{8\Sigma^2}} e^{\frac{m\sigma^2}{4\Sigma^2}\left(2G_u(1,2)+P(\sigma,1,2,T)\right)+\frac{L(1,2,T)}{2\sigma^2}}$$

The proof is deferred to 30 in the appendix. It follows a straightforward application of our generalization bound for $\omega = 1$ and $D_f = 2, T$, and uses the fact that all of the functions $G_u(), P()$, and $L()$ are superlinear in T. We defer the reader to [8] for a mechanistic interpretability study of similar, simple"iteration heads" in transformers. In contrast, when learning to compute Parity without intermediate steps, a substantially less favorable bound remains:

Theorem 8. *Let $\widehat{\Theta_{OP}}(\alpha, \beta)$ be the solution returned by the general learning procedure on the training set of size m consisting of only the inputs and outputs of the complete Parity task of length T, i.e. the functions $f(x_1 \dots x_T) = \bigoplus_{j \leq T} x_j$. Let $\delta_{OP}(\alpha, \beta)$ be the expected error for a single pass of a transformer parameterized by $\widehat{\Theta_{OP}}(\alpha, \beta)$ on the parity task of length T. Then there exist regularization parameters α_{OP}, β_{OP} such that*

$$\begin{aligned}\delta_{OP} &\leq e^{-\frac{m}{8\Sigma^2}} e^{\frac{m\sigma^2}{4\Sigma^2}\left(2G_u(1,T)+P(\sigma,1,T,T)\right)+\frac{L(1,T,T)}{2\sigma^2}} \\ &\leq e^{-\frac{m}{8\Sigma^2}} e^{\frac{m\sigma^2 T}{8\Sigma^2}\left(2G_u(1,2)+P(\sigma,1,2,T)\right)+\frac{TL(1,2,T)}{2\sigma^2}}\end{aligned}$$

In other words, our bound on the error for Parity increases exponentially with length when using the one-pass approach, whereas using the CoT approach, the error increases only linearly with T.

It is worth comparing to prior theoretical accounts for the benefit of chain-of-thought for Parity. Abbé et al. [2] argue that functions with high Globality Degree require a chain-of-thought to be learned efficiently, though the general version of this claim remains conjectural. Hahn & Rofin [15] show that transformers face a very steep loss landscape when learning Parity, but do not show a converse positive result showing that chain-of-thought makes learning easy. Kim & Suzuki [18] show that chain-of-thought helps transformers learn *sparse (subset)* parities, whereas our result applies to the full Parity function applying to all input bits.

F Further Discussion

F.1 Implications

Much recent work has analyzed the ability of transformers to learn boolean functions, and linked certain properties such as low-degree or low-sensitivity of the target function to the hardness of learning. While these studies showed strong empirical and theoretical links between boolean degree and sensitivity to generalization in transformers [7, 1, 15], ours directly uses the properties of these functions in a generalization bound. Furthermore, there is a link between boolean degree and sensitivity, in that low-sensitivity functions are "close" to juntas with high probability, and are therefore capped in their maximum degree [11]. Thus, our bound further cements both boolean degree and sensitivity as complexity measures for transformers. We also introduce the Fourier sparsity as a secondary factor in generalization, adding to our understanding of precisely what makes certain boolean functions hard for transformers to learn.

To our knowledge, this is the first generalization bound that relies on explicitly upper bounding the gradients and hessian of a given class of transformers. For a certain class of low-sharpness interpolators of these constructed transformers, we were able to derive non-vacuous limits on the global quadratic loss of our transformer for arbitrary functions on a boolean domain. We believe our overall methodology offers a promising new approach to explaining generalization in transformers for which there is a specific, known construction that is plausibly learned by them in practice (ideally, supported with mechanistic interpretability studies). For other kinds of algorithmic problems, one could design a transformer to solve the task (or use a specialized programming language for constructing transformers, such as RASP [27]), and then employ automated symbolic solvers to calculate derivatives of the construction.

Beyond the academic and theoretical, our work has potentially broad societal implications. For example, we might suppose that a transformer-based AI targeting system for warfare is trained

on some *targeting* function. When trained on *targeting*, after some image processing in the lower layers, we hypothesize that at some point in the network's layers there is a circuit implementing a high-degree boolean function – in other words, one involving a conjunction of many distinct logical conditions which must be met – on some abstract input space involving concepts about the image. Our theoretical results suggest that "vanilla" transformers likely cannot generalize well on such tasks without CoT. This could have major societal implications that affect e.g. how one should conduct training, and how one should perform testing and evaluation for such tasks.

F.2 Related Work

The work of [9] showed that, even though transformers can be explicitly constructed to learn highly sensitive (and high degree) boolean functions such as parity, hence are not limited by their expressivity, these functions remained difficult for a transformer to learn. In particular, they showed that the reason for this difficulty learning mechanistically speaking is that they induce a strongly oscillating loss curve in the immediately vicinity of some of the key parameters. Still other work such as [19] showed an explicit construction for how parity was learned by transformers using a "shortcut", which explicitly memorized the mod-2 function in the MLP. These two observations, when combined, led to the key insight behind our work: that for higher degree boolean functions, the oscillating function approximated by the MLP will have higher curvature, which should lead to a larger magnitude of the hessian than for lower-degree functions. This intuition turned out to be correct, and serves as the primary basis for our PAC-Bayes bound.

Some of our methods are inspired by techniques in Deora et al. [10], but substantial changes were needed. Their construction did not have an MLP or second attention layer like ours, and differed in some other ways, but demonstrated key techniques such as using finite differences to calculate the Gradients and the Hessian. In addition, the techniques used in bounding the perturbed gradients share many commonalities with the techniques used by Edelman et al. [12] .

F.3 Limitations

One of the main limitations of our bound is the perturbation term $P(\sigma)$ appearing in our bound on the trace of the perturbed loss hessian. While this enabled a fully analytical approach, these bounds on the perturbation terms are likely far from tight, based on our analysis of its outsized role in our bound in H.6. One possible remedy to this situation would be to use parameter- or parameter-group specific perturbation variances, so that parameters for which the scale is smaller or the overall sensitivity of the transformer is higher can have a smaller variance, without causing the KL term to blow up. Another possible remedy is to use a higher order approximation to the sharpness term. If it were a third order approximation, we would be able to use the exact bound on the hessian, which would simply be $G_u(\omega, D_f) \in O(\omega D_f^3)$. Meanwhile, it would be the third derivative that needs to be evaluated at a perturbed point; but these perturbations would now be multiplied by a factor of σ^3 rather than σ^2, alleviating the constraint on σ to be as small compared to other undesired factors such as T.

Additionally, we note that for larger networks, the $|\Theta|$ term becomes prohibitive, even if we are using tricks like random projections. Using automated solvers to approximate the trace of the loss hessian for larger networks directly, rather than going through the operator norm, would be more scalable.

G Experimental Implementation Details

For our experiments, we trained our simple transformer on 5 randomly sampled functions from each (sparsity,degree) complexity class, with the degree ranging from 1 to 5, and the sparsity ranging from 1 to 20; specifically, we let $\omega \in \{1, 7, 14, 20\}$. We chose as our context length $T = 50$ In order to help in debugging and reproducibility, we set the seed for the pseudorandom generator involved in generating the above functions, based on the concatenation of (i, degree, sparsity) interpreted as an integer, where i is the index of the function sampled within each class. All functions are trained until they have a training loss below 0.01, or else they are discarded from our analysis. We found that there was significant variability in the hardness of learning of these random functions even within each function class, and our model did not converge randomly for some functions scattered across various degrees and sparsities.

In order to instantiate our bound, we needed to estimate the value of the subgaussian constant. We did so for several different functions trained to nearly perfect training loss, by analyzing the distribution of the validation losses, calculating the moment generating function, and then finding the smallest σ such that a centered gaussian with standard deviation σ has an MGF that strictly dominates that of our losses. We found Σ empirically to fall in the [.01,.1] range. Thus we use $\Sigma = 0.01$ as a reasonable setting for this parameter when instantiating our bound.

As a next step in instantiating our bound, we needed to set m. To do so, we created a grid over the values of D_f, ω that are described above, with $T = 50$. We then slowly increased m until our bound yielded a generalization error of less than 1, and chose this m as our minimal sample complexity.

Training was performed using a dataset of 8192 randomly generated binary strings of length 50 on 8 A100 GPUs using the DistributedDataParallel package with the "nccl" backend. We found that the rate and likelihood of convergence was strongly dependent on the batch size. While we would have preferred to use larger batch sizes for greater efficiency, we found that convergence to lower losses was much less likely with batches larger than 64 (per GPU), and therefore used this as our batch size. We trained each model until it either reached a loss of 0.01 or training had reached 15,000 epochs. We used a learning rate of 5×10^{-3}, a weight decay of 0.0001, no dropout. The hidden dimension used was 22, and the feed forward dimension was 32. Using these hyperparameters, it took roughly one day to train all 5 random functions from each of the 20 complexity classes.

Our transformer was written using the pytorch library, and was based on the standard implementation of a 1-layer, 1-head transformer from the transformers python library. However we made the appropriate modifications to match our construction above, including the removal of the layernorm, explicit encoding of the positional embedding to match our one-hot positional encoding, a final matrix projection layer, and log scaling of the attention logits. during an ablation study, we found that the results of our experiments did not change qualitatively when we added or removed the layernorm before and/or after the MLP sublayer.

One difference between our construction above and the transformer used in our experiments is that our transformer uses both a larger hidden dimension and a larger MLP width than the theoretical construction. The main reason for this was simply that we found convergence to be much faster with a larger model. We were able to confirm that for smaller degrees and sparsities, convergence was achieved with the minimum number of dimensions described by our construction, i.e. N dimensions for the positional encoding, and only 2 additional dimensions. Indeed, this minimal transformer is what was used during our mechanistic interpretability study. Still, due to the hardness of learning functions with larger degrees and sparsities, we found a larger model to be necessary to obtain our main results showing generalization over various degrees and sparsities in a reasonable amount of time.

H Complete Construction

We define a transformer that takes bitstrings of length T as input. We reserve a special token at the end of the sequence called the CLS token, which is where the final output will be read from. Therefore, the actual input has $T + 1$ positions. For each position $t \in [T + 1]$, we define our input vector to be

$$y_t = \{I[t = 1], ..., I[t = T + 1], z_t\} \in \mathbb{R}^{T+2}$$

where $z_t \in \{0, 1\}$ is the value of the t^{th} bit. For the CLS token, we always fix $z_{T+1} = 0$. Note that in our initial construction, the input dimension is $T + 2$. We are working towards a construction that approximates this one with fewer dimensions, which we will define at the end of this section.

In order to calculate the value of each Fourier component χ_{S_t} at each position in the first layer, we will make use of a purely position-aware attention mechanism with $O(log(T))$ scaling. For each position in the first ω rows of the attention matrix, this puts weights of approximately $\frac{1}{D_f}$ on all positions $j \in S_t$, and 0 elsewhere. We want both the CLS token and the non-active positions (which will all have uniform attention patterns) to pass through the attention layer with no operation. This can be accomplished by forcing these positions to attend only to the CLS token. I.e.,

$$a_{i,t} = \begin{cases} 2log(T) & \text{if i } \in S_t \vee (i = T+1 \wedge t > \omega) = \\ 0 & \text{if i } \notin S_t \wedge (i \neq T+1 \vee t \leq \omega) \end{cases} \tag{4}$$

so that

$$\hat{a}_{i,t} = \begin{cases} \frac{1}{\frac{T+1-D_f}{T^2}+D_f} & \text{if i} \in S_t \\ \frac{1}{T+1-D_f+D_fT^2} & \text{if } i \notin S_t \wedge (i \neq T+1 \vee t \leq \omega) \\ \frac{T^2}{T+T^2} & \text{if } i = T+1 \wedge t > \omega \end{cases} \tag{5}$$

In the limit of T large, this becomes

$$\hat{a}_{i,t} = \begin{cases} \frac{1}{D_f} & \text{if i} \in S_t \\ 0 & \text{if } i \notin S_t \wedge (i \neq T+1 \vee t \leq \omega) \\ 1 & \text{if } i = T+1 \wedge t > \omega \end{cases} \tag{6}$$

H.1 $\bar{Q}^{(1)}$, $\bar{K}^{(1)}$

We let $\bar{Q}^{(1)}, \bar{K}^{(1)} \in \mathbf{R}^{(T+2)\times(T+1)}$ We define $\bar{Q}^{(1)}$ to be a diagonal matrix which holds a scaling factor of $2log(T)$ on each of the diagonals.

$$\bar{Q}^{(1)}_{s,t} = \begin{cases} 2log(T) & s = t \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

We define $\bar{K}^{(1)}$ to be a matrix such that the first ω columns contain inclusion indicators for each component in the first T rows, followed by $T+1-(\omega)$ columns which have all zeros except for the, $T+1^{th}$ row, which has a 1. The final, $(T+2)^{th}$ column and row are all $0s$.

$$\bar{K}^{(1)}_{s,t} = \begin{cases} 1 & \text{if } (s \in \chi_{s_t}, t \in [\omega]) \vee s = T+1, t \in [\omega+1, T+1] \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

Note that

$$(y_i^T \bar{Q}^{(1)})_t = \sum_{s=1}^{T+1} 2log(T) I[i = s] I[s = t]$$

$$= 2log(T) I[i = t\}]$$

and

$$(y_j^T \bar{K}^{(1)})_t = \sum_{s=1}^{T} I[j = s] I[(s \in \chi_{S_t} \wedge t \in [\omega]) \vee (s = T+1, t \in [\omega+1, T+1])]$$

$$= I[(j \in \chi_{S_t} \wedge t \in [\omega]) \vee (j = T+1, t \in [\omega+1, T+1])]$$

Thus

$$a_{i,j} = (y_i^T \bar{Q}^{(1)})(y_j^T \bar{K}^{(1)})^T = 2log(T)\Big(I[i > \omega] \wedge j = T+1] + \sum_{t=1}^{\omega} I[i = t] I[j \in \chi_{S_t}]\Big)$$

$$= 2log(T)\big(I[j \in \chi_{S_i} \vee (i \in [\omega+1, T+1] \wedge j = T+1)]\big),$$

which matches our the desired behavior for attention weights in the first layer. The output of this step is a $(T+1) \times (T+1)$ attention matrix, where each of the first ω rows corresponds to the i^{th} Fourier component, and has $\frac{1}{|\chi_{S_r}|}$ in column j if position j is a member of the i^{th} Fourier component. All other rows simply give uniform attention, except the last row, which is is an identity mapping (corresponding to a no-op in the case of the CLS token).

H.2 Random projections for $log(T)$ width

Suppose we add a projection matrix into our attention mechanism: define $a^*_{i,j} = y_i^T JJ^T \bar{Q}^{(1)} (\bar{K}^{(1)})^T JJ^T y_j$, where $J \in \mathbb{R}^{T+2 \times d+1}$. is an additional inner projection matrix with entries in the first $T+1$ rows and d columns that are i.i.d. normally distributed with standard deviation $\frac{1}{\sqrt{d}}$, and a 1 in bottom right element, and all else 0 in the last row and column, i.e. $J_{ij} \sim \frac{1}{\sqrt{d}}\mathcal{N}(0,1)$. for $i \in [T+1], j \in [d]$, It is an immediate consequence of the Johnson-Lindenstrauss Lemma that for any fixed $\epsilon_p > 0$, for any integer $d > \frac{8log(T)}{\epsilon_p^2}$.

$$Pr\Big[|\langle (\bar{Q}^{(1)})^T JJ^T y_i, (K^{(1)})^T JJ^T y_j \rangle - \langle (Q^{(1)})^T y_i, K^T y_j \rangle| < \epsilon_p\Big] > 1 - 2e^{-\frac{d}{2}(\frac{\epsilon_p^2}{2} - \frac{\epsilon_p^3}{3})} = \delta_p$$

This trick allows us to define our projected inputs, $x_t = J^T y_t \in \mathbb{R}^n$, where $n \in O(log(T))$, using modified attention matrices $Q^{(1)} = J^T \bar{Q}^{(1)}$ and $K^{(1)} = J^T \bar{K}^{(1)}$ so that the overall output of the network is approximately the same as for the original network.

H.3 $\bar{V}^{(1)}$

For our value matrix, we define $\bar{V}^{(1)} \in \mathbb{R}^{(T+2)\times(T+2)}$

$$\bar{V}^{(1)}_{s,i} = \begin{cases} 1 & \text{if s=T+2 and i=T+2} \\ 0 & otherwise \end{cases} \tag{9}$$

Thus

$$y_i^T \bar{V}^{(1)} = \sum_{s=1}^{T+1} y_{i,s} \bar{V}^{(1)}_{s,T+1} = concat(\mathbf{0}_{T+1}, z_i\}$$

Now, for notational simplicity we write $\bar{W}^{(1)} = \bar{Q}^{(1)}(\bar{K}^{(1)})^T$, and we define the output at position t after the first attention layer as $\bar{u}_t$, where we are representing $\bar{u}_t$ as a $(T+2) \times 1$ column vector. In order to express this desired quantity as a column vector, we find it useful to transpose our transformer construction, so that

$$\bar{u}_t = (\bar{V}^{(1)})^T Y^T \phi(Y(\bar{W}^{(1)})^T y_t) = concat(\mathbf{0}_{T+1}, \frac{k_t}{D_f}\} \in \mathbb{R}^{(T+2)}$$

We can easily modify our definition of $\bar{V}$ to work with our dimension-reduced inputs instead. We can replace $\bar{V}^{(1)}$ with $V^{(1)} = J^T \bar{V}^{(1)} J \in \mathbb{R}^{(d+1)\times(d+1)}$ and rewrite the output of our attention layer in terms of our dimension-reduced inputs $X = YJ$ as

$$u_t = (\bar{V}^{(1)})^T JJ^T Y^T \phi(YJJ^T (\bar{W}^{(1)})^T JJ^T y_t) \approx \bar{u}_t$$

We work with the attention output in the dimension reduced space:

$$b_t := J^T u_t \in \mathbb{R}^{d+1}$$
$$= (V^{(1)})^T X^T \phi(XW^{(1)} x_t)$$

We note that using our definition of $V^{\bar{(1)}}$, the only effect of applying the random projection is to reduce the matrix's dimensions: since all of the upper left $(T+1) \times (T+1)$ block is 0 anyways, and this is the the only portion subject to random projections, this block simply gets replaced with a $d \times d$ block of zeros. This approximation allows us to achieve a correct construction with a width of only $d \in O(log(T))$.

H.4 M, F, Γ in MLP layer

Next, we make use of a wide MLP layer which interpolates the "mod 2" function, or alternatively, interpolates the function

$$m : \mathbb{R} \to \mathbb{R}, m(x) = \frac{1}{2}\left(1 + sin\Big(\pi\Big(D_f * x + \frac{1}{2}\Big)\Big)\right)$$

on the domain $x \in \{0, \frac{1}{D_f}, ..., \frac{D_f-1}{D_f}, 1\}$. Below, we sometimes write $m(M)$ where M is a vector or a matrix. In these cases, we can understand this to mean m applied element-wise.

Note that the maximum value for k_t is D_f, and thus the domain of the function needing to be memorized has cardinality at most $D_f + 1$. In order to memorize such a function, using a construction similar to that of [19] or [9], we require $4(D_f + 1)$ MLP units. This creates a first layer that acts as an indicator function $I[k = i]$ in a lookup table, with the second layer of the MLP storing the function values.

Theorem 9. *Let* $\bar{M} \in \mathbb{R}^{(T+2)\times 4(D_f+1)}$, $\bar{\Gamma} \in \mathbb{R}^{4(D_f+1)}$, *and* $\bar{F} \in \mathbb{R}^{(4(D_f+1))\times(T+2)}$. *For each* $i \in [D_f], h_i \in \{0, \frac{1}{D_f}, ..., \frac{D_f-1}{|\chi_{S_r}|}, 1\}$, *Let*

$$\bar{M}_{i,t} = 0, \forall t, i \in [T+1]$$
$$\bar{M}_{i,t} = 1, \forall t, i = T+2$$
$$\bar{\Gamma}_{4i-3} = -h_i - \frac{2}{4D_f}$$
$$\bar{\Gamma}_{4i-2} = -h_i - \frac{1}{4D_f}$$
$$\bar{\Gamma}_{4i-1} = -h_i + \frac{1}{4D_f}$$
$$\bar{\Gamma}_{4i} = -h_i + \frac{2}{4D_f}$$

and let

$$\bar{F}_{i,t} = 0, \forall i \in [4(D+f+1)], t \in [T+1]$$
$$\bar{F}_{4i-3,T+2} = 4m(h_i)D_f \forall i \in [D_f]$$
$$\bar{F}_{4i-2,T+2} = -4m(h_i)D_f \forall i \in [D_f]$$
$$\bar{F}_{4i-1,T+2} = -4m(h_i)D_f \forall i \in [D_f]$$
$$\bar{F}_{4i,T+2} = 4m(h_i)D_f \forall i \in [D_f]$$

Then the function

$$g_t = \bar{F}^T(\bar{M}^T b_t + \bar{\Gamma})_+$$

interpolates

$$\approx \frac{1}{2}\left(1 + sin\Big(\pi\Big(D_f * b_t + \frac{1}{2}\Big)\Big)\right) = \frac{1}{2}\left(1 + sin\Big(\pi\Big(k_t + \frac{1}{2}\Big)\Big)\right),$$

on the domain of possible values of b_t, *a grid of the possible values of our scaled prefix sums,* h_i.

Proof: First, we look at the product $\bar{M}^T b_t$.

$$\left[\bar{M}^T b_t\right]_i = \frac{k_t}{D_f} \implies \bar{M}^T b_t = \frac{k_t}{D_f} \not\Vdash^T_{4(D_f+1)}$$

It is easy to see that

$$max\Big(\Big[\bar{M}^T b_t + \bar{\Gamma}\Big]_{4i-2}, 0\Big) = \frac{(k_t - (i - \frac{6}{4}))_+}{D_f}$$
$$max\Big(\Big[\bar{M}^T b_t + \bar{\Gamma}\Big]_{4i-2}, 0\Big) = \frac{(k_t - (i - \frac{5}{4}))_+}{D_f}$$
$$max\Big(\Big[\bar{M}^T b_t + \bar{\Gamma}\Big]_{4i-1}, 0\Big) = \frac{(k_t - (i - \frac{3}{4}))_+}{D_f}$$
$$max\Big(\Big[\bar{M}^T b_t + \bar{\Gamma}\Big]_{4i}, 0\Big) = \frac{(k_t - (i - \frac{2}{4}))_+}{D_f}$$

Finally, we calculate the product as three separate sums:

$$max(b_t\bar{M} + \bar{\Gamma}^T, 0)F$$

$$= 4D_f m(h_i) \sum_{i=1}^{D_f+1} (b_t - \bar{\Gamma}_{4i-3})_+ - (b_t - \bar{\Gamma}_{4i-2})_+ - (b_t - \bar{\Gamma}_{4i-1})_+ + (b_t - \bar{\Gamma}_{4i})_+$$

$$= 4m(h_i) \sum_{i=1}^{D_f+1} (k_t - (i - \frac{6}{4})_+ - (k_t - (i - \frac{5}{4})_+ - (k_t - (i - \frac{3}{4})_+ + (k_t - (i - \frac{2}{4}))_+$$

The above expression is only nonzero on the integers when $k_t = i - 1$, in which case its value is $\bar{F}^T max(\bar{M}^T b_t + \bar{\Gamma}, 0) = m(h_i)$, proving the claim. The function is a piecewise-linear 'trapezoidal' wave alternating between 0 and 1.

Let us denote this trapezoidal wave function of our MLP by $h(b_t) = \bar{F}^T(\bar{M}^T b_t + \bar{\Gamma})_+$. The derivative $h'(b_t)$ is easily shown to be a square wave that alternates between $-4D_f, 0$ and $4D_f$. The second derivative of h is 0 almost everywhere. We will see later that we avoid dealing with the second derivative entirely, by using first order Taylor approximations and finite differences rather than exact derivatives. After the MLP layer, we now have $k_t mod 2$ stored in each post-MLP activation, h_t, as desired.

Just as the other matrices were compressed using random projections, we similarly can define the dimension-reduced matrices $M = J^T\bar{M}$ and $F = \bar{F}J$. We note that in the case of the MLP matrices, the random projection portion of the J matrix has no effect other than to shrink the size of the 0 blocks of these matrices from having one dimension of size $T + 1$ to that dimension now having size d, while maintaining their original values in the final row and column, respectively.

We will find it useful to refer to the MLP outputs in the dimension-reduced space,

$$g_t = J^T h_t \in \mathbb{R}^{(d+1)\times 1}$$

H.5 $\bar{W^{(2)}}$,$\bar{V^{(2)}}$

To get out final output, we linearly combine the individual Fourier components from the MLP function, g_t. To do so, we implement a final transformer "half-layer", or, a layer consisting of just an attention sub-layer (but no MLP sub-layer). In this sub-layer, The inner attention projection matrix for the full dimension inputs, $\bar{W}^{(2)}$, is a $(T + 2) \times (T + 2)$ matrix with only the penultimate row nonzero. This row contains $log(c_t T^2)$, in the first ω entries, and 0 elsewhere. The value matrix in the second layer is $\bar{V}^{(2)}$, which is a $(T + 2) \times 1$ vector which selects the final dimension from the incoming vector and scales it by a factor of $D = \sum_{t=1}^{\omega} c_t$.

Note that the second layer's pre-softmax attention output at the (t, s) position is zero in all but the last row, which has entries given by $a^{(2)}_{T+1,t} = g^T_{T+1}\hat{W^{(2)}}g_t = log(c_t T^2), \forall t \leq \omega$. Taking the softmax of this first row of the attention matrix yields:

$$\hat{a}^{(2)}_{T+1,t} = \frac{e^{log(c_t T^2)}}{T + 1 - \omega + \sum_{l=1}^{\omega} e^{log(c_l T^2)}} = \frac{c_t T^2}{T + 1 - \omega + T^2 \sum_{l=1}^{\omega} c_l} = \frac{c_t}{\frac{T+1-\omega}{T^2} + D} \to \frac{c_t}{D}$$

Finally, the $\bar{V}^{(2)}$ is defined to have $0s$ everywhere except the final dimension, which has a value of D, so that the final activation is equal to:

$$\sum_{t=1}^{T} \hat{a}^{(2)}_{T+1,t} g_t^T \bar{V}^{(2)} = \sum_{t=1}^{T} \frac{c_t}{D} g_{t,T+2} D = \sum_{t=1}^{T} c_t g_{t,T+2}$$

Recall from above that in the final dimension of the output of the MLP, $g_{t,T+2}$, is stored the parity of component t for the input string. Thus the above formula matches the definition of our target function.

The dimension-reduced versions of $\bar{V}^{(2)}$ and $\bar{W}^{(2)}$ are obtained through the same overall pattern as the other matrices:

$$V^{(2)} = J^T\bar{V}^{(2)}, W^{(2)} = J^T\bar{W}^{(2)}J$$

Our complete transformer output can then be written in a series of the following "blocks" corresponding to the three sub-layers involved in our construction:

$$\mathcal{T}(X, W, V, M, F, \Gamma, W^{(2)}, V^{(2)}) = (V^{(2)})^T G^T(\phi(G(W^{(2)})^T g_{T+1}))$$

where

$$g_t := b_t + F^T\Big(M^T b_t + \Gamma\Big)_+, G := \begin{bmatrix} g_1^T \\ \dots \\ g_{T+1}^T \end{bmatrix} \in \mathbb{R}^{(T+1)\times(d+1)}$$

$$b_t = (V^{(1)})^T X^T \phi\big(X^T (W^{(1)})^T x_t\big), B := \begin{bmatrix} b_1^T \\ \dots \\ b_{T+1}^T \end{bmatrix} \in \mathbb{R}^{(T+1)\times(d+1)}$$

Note that we have a residual connection after the MLP, but not after the first attention layer; in other words, the matrix B is missing the (randomly projected) positional encoding bits in the first d dimensions, unlike X and G.

Theorem 10. *Let $\mathcal{T}(X, \Theta) : \mathbb{R}^{(T+1)\times(d+1)} \to \mathbb{R}$ be the 1-layer transformer as defined above. Let $f : \{0,1\}^T \to \mathbb{R}$ be a function with a Fourier spectrum as described above: with maximum degree D_f, and with a sparsity ω at most T. Then we have the following bounds on the input and parameter norms:*

$$\|X\|_{2,\infty} \lessapprox \sqrt{2},$$
$$\|XV^{(1)}\|_{2,\infty} \lessapprox 1.$$
$$\|XV^{(1)}e_{d+1}\|_\infty \lessapprox 1.$$
$$\|G\|_{2,\infty} \lessapprox \sqrt{2}$$
$$\|G\|_{1,\infty} \lessapprox 1$$
$$||g_{T+1}|| \lessapprox 1$$
$$||GV^{(2)}||_\infty \lessapprox \sqrt{\omega}$$
$$||GV^{(2)}|| \lessapprox \omega$$
$$||V^{(2)}|| \lessapprox \sqrt{\omega}$$
$$||B||_{2,\infty} \lessapprox 1$$
$$||B||_{1,\infty} \lessapprox 1.$$
$$||(W^{(2)})^T g_{T+1}|| \lessapprox 2\sqrt{\omega} log(T)$$
$$||G(W^{(2)})^T||_{2,\infty} \lessapprox 2log(T)$$
$$||W^{(2)}||_2 \lessapprox 2\sqrt{\omega} log(T)$$

Proof. $\|X\|_{2,\infty}$ is the maximum row 2-norm. $X = YJ$ is a $((T+1)\times(d+1))$ matrix where each entry in the t^{th} row is a gaussian R.V. with standard deviation $\frac{1}{\sqrt{d}}$, with the exception of the final column, which stores the bit value. Therefore the 2-norm of each row of YJ is upper bounded by

$$||Y_{t,:}J|| = \sqrt{z_t^2 + \sum_{j=1}^{d}\Big(\sum_{i=1}^{d} Y_{t,i}J_{i,j}\Big)^2} = \sqrt{z_t^2 + \sum_{j=1}^{d} J_{t,j}^2}$$

$$\leq \sqrt{1 + \sum_{j=1}^{d} J_{t,j}^2}$$

Note that $\sum_{j=1}^{d} J_{t,j}^2$ is a chi-squared R.V. with d degrees of freedom. In 28 we show that with probability $1-\delta$, such an R.V. is upper bounded as

$$\|\epsilon_{t,:}\| \leq \frac{1}{\sqrt{d}}\sqrt{d + 2\sqrt{dlog(\frac{1}{\delta})} + 2log(\frac{1}{\delta})}$$

To upper bound the maximum of T such variables, we apply a union bound:

$$\max_{t \in T} ||\epsilon_{t,:}|| \leq \frac{1}{d}\Big(d + 2\sqrt{dlog(\frac{T}{\delta})} + 2log(\frac{T}{\delta})\Big)$$

Note that per our definition of d, $d > \frac{8log(T)}{\epsilon_p^2}$. Therefore, with probability at least $1 - \delta$,

$$\max_{t \in T} ||\epsilon_{t,:}|| \leq 1 + 2\epsilon_p\sqrt{\frac{log(\frac{T}{\delta})}{8log(T)}} + 2\epsilon_p^2\frac{log(\frac{T}{\delta})}{8log(T)}$$

$$= 1 + \frac{\epsilon_p}{\sqrt{2}}\sqrt{1 - \frac{\delta}{log(T)}} + \frac{\epsilon_p^2}{4}\Big(1 - \frac{\delta}{log(T)}\Big)$$

$$\lessapprox 1 + \frac{\epsilon_p}{\sqrt{2}} + \frac{\epsilon_p^2}{4}$$

Since we can make ϵ_p arbitrarily small, the above expression approach 1, and then we can approximate $||X||_{2,\infty} \lessapprox \sqrt{1+1} = \sqrt{2}$ (in the case that the bit value z_t is 1). To approximately upper bound $||X||_{1,\infty}$, the maximum element of YJ, we simply consider the composition of the matrix X, which has the upper left $(T \times d)$ sub-matrix as gaussian random variables with standard deviation $\frac{1}{\sqrt{d}}$, and then a final column of the bit values. Since $d \in O(log(T))$, the probability of all terms being less than 1 is overwhelmingly likely. The expectation of the maximum of m random variables is known to be $\sigma\sqrt{2log(m)} + o(1)$ With $\sigma = \frac{1}{log(T)}$ and $m = Tlog(T)$, this becomes $\frac{1}{log(T)\sqrt{2log(Tlog(T))}} \to 0$ for large T. Thus the expectation is in $o(1)$. Therefore we approximate $||X||_{1,\infty} \lessapprox 1$.

To bound $||XV^{(1)}||_{2,\infty}$, note that $Y\bar{V}^{(1)}$ is a matrix similar to Y, in that the final column contains all of the bit values of Y, but all previous columns which held the positional encodings are 0. Therefore, $Y\bar{V}^{(1)}$ could have a maximum row 2-norm of 1.. The matrix $XV^{(1)} = YJJ^T\bar{V}^{(1)}$ approximates $Y\bar{V}^{(1)}$, since the error from our random projections approximation ϵ_p can be made arbitrarily small. In particular,

$$||XV^{(1)}||_{2,\infty} = ||YJJ^T\bar{V}^{(1)}||_{2,\infty} \leq ||Y\bar{V}^{(1)}||_{2,\infty} + ||YJJ^T\bar{V}^{(1)} - Y\bar{V}^{(1)}||_{2,\infty}$$

$$\leq ||Y\bar{V}^{(1)}||_{2,\infty} + ||Y||_{2,\infty}|JJ^T\bar{V}^{(1)} - \bar{V}^{(1)}||_2 \leq ||Y\bar{V}^{(1)}||_{2,\infty} + \epsilon_p||Y||_{2,\infty}||\bar{V}^{(1)}||_2,$$

$$\leq ||Y\bar{V}^{(1)}||_{2,\infty} + \sqrt{2}\epsilon_p$$

where the second to last inequality follows from JLL, and the last step follows from our bound on $||Y||_{2,\infty}$, and the fact that $\bar{V}^{(1)}$ has only a single nonzero value, which is 1, and therefore its maximum singular value can be no larger than 1. Thus,

$$||XV||_{2,\infty} \lessapprox 1$$

Similarly, we consider the $||XV^{(1)}e_{d+1}||_\infty$, which is approximated by $||Y\bar{V}^{(1)}e_{d+1}||_\infty$ due to JLL. the maximum value in the final column of $\bar{Y}V^{(1)}$, as described above, is 1. Thus

$$||XV^{(1)}e_{d+1}||_\infty \lessapprox 1$$

We now consider the norms relating to our MLP output matrix, $G \in \mathbb{R}^{(T+1)\times(d+1)}$. Let $\bar{H} \in \mathbb{R}^{(T+1)\times(T+2)}$ be the full-dimension MLP output matrix using the full-dimensional parameter matrices. The maximum column $2-$norm of $\bar{H}$ is at most $\sqrt{T+1}$, since there could be $1s$ in any element of the final column. The maximum row 2-norm is $\sqrt{2}$ because there are at most 2 nonzero bits per row. The output of the dimension reduced transformer $H \in \mathbb{R}^{(T+1)\times(T+2)}$ can be made arbitrarily close to $\bar{H}$, and thus we have

$$||H||_{1,2} \lessapprox \sqrt{T+1}$$

$$||H||_{2,\infty} \lessapprox \sqrt{2}$$

To bound these norms for the dimension-reduced MLP output matrix G, we note that $G = HJ$. and that the overall structure of H is exactly the same as that of Y above: each row has two possible nonzero elements: one for the positional encoding, and another variable bit value in the $T + 2^{th}$ column. therefore we can apply the exact same arguments to conclude that

$$||G||_{2,\infty} = ||HJ||_{2,\infty} \lessapprox ||H||_{2,\infty} = \sqrt{2}$$

We can apply this argument once more to approximately upper bound the maximum element of G :

$$||G||_{1,\infty} = ||HJ||_{1,\infty} \lessapprox ||H||_{1,\infty} = 1.$$

' We will also find it useful to bound the 2-norm of the final post-MLP activation, $||g_{T+1}||$. Note that the $d + 1^{th}$ dimension for the CLS token is always 0. Therefore $||g_{T+1}||$ is a chi-distributed variable with d degrees of freedom and standard deviation $\frac{1}{\sqrt{d}}$. Thus with high probability, for large d it becomes tightly concentrated around 1. We thus approximate $||g_{T+1}|| \lessapprox 1$.

We consider the maximum row 2-norm of the matrix $GV^{(2)}$. We can argue the same way as above in our bound on $||XV^{(1)}||_{2,\infty}$, that $||GV^{(2)}||_\infty = ||HJJ^T\bar{V}^{(2)}||_\infty$ approximates $||H\bar{V}^{(2)}||_\infty$ arbitrarily well. Since $H\bar{V}^{(2)}$ is a vector containing a copy of the final column of H, which are the MLP outputs in $\{0, 1\}$, but scaled by a factor of D, we can conclude that $||H\bar{V}^{(2)}||_\infty \leq D$. Therefore we conclude that

$$||GV^{(2)}||_\infty \lessapprox D$$

Now, note that

$$D = \sum_{t=1}^{\omega} c_t \leq \sum_{t=1}^{\omega} |c_t| \leq \sqrt{\omega}$$

Thus

$$||GV^{(2)}||_\infty \lessapprox \sqrt{\omega}$$

Noting that the final column of G is nonzero for only ω entries, we use Cauchy-Schwarz to conclude that

$$||GV^{(2)}|| \lessapprox \omega$$

Bounding $||V^{(2)}||$ is trivial, since $||V^{(2)}|| = ||J^T\bar{V}^{(2)}|| \leq D||J_{T+2,:}|| \leq D$, where we have used the fact that the final, $T + 2^{th}$ column of J is e_{d+1}^T, which has norm 1. Thus $||V^{(2)}|| \leq D \leq \sqrt{\omega}$.

Note that the matrix U has a similar overall structure as both Y and H above, except that due to the lack of residual connection after the first attention layer, B does not contain any positional encodings. Following the similar arguments with this assumption, we can bound

$$||B||_{2,\infty} = ||UJ||_{2,\infty} \lessapprox ||U||_{2,\infty} = 1$$
$$||B||_{1,\infty} = ||UJ||_{1,\infty} \lessapprox ||U||_{1,\infty} = 1.$$

Similarly, JLL says that $||G(W^{(2)})^T||_{2,\infty}$ approximates $||H(\bar{W}^{(2)})^T||_{2,\infty}$ arbitrarily well. Since the first T columns and final column of $(\bar{W}^{(2)})^T$ are $0s$, the same is true of $H(\bar{W}^{(2)})^T$. In the penultimate column of this matrix, we have the vector $(log(c_1T^2), \ldots, log(c_\omega T^2), 0, \ldots, 0)$. Each row therefore has a 2-norm that is bounded by $log(T^2) = 2log(T)$. Thus

$$||G(W^{(2)})^T||_{2,\infty} \lessapprox 2log(T)$$

Following similar arguments, it follows from JLL that $||(W^{(2)})^T g_{T+1}||$ approximates $||(\bar{W}^{(2)})^T h_{T+1}||$ arbitrarily well. Recall that the matrix $(\bar{W}^{(2)})^T$ is only nonzero in the penultimate column, which has $log(c_tT^2)$ in its first ω entries. h_{T+1} is only nonzero in its positional encoding bit, which is the penultimate bit, and therefore the vector $(W^{(2)})^T g_{T+1} = (log(c_1T^2), \ldots, log(c_\omega T^2), 0, \ldots, 0)$. It follows that $||(W^{(2)})^T g_{T+1}|| \leq \sqrt{\sum_{t=1}^{\omega} log(c_tT^2)^2} \leq \sqrt{\sum_{t=1}^{\omega} 4log(T)^2} \leq 2\sqrt{\omega}log(T)$

$||W^{(2)}||_2 = ||J^T W^{(2)} J||_2$, the operator norm of $W^{(2)}$, approximates $||\bar{W}^{(2)}||_2$. Recall that $\bar{W}^{(2)}$ is nonzero only in its penultimate row, which has $log(c_tT^2)$ in the first ω columns. Let $l := (log(c_1T^2), \ldots, log(c_\omega T^2), 0, \ldots, 0)$ be the vector representing this nonzero row. it is easy to show that $(\bar{W}^{(2)})^T\bar{W}^{(2)} = ll^T$. There is one nonzero eigenvalue of this rank-1 outer product, which is $||l||^2$. Thus the maximum singular value is $||l||$. We can easily bound this using Cauchy-Schwarz to conclude that $||\bar{W}^{(2)}||_2 \leq 2\sqrt{\omega}log(T)$. □

H.6 Details of PAC-Bayes Bound derivation

Recall that $L_f(X, \Theta)$ is the loss for our transformer learning the function f on input X. Define $L(f_{\boldsymbol{\Theta}})$ to be the global loss, while $\hat{L}(f_{\boldsymbol{\Theta}})$ is the empirical loss on our training set of size m.

The original bound of Macallester (1998) claimed that starting with any data-independent prior distribution P over hypotheses, and a data-dependent (specifically, the outcome of training a neural network) posterior distribution Q, for any $\delta > 0$, the following holds with probability at least $1 - \delta$:

$$\mathbb{E}_{\Theta \sim Q}[L(f_\Theta)] \leq \mathbb{E}_{\Theta \sim Q}[\hat{L}(f_\Theta)] + \sqrt{\frac{KL(Q||P) + ln(\frac{m}{\delta})}{2(m-1)}}$$

In Neybashur et. al. (2017), they consider the bound for a distribution of perturbations around a specific set of weights, and in so doing, derive a bound with a term representing the sharpness of the loss in parameter space, as well a term that depends on the norm of the parameters. If Θ is our learned weights, and f_θ our learned min imum, then we write $f_{\Theta+\epsilon}$ for our posterior distribution of predictors, where ϵ is some small Gaussian perturbation of the parameters. We start with the classical bound of McAllester, plugging in the above assumptions about our Gaussian prior and posterior distributions Θ, ϵ.

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)^n}[L(f_{\Theta+\epsilon})] \leq \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)^n}[\hat{L}(f_\Theta)] + \sqrt{\frac{KL(\Theta+\epsilon||\Theta) + ln(\frac{m}{\delta})}{2(m-1)}}$$

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)^n}\big[L(f_{\Theta+\epsilon})\big] \leq \hat{L}(f_\Theta) + \underbrace{\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)^n}\big[\hat{L}(f_{\Theta+\epsilon})\big] - \hat{L}(f_\Theta)}_{sharpness} + \sqrt{\frac{1}{m}\Big(\underbrace{\frac{||\Theta||^2}{2\sigma^2}}_{norm} + ln\frac{2m}{\delta}\Big)}$$

where $0 < \delta < 1$ is our probability of the bound being violated, m is the number of training points, and σ is the standard deviation of our parameter perturbation (assumed to be equal for all parameters in this version of the bound). Note that the prior distribution P is assumed to be a centered Gaussian with the same noise variance as the posterior distribution Q around its mean, Θ.

Our case is slightly different from this, as we are not using a binary loss bounded within $[0, 1]$.Rather, we consider functions f on a boolean domain, but with a real-valued range, and a natural choice of our loss is the quadratic loss, i.e. $L_f(X, \Theta) = (T(\Theta, X) - f(X))^2$. Several variants of the original PAC-Bayes bounds by McAllester have been derived that allow us to adapt Neyshabur's perturbation bound to our present situation.

[4] provide a bound intended for losses that are possibly unbounded, but have a tail distribution that is sub-gaussian. Mathematically, we assume that for some positive constant Σ, our loss function satisfies:

$$\mathbb{E}[e^{t[l(\Theta,x,f) - \mathbb{E}[l(\Theta,x,f)]]}] \leq e^{\frac{\Sigma^2 t^2}{2}}.$$

Then with probability at least $1 - \delta$, the following bound holds:

$$\mathbb{E}_{\Theta \sim Q}[L(f_\Theta)] \leq \mathbb{E}_{\Theta \sim Q}[\hat{L}(f_\Theta)] + \frac{\lambda \Sigma^2}{2m} + \frac{KL(Q||P) + ln\frac{1}{\delta}}{\lambda}$$

In order to estimate the the sub-gaussian constant for a trained model, we calculate the MGF of our validation loss distribution, and find the lowest Σ such that the gaussian MGF still dominates the true loss MGF. In our case, we found that for models trained to a loss of .01 on functions of various degrees and sparsities, Σ generally falls in the range $[0.01, 0.1]$. While in the

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)}[L(f_{\Theta+\epsilon})] \leq \hat{L}(f_\Theta) + \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)}[\hat{L}(f_{\Theta+\epsilon})] - \hat{L}(f_\Theta) + \frac{\lambda \Sigma^2}{2m} + \frac{\frac{||\Theta||^2}{2\sigma^2} + ln\frac{1}{\delta}}{\lambda}$$

We combine this with 12, which tells us that

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)^n}\Big[L(f_{\Theta+\epsilon})\Big] = \frac{\sigma^2}{2} Tr\Big(\nabla^2\big[L(f_{\Theta+\zeta})\big]\Big)$$

Under the assumption that the training loss is zero, i.e. $\hat{L}(f_\Theta) = 0$, this gives us the generalization bound:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[L(f_{\Theta+\epsilon})] \leq \frac{\sigma^2}{2} Tr\Big(\nabla^2\big[L(f_{\Theta+\zeta})\big]\Big) + \frac{\lambda\Sigma^2}{2m} + \frac{\frac{||\Theta||^2}{2\sigma^2} + ln\frac{1}{\delta}}{\lambda}$$

We now use 13 and 14showing the existence of a low-sharpness solution and and low-norm solution for any boolean function of Fourier sparsity $\omega < T$, degree D_f, and context length T, combined with our abstract regularized learning model, to replace the trace and norm terms with upper bounds on them.

$$Tr\Big(\nabla^2\big[L(f_{\Theta+\zeta})\big]\Big) \leq 2G_u(\omega, D_f) + P(\sigma) \wedge ||\mathbf{\Theta}||_F^2 \leq L(\omega, D_f, T)$$

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[L(f_{\Theta+\epsilon})] \leq \sigma^2\Big(G_u(\omega, D_f) + \frac{1}{2}P(\sigma)\Big) + \frac{\lambda\Sigma^2}{2m} + \frac{\frac{L(\omega,D_f,T)}{2\sigma^2} + ln\frac{1}{\delta}}{\lambda}$$

minimizing the RHS with respect to λ, we obtain

$$\frac{\Sigma^2}{2m} = \frac{\frac{L(\omega,D_f,T)}{2\sigma^2} + ln\frac{1}{\delta}}{\lambda^2} \implies \lambda = \sqrt{\frac{2m}{\Sigma^2}\Big(\frac{L(\omega, D_f, T)}{2\sigma^2} + ln\frac{1}{\delta}\Big)}$$

$$\implies \frac{\lambda\Sigma^2}{2m} + \frac{\frac{L(\omega,D_f,T)}{2\sigma^2} + ln\frac{1}{\delta}}{\lambda} = 2\sqrt{\frac{\Sigma^2}{2m}\Big(\frac{L(\omega, D_f, T)}{2\sigma^2} + ln\frac{1}{\delta}\Big)}$$

Hence our bound becomes:

$$\implies \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[L(f_{\Theta+\epsilon})] \leq \sigma^2\Big(G_u(\omega, D_f) + \frac{1}{2}P(\sigma)\Big) + 2\sqrt{\frac{\Sigma^2}{2m}\Big(\frac{L(\omega, D_f, T)}{2\sigma^2} + ln\frac{1}{\delta}\Big)},$$

The natural next step in simplifying our bound is to optimize over σ. If we were to momentarily ignore the perturbation term $P(\sigma)$ and the $ln\frac{1}{\delta}$ term, we would end up with the following formula for the optimal σ^* :

$$\sigma^* = \Big(\frac{\Sigma^2 L(\omega, D_f, T)}{4mG_u(\omega, D_f)^2}\Big)^{\frac{1}{6}}$$

Unfortunately, if we use the full bound including the perturbation term, the larger constants and stronger dependency on D_f, ω mean that we are unable to leverage the optimal $\sigma*$ ends up being much smaller in reality. We return to the analysis of the optimization of σ and tightness of the full bound after a brief digression that continues with the analysis of this truncated bound in order to contextualize our overall approach against the behavior of previous bounds for transformers on boolean domains.

H.7 Comparison with Norm-Based Bound of Edelman et. al. (2022)

If we were to plug the optimal σ into our bound (still excluding the perturbation term), it would yield an optimized bound that grows like $O(m^{-\frac{1}{3}})$. If we compare this to a purely norm-based capacity bound such as in [12] for sparse boolean functions, our norm term alone shares the $O(m^{-\frac{1}{2}})$ dependency. However, the sharpness term makes our bound $O(m^{-\frac{1}{3}})$ overall, and it is therefore asymptotically dominated by the norm bound. As with many PAC-Bayes style bounds, the hope is that the flatness of the loss landscape around a learned minimum (low G_u) can be exploited in order to obtain relatively small sharpness constants even for large σ, keeping both the sharpness and norm terms numerically small for small m. While offering a strong baseline, norm-based bounds depend exclusively on global complexity parameters such as parameter norms, and are independent of the curvature of the loss landscape.

We now instantiate our generalization bound and compare it directly with the Edelman-style covering-number bound under identical architectural assumptions. We consider a target function on T bits that is a linear combination of a tiling of 2-parities. In this case

$$D_f = 2, \qquad \omega = \frac{T}{2}.$$

This function is not Boolean-valued, but both our PAC-Bayes analysis and the norm-based analysis of [12] apply unchanged. Throughout this subsection we fix

$$T = 20, \qquad \Sigma = 0.01, \qquad m = 10^6, \qquad \delta = 0.05,$$

and we use the same architectural budgets for both bounds: a 1-head, 2-layer transformer without layer normalization and hidden dimension $d = T + 2 = 22$.

H.8 Our Truncated PAC-Bayes Bound (Without Perturbation Remainder)

Ignoring the perturbation term $P(\sigma)$, our oracle PAC-Bayes bound specializes to

$$\mathrm{gap}_{\mathrm{ours}}(\sigma, m) = \underbrace{\sigma^2\, G_u(\omega, D_f)}_{\text{Sharpness term}} + \underbrace{2\sqrt{\frac{\Sigma^2}{2m}\left(\frac{L(\omega, D_f, T)^2}{2\sigma^2} + \ln\frac{1}{\delta}\right)}}_{\text{PAC-Bayes term}}. \tag{10}$$

The sharpness coefficient is

$$G_u(\omega, D_f) = 4 + 4\omega\big(2 + D_f + 32D_f^2 + 32D_f^3\big).$$

For $D_f = 2$ and $\omega = T/2 = 10$,

$$G_u = 4 + 4 \cdot 10(2 + 2 + 128 + 256) = 15524.$$

The norm proxy appearing in (10) is

$$L(\omega, D_f, T)^2 = 21 + 16(D_f + 1)D_f^2 + 8(D_f + 1) + 4\ln^2(T)\big(D_f\omega + T + 1 - \omega\big),$$

and for $(D_f, \omega, T) = (2, 10, 20)$ this evaluates to

$$L^2 = 1349.827, \qquad L \approx 36.74.$$

Substituting these values into (10) yields the explicit one-dimensional objective

$$\mathrm{gap}_{\mathrm{ours}}(\sigma, 10^6) = 15524\,\sigma^2 + 2\sqrt{\frac{10^{-4}}{2 \cdot 10^6}\Big(\frac{1349.827}{2\sigma^2} + \ln 20\Big)}.$$

Optimizing over $\sigma > 0$ numerically gives

$$\sigma^\star \approx 2.51 \times 10^{-3}.$$

Evaluating each term at $\sigma^\star$:

$$\text{Sharpness} = 15524(\sigma^\star)^2 \approx 0.09795, \qquad \text{PAC-Bayes} \approx 0.14626.$$

Thus our bound yields

$$\mathrm{gap}_{\mathrm{ours}}(m = 10^6) \approx 0.244. \tag{11}$$

H.9 Edelman et. al. for our Construction

In Theorem 14, we obtained explicit Frobenius norm bounds for each parameter matrix in our construction. To compare with the norm-based covering-number analysis of [12], we convert each Frobenius bound to a $(2, 1)$ bound using

$$\|W\|_{2,1} = \sum_{j=1}^{c} \|W_{:,j}\|_2 \;\le\; \sqrt{c}\,\|W\|_F, \qquad W \in \mathbb{R}^{r \times c}.$$

From Theorem 14 and the structure of the construction, we obtain:

$$\begin{aligned} \|F\|_{2,1} &= 4D_f\sqrt{D_f + 1}, \\ \|\Gamma\|_{2,1} &\le 2\sqrt{D_f + 1}, \\ \|M\|_{2,1} &= 4(D_f + 1), \end{aligned}$$

$$\|V^{(1)}\|_{2,1} = 1,$$
$$\|W^{(1)}\|_{2,1} \le \sqrt{T+2} \cdot 2\log(T)\sqrt{D_f\omega + (T+1-\omega)},$$
$$\|W^{(2)}\|_{2,1} \le \sqrt{T+2}\,\sqrt{2\omega\log(T)},$$
$$\|V^{(2)}\|_{2,1} \le \sqrt{\omega}.$$

For $(D_f, \omega, T, d) = (2, 10, 20, 22)$, these yield

$$C_{21}(20,22) = \|F\|_{2,1}+\|\Gamma\|_{2,1}+\|M\|_{2,1}+\|V^{(1)}\|_{2,1}+\|W^{(1)}\|_{2,1}+\|W^{(2)}\|_{2,1}+\|V^{(2)}\|_{2,1} \approx 226.26.$$

H.10 Edelman-Style Covering-Number Bound

The generalization bound of [12], when instantiated with our architectural hyperparameters and $(2,1)$-norm bounds, takes the form

$$\mathrm{gap}_{\mathrm{Edelman}}(m) \;\lesssim\; C_{21}(T,d)\,\sqrt{\frac{\ln(d\,m\,T)}{m}}. \tag{12}$$

For $T = 20,\ d = 22,\ m = 10^6$,

$$\sqrt{\frac{\ln(d\,m\,T)}{m}} = \sqrt{\frac{\ln(4.4\times 10^8)}{10^6}} \approx 4.46 \times 10^{-3},$$

and therefore

$$\mathrm{gap}_{\mathrm{Edelman}}(m = 10^6) \approx 226.26 \times 4.46 \times 10^{-3} \approx 1.01. \tag{13}$$

H.11 Summary of the Comparison

For the $T = 20$ tiling of 2-parities described above, and using identical architectural and norm budgets, we obtain:

$$\mathrm{gap}_{\mathrm{ours}}(m = 10^6) \approx 0.244, \qquad \mathrm{gap}_{\mathrm{Edelman}}(m = 10^6) \approx 1.01.$$

Thus, in this concrete low-degree, width-$\omega = T/2$ example, our PAC-Bayes flatness-based bound is numerically tighter than the Edelman-style covering-number bound, despite the latter using the same model class and norm budgets. Note that this choice of target function intentionally leverages another advantage of our bound: that it naturally applies for functions with low-degree but non-sparse Fourier spectra, a situation in which the

As mentioned above, including the perturbation term introduces a significant numerical penalty in our bound, driving the optimal σ lower, making our bound much less competitive. For target functions of high sharpness, the constants may be so large that there may not be any m below which our bound has an advantage. If our perturbation analysis were tightened – for example by using perturbations with variances that are parameter-specific – we might be able to overcome this issue.

H.12 The Outsized Role of $P(\sigma;)$

Aside from this impact on the tightness of the bound, $P(\sigma;)$ also complicates the algebra, precluding an analytic expression for the optimal σ. Furthermore, in contrast to the unique optimal σ^* monotonically decreasing in D_f, ω, T, as shown in the formula above, the argument of the minimum may exhibit non-monotonic behavior. See 5 for a surface plot showing the optimal sigma over the choices of D_f, ω that we used in our experiments. As you can see, it exhibits non-monotonic behavior in the partial order of (ω, D_f). For this reason, we opt to show our generalization bound not using the optimal sigma for each unique (ω, D_f), but rather using a single common σ that we deem to be sensible. In general, we found that using the smallest optimal σover all (ω, D_f) in our set of complexity classes used in our experiments works well, most likely because the sharpness bound increases much faster as a function of the complexity than does the norm. Whereas the largest optimal σ — generally corresponding to the simplest complexity class – tend to not work very well for the more complex functions, since the sharpness term will blow up for them, yielding a vacuous bound. In

the case of 1, we did not use the actual smallest value of σ, since our optimizer sometimes converged on our lower bound on σ of 10^{-20}, which we consider to be an artifact of the optimization process. Since these solutions also required an extremely large m, we instead opted to use a $\sigma = 10^{-9}$ which was at the low end of the bulk of the distribution of optimal σ, which was in the range $[10^{-9}, 10^{-6}]$. We then used the minimum m corresponding to that "reasonably small σ" to achieve a bound < 1 as the sample complexity used on all of the bounds shown in 1.

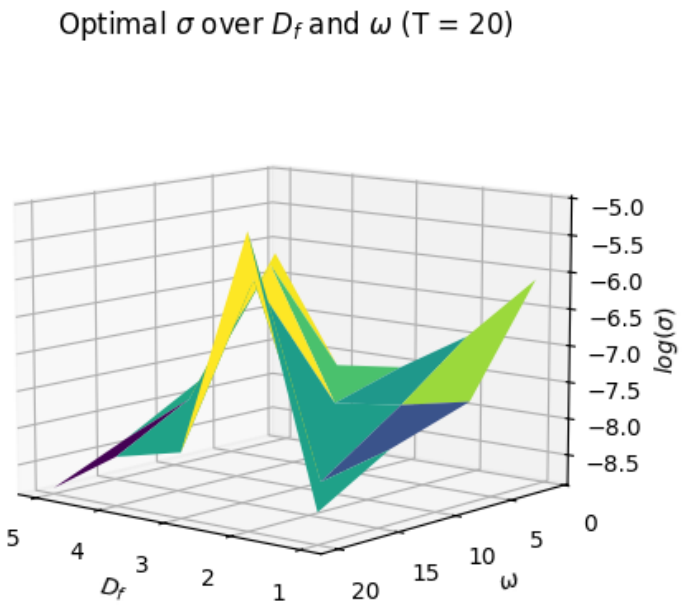

Figure 5: A surface plot showing the landscape of optimal σ over D_f, ω , with the perturbnation term. We note that the landscape is non-monotonic in the partial ordering of D_f, ω. This non-monotonicity ultimately can lead to non-monotonicity in the resulting theoretical generalization bound over the same partial order.

To analyze the impacts of the perturbation term, and to get a sense of how much room for improvement there is in our bound, 6 in the appendix shows our bound with the perturbation term truncated, and calculates the sample complexity that would be required to make the expression non-vacuous if it were actually our formal bound. In that scenario, the optimal σ would become larger, and we can get a non-vacuous bound with a sample complexity of only $m \sim 20,000$. While this analysis is informal, it highlights the fact that the interaction between the unperturbed hessian G_u and the norm L already captures a fundamental aspect of our complexity class, and requires a (pseudo-)sample complexity that is of the correct order of magnitude (our experiments used $m \sim 8,000$).

We note that the sample complexity that would result from this pseudo-bound to achieve non-vacuousness is ten orders of magnitude smaller than our actual bound, and in the range of training set sizes actually used for training this task in our experiments. This indicates that our overall approach could lead to much tighter bounds with far smaller sample complexity, if the analysis of perturbations were tightened.

Finally, we present the same bound as in 1 but with different values of σ, all of which are still in the range of the optimal σs we found over the full set of complexity classes used in our experiments. This shows the impacts of varying σ on the shape of the bound, and its vacuousness.

Theorem 11. *Let f be a target boolean function of sparsity $\omega \leq T$ and maximum degree D_f, and let $\mathcal{T}(X, \Theta)$ be a transformer of context length T implementing it exactly according to our construction.*

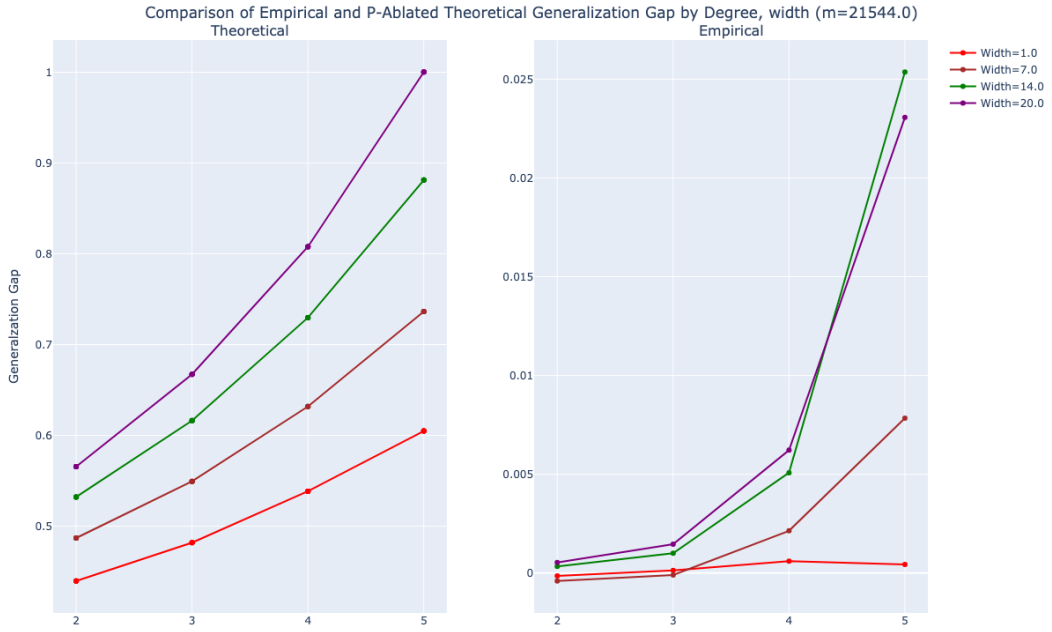

Figure 6: (left) A plot showing our psuedo-generalization bound for a fixed $\sigma = 8.8 \times 10^{-4}, m = 21544$. (right) A plot showing the average empirical generalization gap over a sample of 5 functions from each Degree, Sparsity class. We note that even without the perturbation term, our bound captures the overall shape of the generalization error found empirically. In addition, we note that the smallest reasonable σ in this scenario is four orders of magnitude larger than that of our actual bound over the same complexity classes.

Let $L_f(X, \Theta) = (T(\Theta, X) - f(X))^2$ be the unbounded, quadratic loss for our transformer learning the function f. Then under the quadratic loss, the trace of the loss Hessian is given by:

$$Tr\Big(\nabla^2 L(\mathcal{T}(X, \Theta))\Big) = 2||\nabla \mathcal{T}(X, \Theta)||^2 \in O(\omega D_f^3)$$

Proof. For a quadratic loss, the gradient of the loss is given by:

$$\nabla L_f(X, \Theta + \zeta) = 2(f(X) - \mathcal{T}(X, \Theta))\nabla \mathcal{T}(X, \Theta)$$

The Hessian is then given by:

$$\nabla^2 L_f(X, \Theta) = -2\nabla \mathcal{T}(X, \Theta)\nabla \mathcal{T}(X, \Theta)^T + 2(f(X) - \mathcal{T}(X, \Theta))\nabla^2 \mathcal{T}(X, \Theta)$$

The second term on the RHS vanishes when the test loss is 0, i.e. $f(X) = \mathcal{T}(X, \Theta)$. Therefore

$$Tr\Big(\nabla^2 L_f(X, \Theta)\Big) = -2Tr\Big(\nabla \mathcal{T}(X, \Theta)\nabla \mathcal{T}(X, \Theta)^T\Big) = 2\Big\|\nabla \mathcal{T}(X, \Theta)\Big\|^2$$

The second inequality on the right follows from the fact that the argument to the trace function is a rank-1 outer product, and thus has only a single nonzero eigenvalue making the operator norm and the trace both equal to $\Big\|\nabla \mathcal{T}(X, \Theta)\Big\|^2$. In 15, we establish the following bounds on the norms of the gradients for the exact construction.

$$||\nabla \mathcal{T}(X, \Theta)||^2 \leq G_u(\omega, D_f) := 2\omega + 8D_f^{\frac{3}{2}} + 288\omega D_f^2 + 12\omega D_f^3 \in O(\omega D_f^3)$$

It follows that

$$Tr\Big(\nabla^2 L(\mathcal{T}(X, \Theta))\Big) = 2||\nabla \mathcal{T}(X, \Theta)||^2 \leq 2G_u(\omega, D_g, T)$$
$$\in O(\omega D_f^3)$$

□

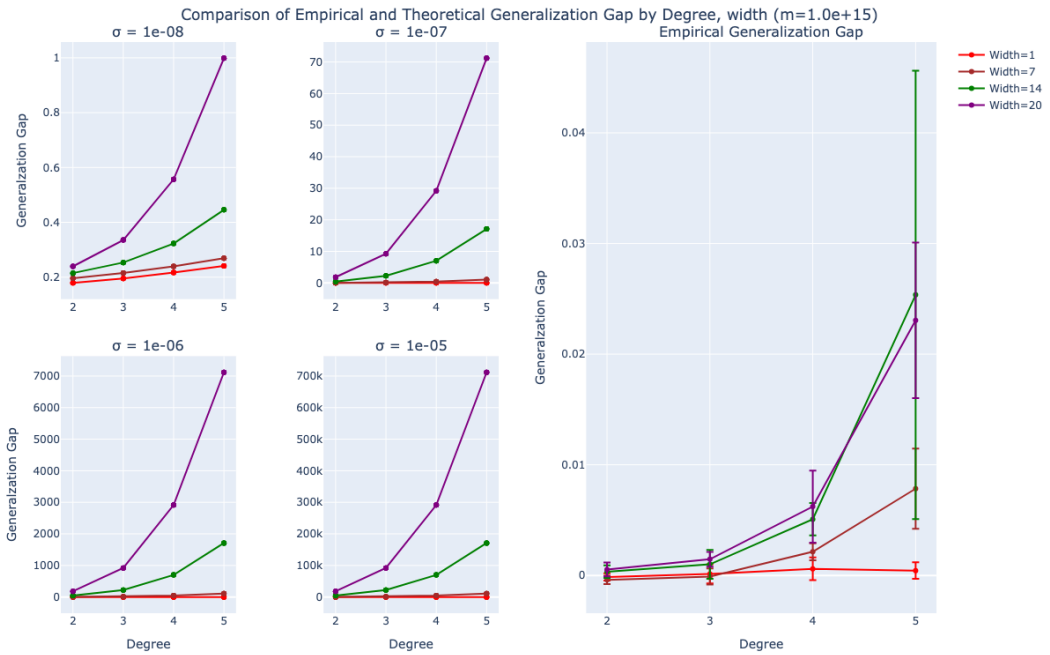

Figure 7: A plot showing the theoretical generalization gap for different values of σ that were still in the normal range of optimal sigmas for our set of function classes. Note that values of σ even slightly above the smallest of the optimal σs lead to vacuous generalization bounds, due to the stronger dependence of the sharpness term on D_f, ω than the KL term. We also include error bars in this plot, indicating one standard deviation from the mean in either direction.

Theorem 12. *Suppose our transformer $\mathcal{T}(X, \Theta + \epsilon)$ approximately represents f using a parameter set $\Theta + \epsilon$, where epsilon a normal perturbation $\epsilon \in \mathcal{N}(0, \sigma^2)^{|\Theta|}$, where $|\Theta|$ is the parameter count. Define $L(f_{\Theta})$ to be the global average (quadratic) loss of the transformer with parameters Θ evaluated against the target function f. The following equation bounds the expected global loss under ϵ with the trace of the perturbed hessian of the loss:*

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)^n}\Big[L(f_{\Theta+\epsilon})\Big] = \frac{\sigma^2}{2} Tr\Big(\nabla^2\big[L(f_{\Theta+\zeta})\big]\Big)$$

where $\zeta \in \mathbb{R}^n$ are the "remainder perturbations", i.e. $\zeta_i \in [0, \epsilon_i]$ is some small perturbation of the parameters around the true minimum that makes the Taylor expansion exact (using the Lagrange form of the remainder), conditional on each Gaussian perturbation ϵ_i.

Proof. We start off by taking a second-order Taylor expansion of $L(f_{\Theta+\epsilon})$ with the Lagrange form of the remainder:

$$L(f_{\Theta+\epsilon}) - L(f_{\Theta}) = \nabla L(f_{\Theta})\epsilon + \frac{1}{2}\epsilon^T \nabla^2\big[L(f_{\Theta+\zeta})\big]\epsilon,$$

where $\zeta \in \mathbb{R}^n$ are the "remainder perturbations", i.e. $\zeta_i \in [0, \epsilon_i]$ is some small perturbation of the parameters around the true minimum that makes the Taylor expansion exact, conditional on each Gaussian perturbation ϵ_i. Since the unperturbed transformer exactly represents the target function, we have $L(f_{\Theta}) = 0$ Assuming our function has attained a local minimum, we can set the first term on the RHS above to 0. Therefore, we can simplify the equation to:

$$L(f_{\Theta+\epsilon}) = \frac{1}{2}\epsilon^T \nabla^2\big[L(f_{\Theta+\zeta})\big]\epsilon$$

Taking the expectation of both sides with respect to ϵ, we get

$$\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)^{|\Theta|}}\Big[L(f_{\Theta+\epsilon})\Big] = \frac{1}{2}\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2)^{|\Theta|}}\Big[\epsilon^T \nabla^2\big[L(f_{\Theta+\zeta})\big]\epsilon\Big]$$

Following the proof of correctness of Hutchinson's Trace Estimator, the second term on the right is equal to $\frac{\sigma^2}{2} Tr\Big(\nabla^2\big[L(f_{\Theta+\zeta})\big]\Big)$. This leads to the following equation for the global test loss in terms of the trace of the Hessian of the loss of the transformer evaluated at the "remainder perturbations":

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)^{|\Theta|}}\Big[L(f_{\Theta+\epsilon})\Big] = \frac{\sigma^2}{2} Tr\Big(\nabla^2\big[L(f_{\Theta+\zeta})\big]\Big)$$

□

Theorem 13. *Given our perturbed transformer $\mathcal{T}(X, \Theta+\zeta)$, we can upper bound the trace of the hessian as follows:*

$$Tr\Big(\nabla^2[L_f(\mathcal{T}(X,\Theta+\zeta))]\Big) \leq 2G_u(\omega, D_f) + P(\sigma,\omega,D_f,T,d),$$

where

$$P(\sigma,\omega,D_f,T,d) :=$$
$$+2G_p(\sigma,\omega,D_f,T,d)\Big(2\sqrt{G_u(\omega,D_f)} + G_p(\sigma,\omega,D_f,T,d)\Big)$$
$$+T_p(\sigma,\omega,D_f,T,d)\Big(H_u(\omega,D_f,T,d,\sigma) + H_p(\sigma,\omega,D_f,T,d)\Big)|\Theta|$$
$$\in o\Big(\sigma D_f^5\omega^2\big(log^{\frac{7}{2}}(T) + D_f log^{\frac{5}{2}}(T)\big)\Big)$$

Proof. We need to upper bound the trace of the perturbed loss hessian, $Tr\Big(\nabla^2\big[L_f(\mathcal{T}(X,\Theta+\zeta))\big]\Big)$. Unfortunately, we cannot reuse the simple bound on the trace for the exact construction from 5, since the loss of our perturbed transformer is no longer 0, and thus the term $(f(x) - \mathcal{T}(X,\Theta+\zeta))\nabla^2\mathcal{T}(X,\Theta+\zeta)$ can not be ignored. In general, the gradient of the loss of the perturbed transformer is given by:

Hessian of the loss of the perturbed transformer is given by:

$$\nabla^2[L_f(\mathcal{T}(X,\Theta+\zeta))] = -2\nabla\mathcal{T}(X,\Theta+\zeta)\nabla\mathcal{T}(X,\Theta+\zeta)^T + 2(f(x)-\mathcal{T}(X,\Theta+\zeta))\nabla^2\mathcal{T}(X,\Theta+\zeta)$$

Therefore

$$Tr\Big(\nabla^2[L_f(\mathcal{T}(X,\Theta+\zeta))]\Big) = -2Tr\Big(\nabla\mathcal{T}(X,\Theta+\zeta)\nabla\mathcal{T}(X,\Theta+\zeta)^T\Big)$$
$$+2(f(x)-\mathcal{T}(X,\Theta+\zeta))Tr(\nabla^2\mathcal{T}(X,\Theta+\zeta))$$
$$\leq 2\Big\|\nabla\mathcal{T}(X,\Theta+\zeta)\Big\|^2 + 2(f(x)-\mathcal{T}(X,\Theta+\zeta))\Big\|\nabla^2\mathcal{T}(X,\Theta+\zeta))\Big\||\Theta|$$

For analytical tractability, we have opted to upper bound the trace of the Hessian using the maximum eigenvalue, incurring a factor of $|\Theta|$. This need not be too costly if we are using random projections, using an inner dimension of $d \in O(log(T))$. Since the number of parameters in our transformer is in $O(d^2 + dD_f)$, this translates to a total parameter count in $O(log^2(T) + D_f log(T))$. In 15 we show that $||\nabla_\Theta\mathcal{T}(X,\Theta)||_F^2 \leq G_u(\omega, D_f)$ where $G_u(\omega, D_f) \in O(\omega D_f^3)$. In 28 we demonstrate that

$$||\nabla\mathcal{T}(X,\Theta+\zeta) - \nabla\mathcal{T}(X,\Theta)|| \lessapprox G_p(\sigma,\omega,D_f,T)$$

where

$$G_p(\omega,D_f,T) \in o(\sigma D_f^{\frac{7}{2}}\omega^{\frac{3}{2}}log^2(T))$$

By the triangle inequality,

$$||\nabla\mathcal{T}(X,\Theta+\zeta)|| \lessapprox G_u(\omega,D_f) + G_p(\sigma,\omega,D_f,T)$$

In 22 it is shown that $|f(x) - \mathcal{T}(X,\Theta+\zeta)| \leq T_p(\sigma,\omega,D_f,T)$, where $T_p(\sigma,\omega,D_f,T) \in O(\sigma D_f^2\omega log(T))$

it is shown in 17 that

$$||\nabla_\Theta^2\mathcal{T}(X,\Theta)|| \leq H_u(\omega,D_f,T)$$

where

$$H_u(\omega, D_f, T, d) \in O\Big(D_f^{\frac{3}{2}}\sqrt{\omega log(T)}\Big)$$

Finally, in 29, it is shown that

$$||\nabla^2_\Theta \mathcal{T}(X, \Theta + \zeta) - \nabla^2_\Theta \mathcal{T}(X, \Theta)|| \lessapprox H_p(\sigma, \omega, D_f, T)$$
$$\in o(\sigma D_f^{\frac{7}{2}} \omega^{\frac{5}{2}} log(T)^{\frac{7}{2}})$$

Thus, by the triangle inequality, the norm of the perturbed hessian is upper bounded as

$$||\nabla^2_\Theta \mathcal{T}(X, \Theta + \zeta)|| \leq ||\nabla^2_\Theta \mathcal{T}(X, \Theta)|| + ||\nabla^2_\Theta \mathcal{T}(X, \Theta + \zeta) - \nabla^2_\Theta \mathcal{T}(X, \Theta)||$$
$$\lessapprox H_u(\omega, D_f, T) + H_p(\sigma, \omega, D_f, T).$$

Similarly, we write the square of the perturbed gradient norm in terms of the unperturbed gradient norm (keeping in mind that G_u is defined as its square) and the perturbation to the gradient norm:

$$\Big|\Big|\nabla \mathcal{T}(X, \Theta + \zeta)\Big|\Big|^2 \leq \Big(||\nabla \mathcal{T}(X, \Theta)|| + ||\nabla \mathcal{T}(X, \Theta + \zeta) - \nabla \mathcal{T}(X, \Theta)|\Big)^2$$
$$\leq \Big(\sqrt{G_u(\omega, D_f)} + G_p(\sigma, \omega, D_f, T)\Big)^2$$
$$= G_u(\omega, D_f) + 2\sqrt{G_u(\omega, D_f)} G_p(\sigma, \omega, D_f, T) + G_p(\sigma, \omega, D_f, T)^2$$

With these ingredients in mind, we can rewrite our bound for the trace of the perturbed loss hessian as

$$Tr\Big(\nabla^2[L_f(\mathcal{T}(X, \Theta + \zeta))]\Big) \leq 2G_u(\omega, D_f))$$
$$+2G_p(\sigma, \omega, D_f, T, d)\Big(2\sqrt{G_u(\omega, D_f)} + G_p(\sigma, \omega, D_f, T, d)\Big)$$
$$+T_p(\sigma, \omega, D_f, T, d)\Big(H_u(\omega, D_f, T, d, \sigma) + H_p(\sigma, \omega, D_f, T, d)\Big)|\Theta|$$

Note that all additive terms in the above expression depend on σ (and also T except for the unperturbed gradients term, which also represents the hessian of the loss for the exact transformer). Thus, we prefer to write this succinctly as

$$Tr\Big(\nabla^2[L_f(\mathcal{T}(X, \Theta + \zeta))]\Big) \leq 2G_u(\omega, D_f) + P(\sigma, \omega, D_f, T),$$

where

$$P(\sigma, \omega, D_f, T) :=$$
$$+2G_p(\sigma, \omega, D_f, T)\Big(2\sqrt{G_u(\omega, D_f)} + G_p(\sigma, \omega, D_f, T)\Big)$$
$$+T_p(\sigma, \omega, D_f, T, d)\Big(H_u(\omega, D_f, T, d, \sigma) + H_p(\sigma, \omega, D_f, T, d)\Big)|\Theta|$$
$$\in o\Big(\sigma D_f^5 \omega^2 log^2(T) + \sigma^2 D_f^7 \omega^3 log(T)^4 + \big(\sigma D_f^{\frac{7}{2}} \omega^{\frac{3}{2}} log^{\frac{3}{2}}(T) + \sigma^2 D_f^{\frac{11}{2}} \omega^{\frac{7}{2}} log(T)^{\frac{9}{2}}\big)\big(log^2(T) + D_f log(T)\big)\Big)$$

Keeping only terms that are first-order in σ, we have

$$P(\sigma, \omega, D_f, T) \in o\Big(\sigma D_f^5 \omega^2 \big(log^{\frac{7}{2}}(T) + D_f log^{\frac{5}{2}}(T)\big)\Big)$$

□

Theorem 14. *Norms of the weight matrix. Using the above construction, we have the following bound on the (vector) 2-norm of our weight matrix:*

$$||\mathbf{\Theta}||_F^2 \leq L(\omega, D_f, T) := 1 + 16(D_f + 1)D_f^2 + 8(D_f + 1) + 1 + 4log^2(T)(D_f\omega + (T + 1 - \omega))$$
$$\in O\Big(D_f^3 + log(T)^2(\omega D_f + T - \omega)\Big)$$

Proof. First, we consider the MLP matrix F. F consists of 1 column with maximum absolute value of $4D_f$. In fact, every 4-tuple of parameters corresponding to a unique function value in the domain has total squared sum of $16D_f^2$. There are $(D_f + 1)$ such quadruplets, and therefore the total squared 2-norm of this matrix is exactly $16(D_f + 1)D_f^2$.

Next, consider the bias vector Γ. This vector has $4(D_f + 1)$ entries, each of which is upper bounded by 1. Therefore $||\Gamma||_F^2 \leq 4(D_f + 1)$.

The matrix M consists of 1 row, which is all 1s. Therefore, $||M||_F^2 = 4(D_f + 1)$

For the matrix $\bar{V}^{(1)}$, all entries except the bottom right are zero. For the dimension-reduced version $V^{(1)} = J^T \bar{V}^{(1)} J$, the same is true, and that the matrix is now $(d + 1) \times (d + 1)$ with a 1 in the bottom right. Thus $||V^{(1)}|| = 1$.

The matrix $\bar{W}^{(1)} \in \mathbb{R}^{(T+2)\times(T+2)}$ has in each of its first ω columns, D_f-many elements with value $2log(T)$ representing the positional encodings of each Fourier component. In the next $T + 1 - \omega$ columns, there is a single nonzero entry of $2log(T)$ in the $(T + 1)^{th}$ row. Thus we have $D_f\omega + (T + 1 - \omega)$ entries with value $2log(T)$, yielding a Frobenius norm of $||\bar{W}^{(1)}||_F = 2log(T)\sqrt{D_f\omega + (T + 1 - \omega)}$. The dimension-reduced matrix $W^{(1)} = J^T \bar{W}^{(1)} J$ will have a Frobenius norm whose expectations is $||\bar{W}^{(1)}||_F + o(\frac{1}{k})$. To see this, we start with the definition of the Frobenius norm:

$$||J^T \bar{W}^{(1)} J||_F^2 = \sum_{r,s=1}^{d} \left(J_{:,r}^T \bar{W}^{(1)} J_{:,s}\right)^2 = \frac{1}{d^2} \sum_{r,s=1}^{d} \left(\hat{J}_{:,r}^T \bar{W}^{(1)} \hat{J}_{:,s}\right)^2,$$

where $\hat{J}$ is the un-scaled projection matrix, i.e. $\hat{J} = \sqrt{d}J$. To evaluate this, we consider the diagonal and off-diagonal terms separately. When $r = s$, the Isserlis/Wick theorem tells us that the fourth moment of the quadratic guassian form is given by

$$\mathbb{E}\big[\big(\hat{J}_{:,r}^T \bar{W}^{(1)} \hat{J}_{:,r}\big)^2\big] = 2||\bar{W}^{(1)}||_F^2 + Tr(M)^2.$$

For $r \neq s$, we have

$$\mathbb{E}\big[\big(\hat{J}_{:,r}^T \bar{W}^{(1)} \hat{J}_{:,s}\big)^2\big] = ||\bar{W}^{(1)}||_F^2.$$

Since there are $d(d-1)$ terms where $r \neq s$ and k terms where $r = s$, we put these together to obtain

$$\begin{aligned}||\hat{J}^T \bar{W}^{(1)} \hat{J}||_F^2 &= \frac{1}{d^2}\Big(d\big(2||\bar{W}^{(1)}||_F^2 + Tr(M)^2\big) + d(d-1)\big(||\bar{W}^{(1)}||_F^2\big)\Big) \\ &= ||\bar{W}^{(1)}||_F^2 + \frac{1}{d}\Big(||\bar{W}^{(1)}||_F^2 + Tr(\bar{W}^{(1)})^2\Big).\end{aligned}$$

Thus, for d large, we can approximate

$$||W^{(1)}||_F^2 = ||J^T \bar{W}^{(1)} J||_F^2 \approx ||\bar{W}^{(1)}||_F^2 = 4log^2(T)(D_f\omega + (T + 1 - \omega))$$

We can take a similar approach to bounding $||W^{(2)}||_F^2$. By the same arguments, this is well approximated by $||\bar{W}^{(2)}||_F^2$. The matrix $\bar{W}^{(2)}$ has all $0s$ except for ω entries of $log(c_t T^2)$ in the $T + 1^{th}$ (out of $T + 2$) column. Thus the squared Frobenius norm of this matrix is bounded as

$$||W^{(2)}||_F^2 \lessapprox \sum_{t=1}^{\omega} log(c_t T^2) \leq 2\omega log(T)$$

Finally, we bound $||V^{(2)}||_F^2$. Note that, like $\bar{V}^{(2)}$, $V^{(2)}$ is zero in all but the final dimension, on which J^T acts as an identity map. Since the only nonzero entry of $\bar{V}^{(2)}$ has value $D \leq \omega$, we have

$$||V^{(2)}||_F^2 \leq \omega$$

Adding these together, we have that

$$||\mathbf{\Theta}||^2 \leq$$

$$\underbrace{16(D_f + 1)D_f^2}_{||F||_F^2} + \underbrace{4(D_f + 1)}_{||\Gamma||_F^2} + \underbrace{4(D_f + 1)}_{||M||_F^2 +} \underbrace{1}_{||V^{(1)}||_F^2} + \underbrace{4log^2(T)(D_f\omega + (T + 1 - \omega))}_{||W^{(1)}||} + \underbrace{2\omega log(T)}_{||W^{(2)}||_F^2} + \underbrace{\omega}_{||V^{(2)}||_F^2}$$

$$\approx 16D_f^3 + 4log^2(T)(D_f\omega + (T+1-\omega))$$

$$\in O\Big(D_f^3 + log(T)^2(\omega D_f + T - \omega)\Big)$$

We note that the $poly(T)$ dependency in the parameter norm ultimately can be traced back to the fact that we have chosen to make the "inactive positions" in transformer (i.e. with $t > \omega$) attend to the CLS token so that they will have 0 in their final dimension (rather than the result of having all $0s$ in the columns of $\bar{W}$ corresponding to those inactive positions, resulting in a uniform attention pattern for those positions, which in turn would have caused a final dimension value which is nonzero). While leaving those columns of the attention projection matrix blank would have gotten rid of the T dependency in the norm, we will show in 28 that this would have caused the perturbations in the gradients and hessian to carry a $poly(T)$ dependency. Since we deem this to be worse than having a T dependency in the norm, we opt for this construction which "zeros out" those inactive positions, at the cost of a slightly T dependency in the parameter norms. □

Theorem 15. *Suppose our transformer construction with context length T is exactly representing some function f with maximum degree D_f and sparsity ω. Then the following bounds on the gradient norms hold uniformly for all X:*

$$||\nabla_\Theta \mathcal{T}(X,\Theta)||_F^2 \lessapprox G_u(\omega, D_f) := 2 + 8D^2 + 64\omega D_f^3 + 4D_f\omega + 64\omega D_f^3 + 2 + 128\omega D_f^2$$
$$= 2 + 4\omega(2 + D_f + 32D_f^3)$$
$$\in O(\omega D_f^3)$$

Proof. Recall that our complete transformer function is given by

$$\mathcal{T}(X, W^{(1)}, V^{(1)}, M, F, \Gamma, W^{(2)}, V^{(2)}) = (V^{(2)})^T G^T(\phi(G(W^{(2)})^T g_{T+1}))$$

where

$$g_t := F^T\Big(M^T b_t + \Gamma\Big)_+, G := \begin{bmatrix} g_1^T \\ \dots \\ g_{T+1}^T \end{bmatrix} \in \mathbb{R}^{(T+1)\times(d+1)}$$

$$b_t = (V^{(1)})^T X^T \phi(X(W^{(1)})^T x_t), B := \begin{bmatrix} b_1^T \\ \dots \\ b_{T+1}^T \end{bmatrix} \in \mathbb{R}^{(T+1)\times(d+1)}$$

Before we calculate all of the gradients, we first calculate an intermediate derivative that will be useful throughout the following sections, and will demonstrate some of the common techniques used in this paper. We first take the gradient of the transformer with respect to G. Note that we can write $g_{T+1} = (e_{T+1}^T G)^T = G^T e_{T+1}$. G still contains the one-hot positional encodings in the first d dimensions from the residual stream, but these are not dependent on any of the parameter matrices, and thus could not contribute to the gradient of the output with respect to of any of the parameter matrices. Rather, it is only the gradients with respect to $G_{t,d+1}$ that will play a role when calculating the derivatives using the chain rule. The same is true of the final dimension of the first attention sub-layer's outputs $B_{t,d+1}$: we can think of these as bottlenecks in the dependency graph for the overall transformer construction.

To give an example of how we can use these bottlenecks to create useful abstraction in the calculation of gradients, consider the gradient with respect to $W^{(1)}$.

$$\nabla_{W_1}\mathcal{T}(X,\Theta) = \sum_{t=1}^{T+1} \frac{\partial \mathcal{T}(X,\Theta)}{\partial G_{t,d+1}} \nabla_{W_1} G_{t,d+1}$$

Computing this derivative involves two tasks: first, calculating the derivative of the overall transformer output with respect to $G_{t,d+1}$ or $B_{t,d+1}$ downstream of the layer (inter-layer derivatives), and then the gradients of those key bottlenecks (scalars) with respect to the weight matrices in each layer

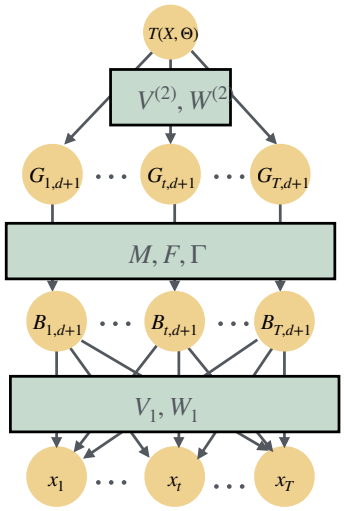

Figure 8: A diagram showing the high-level dependency graph of our 1.5-layer transformer construction.

(intra-layer gradients). To calculate the $\frac{\partial \mathcal{T}(X,\Theta)}{\partial G_{t,d+1}}$, we take the gradient with respect to the entire $G \in \mathbb{R}^{(T+1)\times(d+1)}$ matrix, and then only look at a specific row at a time, and only at the final column. This more general approach will allow us to analyze a perturbed version of our transformer, which might have have contributions from all dimensions. In these cases, we would have to consider the full Jacobian of g_t with respect to $W^{(1)}$, $\mathcal{J}_{g_t}(W^{(1)})$.

$$\nabla_{W^{(1)}} \mathcal{T}(X,\Theta) = \sum_{t=1}^{T+1} \nabla_{g_t}^T \mathcal{T}(X,\Theta) \mathcal{J}_{g_t}(W^{(1)})$$

We begin with the following finite difference equation:

$$(V^{(2)})^T (G+\Delta)^T (\phi((G+\Delta)(W^{(2)})^T (G+\Delta)^T e_{T+1})$$

$$= (V^{(2)})^T \Delta^T \phi(G(W^{(2)})^T g_{T+1}) + (V^{(2)})^T G^T (\phi((G+\Delta)(W^{(2)})^T (G^T+\Delta^T) e_{T+1})$$

First, we handle the term

$$\phi\Big((G+\Delta)(W^{(2)})^T (G^T+\Delta^T) e_{T+1}\Big) = \phi\Big(G(W^{(2)})^T (G^T+\Delta^T) e_{T+1} + \Delta(W^{(2)})^T (G^T+\Delta^T) e_{T+1}\Big)$$

$$= \phi(G(W^{(2)})^T G^T e_{T+1} + G(W^{(2)})^T \Delta^T e_{T+1} + \Delta(W^{(2)})^T G^T e_{T+1} + \Delta(W^{(2)})^T \Delta^T e_{T+1})$$

We can use a first-order Taylor approximation of ϕ to obtain:

$$\phi\Big((G+\Delta)(W^{(2)})^T (G^T+\Delta^T) e_{T+1}\Big)$$

$$= \phi(G(W^{(2)})^T g_{T+1}) + \phi'(G(W^{(2)})^T g_{T+1})(G(W^{(2)})^T \Delta^T e_{T+1} + \Delta(W^{(2)})^T G^T e_{T+1}) + O(\Delta^2)$$

Thus,

$$(V^{(2)})^T (G+\Delta)^T \phi\Big((G+\Delta)(W^{(2)})^T (G^T+\Delta^T) e_{T+1}\Big) - (V^{(2)})^T G^T \phi(G(W^{(2)})^T G^T e_{T+1})$$

$$= (V^{(2)})^T \Delta^T \phi(G(W^{(2)})^T G^T e_{T+1})$$

$$+ (V^{(2)})^T G^T \phi'(G(W^{(2)})^T G^T e_{T+1})(G(W^{(2)})^T \Delta^T e_{T+1} + \Delta(W^{(2)})^T G^T e_{T+1}) + O(\Delta^2)$$

We want to factor out the common factor of Δ^T (for reasons to be explained in the next paragraph), but note that there is an additive term of $(V^{(2)})^T G^T \phi'(G(W^{(2)})^T G^T e_{T+1}) \Delta(W^{(2)})^T G^T e_{T+1}$.

To bring out the Δ^T, we transpose this term (which is just a scalar), to get $e_{T+1}^T GW^{(2)}\Delta^T\phi'(G(W^{(2)})^T G^T e_{T+1})GV^{(2)}$. Here we have used the well-known Jacobian of the softmax function, $\phi'(x) = diag(\phi(x)) - \phi(x)\phi(x)^T$, to deduce that $\phi'(G(W^{(2)})^T G^T e_{T+1})$ is symmetric.) We can now rearrange some terms so that all of the Δ^T are on the right, and express them in terms of the trace operator:

$$= Tr\Bigg(\Big(\phi(G(W^{(2)})^T G^T e_{T+1})(V^{(2)})^T + e_{T+1}(V^{(2)})^T G^T \phi'(G(W^{(2)})^T G^T e_{T+1})G(W^{(2)})^T$$

$$+\phi'(G(W^{(2)})^T G^T e_{T+1})GV^{(2)} e_{T+1}^T GW^{(2)}\Big)\Delta^T\Bigg)$$

Now we can use the Taylor series for matrix derivatives, which says that

$$f(A+\Delta) = f(A) + df(\Delta) + O(|\Delta|^2) = f(A) + Tr((\nabla_A f)\Delta^T) + O(|\Delta|^2).$$

We can extract the gradient from the above difference equation to conclude that

$$\nabla_G \mathcal{T}(X,\Theta) = \phi_{T+1}^{(2)}(V^{(2)})^T + e_{T+1}(V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} G(W^{(2)})^T + \phi_{T+1}^{\prime(2)} GV^{(2)} e_{T+1}^T GW^{(2)}$$

Where for brevity we have written $\phi_{T+1}^{(2)} := \phi(G(W^{(2)})^T G^T e_{T+1})$ for the post-softmax attention weights for the final position in the second attention layer. To take only the last column of the matrix $\nabla_G \mathcal{T}(X,\Theta)$, we right-multiply it by e_{d+1}. To get only the t^{th} row, we left-multiply by e_t^T. The first term in this expression represents the change in the second attention layer's output under an infinitesimal change in the MLP output: if any of the MLP outputs change, we get a linear change in the final activation, with slope equal to the Fourier weight at that position, which are stored in the softmax scores, $\phi_{T+1}^{(2)}$. The second and third terms represent changes in the final activation for the CLS token due to the shifting of attention scores. However, these terms can easily be see to be zero in the $d+1^{th}$ dimension: note that both the final column and final row of $W^{(2)}$ are both 0, and therefore $W^{(2)}e_{d+1} = (W^{(2)})^T e_{d+1} = \mathbf{0}$. Intuitively, this comes from the fact that the attention matrix in the second layer is purely position-aware. Thus, we have simply:

$$e_{t+1}^T \nabla_G \mathcal{T}(X,\Theta) e_{d+1} = c_t$$

For the gradient of the MLP output $G_{t,d+1}$ with respect to W_1, we break the derivative down further, again using the chain rule:

$$\nabla_{W_1} G_{t,d+1} = \sum_{s=1}^{\omega} \frac{\partial G_{t,d+1}}{\partial B_{s,d+1}} \nabla_{W_1} B_{s,d+1}$$

However, since the MLP is applied element-wise, the partial derivative is 0 unless $t = s$, in which case the derivative is simply the slope of the MLP as a function of $B_{t,d+1}$. In our construction, this is always in the range $[-4D_f, 4D_f]$. More simply, we have

$$\nabla_{W_1} G_{t,d+1} = g_t' \nabla_{W_1} B_{t,d+1},$$

Where we have used the shorthand $g_t' := \frac{\partial G_{t,d+1}}{\partial B_{t,d+1}}$. Note that since our MLP matrices M and F are only nonzero in their last row and column (respectively), only the final dimension of $B_{t,:}$ plays a role. When we consider perturbed versions of our construction, we will have to consider the full Jacobian of g_t with respect to b_t. Our total derivative would then be

$$\nabla_{W^{(1)}} \mathcal{T}(X,\Theta) = \sum_{t=1}^{T+1}\sum_{i=1}^{d+1} \nabla_{g_t}^T \mathcal{T}(X,\Theta)\Big[\mathcal{J}_{g_t}(b_t)\Big]_{:,i} \nabla_{W^{(1)}} b_{t,i}$$

For now, since we don't consider perturbations to our construction, we can write the final gradient in terms of these three basic constituents:

$$\nabla_{W_1} \mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X,\Theta)}{\partial G_{t,d+1}} g_t' \nabla_{W_1} B_{t,d+1}$$

As a sanity check, suppose our transformer implemented the final linear combination of Fourier components using a simple linear head, $C \in \mathbb{R}^{(T+1)\times(d+1)}$ instead. Mathematically, suppose that our transformer function were defined simply as

$$\mathcal{T}(X, W, V, M, F, \Gamma, C) = \left\langle \left(M^T b_t + \Gamma \right)_+, C \right\rangle$$

where $C \in \mathbb{R}^{T\times(d+1)}$, with the Fourier coefficients c_t in the last column, and $0s$ elsewhere. In this case, we would have$\nabla_G \mathcal{T}(X, \Theta) = C$, and therefore

$$\frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,d+1}} = e_t^T C e_{d+1} = c_t, \forall t \in [\omega].$$

This is the same as the gradient of our actual construction determined above, corroborating the fact that our additional attention layer is essentially performing the same task as a simple linear head (though with notably better properties when perturbations to the parameters are considered, as we demonstrate in 28.) To continue with our example, using our construction, we have

$$\left[\nabla_{W^{(1)}} \mathcal{T}(X, \Theta)\right]_{i,j} = \sum_{t=1}^{T} g'_t c_t \left(e_i^T \nabla_{W^{(1)}} B_{t,d+1} e_j \right)$$

We can already see that the different cross derivatives in the Hessian can be broken down into gradients of these major components of the derivative. The first component g'_t only depends on the MLP matrices; the second term c_t is a function only of the second layer attention matrices; third term $e_i^T \nabla_{W^{(1)}} B_{t,d+1} e_j$ is specific to the second derivative in questions (the intra-layer gradient).

I Gradients

I.0.1 $\nabla_{V^{(2)}} \mathcal{T}(X, \Theta)$

From basic rules of matrix calculus, we have

$$\nabla_{V^{(2)}} \mathcal{T}(X, \Theta) = G^T \phi_{T+1}^{(2)}$$

We can use the matrix inequality $||Av||_2 \leq ||A||_{1,2}||v||_1$ to obtain

$$||G^T \phi_{T+1}^{(2)}|| \leq ||G||_{2,\infty} ||\phi_{T+1}^{(2)}||_1 = ||G||_{2,\infty} \leq \sqrt{2}$$

I.1 $\nabla_{W^{(2)}} \mathcal{T}(X, \Theta)$

To find the gradient $\nabla_{W^{(2)}} (V^{(2)})^T G^T \phi_{T+1}^{(2)}$, we write the following difference equations for the differential in the transformer function with respect to a small change in $W^{(2)}$.

$$(V^{(2)})^T G^T \phi(G(W^{(2)} + \Delta)^T g_{T+1})$$

$$(V^{(2)})^T G^T \left(\phi_{T+1}^{(2)} + \phi_{T+1}^{\prime(2)} G \Delta^T g_{T+1} \right) + O(\Delta^2) =$$

$$(V^{(2)})^T G^T \phi_{T+1}^{(2)} + Tr\left(g_{T+1} (V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} G \Delta^T \right) + O(\Delta^2) =$$

Once again we use the Taylor series for matrix derivatives, from which we can recover the gradient from the above difference equations:

$$\nabla_{W^{(2)}} \mathcal{T}(X, \Theta) = g_{T+1} (V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} G,$$

I.2 $\nabla_{W^{(1)}} B_{t,i}$:

Note that we have left the index i free, despite the fact that in our exact construction, only the final dimension plays a role. This will become useful when analyzing the perturbed gradients in 28. Back to our present goal, we have already taken the gradient of our transformer function with respect to $G_{t,d+1}$, and the partial derivative of the same with respect to $B_{t,d+1}$; what remains to calculate

$\nabla_{W^{(1)}}\mathcal{T}(X,\Theta)$ is the gradient $\nabla_{W^{(1)}}B_{t,d+1}$. To find it, we apply the method of finite differences. Using our definitions, we have

$$B_{t,i} = e_i^T(V^{(1)})^T X^T \phi(X(W^{(1)}+\Delta)^T)x_t$$
$$= e_i^T(V^{(1)})^T X^T\big(\phi_t + \phi_t' X\Delta^T x_t\big) + O(\Delta^2)$$

Therefore,

$$e_i^T(V^{(1)})^T X^T \phi(X(W^{(1)}+\Delta)^T)x_t - e_i^T(V^{(1)})^T X^T\big(\phi(X(W^{(1)})^T x_t)$$
$$= e_i^T(V^{(1)})^T X^T \phi_t' X\Delta^T x_t = Tr\Big(x_t e_i^T(V^{(1)})^T X^T\phi_t' X\Delta^T\Big)$$

We can now extract the gradient to conclude that

$$\nabla_{W_1}B_{t,i} = x_t e_i^T(V^{(1)})^T X^T\phi_t' X$$

I.3 $\nabla_{V^{(1)}}B_{t,i}$:

The key gradient that remains to calculate $\nabla_{V^{(1)}}\mathcal{T}(X,\Theta)$ is to calculate $\nabla_{V_1}B_{t,d+1}$. Note that

$$B_{t,d+1} = e_i^T(V^{(1)})^T X^T\phi_t.$$

We now take finite differences:

$$e_i^T(V^{(1)}+\Delta)^T X^T\phi_t = e_i^T V_1^T X^T\phi_t + e_i^T\Delta^T X^T\phi_t$$
$$= e_i^T(V^{(1)})^T X^T\phi_t + Tr\Big(X^T\phi_t e_i^T\Delta^T\Big)$$

Therefore,

$$\nabla_{V^{(1)}}B_{t,i} = \nabla_{V_1}(e_i^T(V^{(1)})^T X^T\phi_t) = X^T\phi_t e_i^T \tag{14}$$

I.4 $\nabla_M G_{t,i}$:

We now look at the gradient of $G_{t,i}$ with respect to the first matrix of the MLP, M. Recall that for each t,

$$G_{t,i} = e_i^T F^T(M^T b_t + \Gamma)_+$$

Using the rules of matrix differentiation, we have

$$\nabla_M G_{t,i} = \nabla_M e_i^T F^T max(M^T b_t + \Gamma, 0) = b_t e_i^T F^T diag(I\big[M^T b_t + \Gamma > 0\big])$$

I.5 $\nabla_\Gamma G_{t,i}$:

The unique part of the dependency graph we need for this gradient is the gradient of $G_{t,i}$ with respect to Γ. Again, using basic matrix differentiation rules, we obtain

$$\nabla_\Gamma G_{t,i} = \nabla_\Gamma e_i^T F^T(M^T b_t+\Gamma)_+ = diag(I\big[M^T b_t+\Gamma\geq 0\big])Fe_i$$

I.6 $\nabla_F G_{t,i}$:

Finally, using the fact that the matrix derivative with respect to matrix M of $x^T My$ for vectors x, y is given by xy^T, we have

$$\nabla_F G_{t,i} = \nabla_F e_i^T F^T(M^T b_t+\Gamma)_+ = \nabla_F(M^T b_t+\Gamma)_+^T Fe_{t,i} = \big(M^T b_t+\Gamma\big)_+ e_i^T$$

In summary, we have the following for the total gradient with respect to each of our parameter matrices:

$$\nabla_{V^{(2)}}\mathcal{T}(X,\Theta) = G^T\phi_{T+1}^{(2)}$$
$$\nabla_{W^{(2)}}\mathcal{T}(X,\Theta) = g_{T+1}(V^{(2)})^T G^T\phi_{T+1}^{\prime(2)}G$$
$$\nabla_M\mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega} c_t b_t e_{d+1}^T F^T diag(I\big[M^T b_t+\Gamma>0\big])$$

$$\nabla_F \mathcal{T}(X, \Theta) = \sum_{t=1}^{\omega} c_t \big(M^T b_t + \Gamma\big)_+ e_{d+1}^T$$

$$\nabla_\Gamma \mathcal{T}(X, \Theta) = \sum_{t=1}^{\omega} c_t diag(I\big[M^T b_t + \Gamma \geq 0\big]) F e_{d+1}$$

$$\nabla_{V^{(1)}} \mathcal{T}(X, \Theta) = \sum_{t=1}^{\omega} c_t g_t' X^T \phi_t e_{d+1}^T$$

$$\nabla_{W^{(1)}} \mathcal{T}(X, \Theta) = \sum_{t=1}^{\omega} c_t g_t' x_t e_{d+1}^T V_1^T X^T \phi_t' X$$

Recall that in our construction, the slope of the MLP output with respect to small changes around the points in our 1-D grid is 0, and thus we have $g_t' = \left|\frac{\partial G_{t,d+1}}{\partial B_{t,d+1}}\right| = 0$. Thus we can make the simplification that

$$\nabla_{V^{(1)}} \mathcal{T}(X, \Theta) = \mathbf{0}_{(d+1)\times(d+1)}$$

$$\nabla_{W^{(1)}} \mathcal{T}(X, \Theta) = \mathbf{0}_{(d+1)\times(d+1)}$$

We now state a useful lemma due to Deora et al. [10]:

Lemma 16. *Given an input matrix $X \in \mathbb{R}^{T\times d}$ a post-softmax Jacobian matrix $\phi_t' \in \mathbb{R}^{T\times T}$, and any vector $u_t \in \mathbb{R}^d$, the following bound holds:*

$$\left\| X^T \phi_t' X u_t \right\| = \left\| u_t^T X^T \phi_t' X \right\| \leq 2 ||X u_t||_\infty ||X||_{2,\infty}$$

The proof of this lemma is due to [10] and we defer the reader to the proof of Eq. 42 in that manuscript for further details.

J Gradient Norms

J.1 $||\nabla_{V^{(2)}} \mathcal{T}(X, \Theta)||_F$

We can use the matrix inequality $||Av||_2 \leq ||A||_{1,2} ||v||_1$ to obtain

$$||G^T \phi_{T+1}^{(2)}||_F \leq ||G||_{2,\infty} ||\phi_{T+1}^{(2)}||_1 = ||G||_{2,\infty}$$

Applying our bound on $||G||_{2,\infty}$ from 10, we obtain:

$$||\nabla_{V^{(2)}} \mathcal{T}(X, \Theta)||_F \lessapprox \sqrt{2}$$

J.2 $||\nabla_{W^{(2)}} \mathcal{T}(X, \Theta)||_F$

We calculate the norm of $\nabla_{W^{(2)}} \mathcal{T}(X, \Theta)$ by noting that the entire gradient term is a rank-1 outer product, and since the Frobenius norm of an outer product is the product of 2-norms, we have

$$||g_{T+1}(V^{(2)})^T G^T \phi_{T+1}'^{(2)} G||_F = ||g_{T+1}|| ||(V^{(2)})^T G^T \phi_{T+1}'^{(2)} G||$$

By 16, we can bound the second multiplicative term as:

$$||(V^{(2)})^T G^T \phi_T' G|| \leq 2 ||(V^{(2)})^T G^T||_\infty ||G||_{2,\infty}$$

In 10, we show that $||GV^{(2)}||_\infty \lessapprox D$, $||G||_{2,\infty} \lessapprox \sqrt{2}$, and $||g_{T+1}|| \lessapprox$ It follows that

$$||\nabla_{W^{(2)}} \mathcal{T}(X, \Theta)||_F \lessapprox 2\sqrt{2} D$$

J.3 $||\nabla_M \mathcal{T}(X,\Theta)||_F$

We start with the triangle inequality:

$$||\nabla_M \mathcal{T}(X,\Theta)||_F \leq \sum_{t=1}^{\omega} \left\| c_t b_t e_{d+1}^T F^T diag(I\big[M^T b_t + \Gamma > 0\big]) \right\|_F$$

$$= \sum_{t=1}^{\omega} |c_t| \|b_t\| \left\| diag(I\big[M^T b_t + \Gamma > 0\big]) F e_{d+1} \right\|$$

In 10, we show that $||b_t|| \leq 1$. Additionally we have

$$\left\| diag(I\big[M^T b_t + \Gamma > 0\big]) F e_{d+1} \right\| \leq \left\| F e_{d+1} \right\| = 4D_f \sqrt{4(D_f+1)}$$

We can now apply Cauchy-Schwarz to conclude:

$$||\nabla_M \mathcal{T}(X,\Theta)||_F \leq 8D_f \sqrt{\omega(D_f+1)} \approx 8D_f \sqrt{\omega D_f}$$

We will find it useful for future reference to state the norm of the intra-layer gradient term by itself. The intra-layer gradient is given by

$$\nabla_M G_{t,i} = b_t e_i^T F^T diag(I\big[M^T b_t + \Gamma > 0\big]).$$

Note that in the above calculation we restricted the sum to $t \in [\omega]$ because we knew that $c_t = 0$ for $t > \omega$. However, observe that the intra-layer gradient is also 0 when $t > \omega$. Recall that in the first attention layer, all positions $t > \omega$ attend to the CLS token, which has a final bit value of 0. Since the $V^{(1)}$ matrix only selects the final dimension, and there is no residual connection after the attention layer, the b_t in these inactive postions are all the $0-$vector. Thus for $T > \omega$, we actually have $||b_t|| = 0, \forall t \in [\omega+1, T+1]$. [4]

Thus, following the same arguments above yields

$$\left\| \nabla_M G_{t,i} \right\|_F \lessapprox 8D_f^{\frac{3}{2}} I[i = d+1 \wedge t \leq \omega]$$

J.4 $||\nabla_F \mathcal{T}(X,\Theta)||_F$

$$||\nabla_F \mathcal{T}(X,\Theta)||_F \leq \sum_{t=1}^{\omega} |c_t| \left\| \big(M^T b_t + \Gamma\big)_+ e_{d+1}^T \right\|_F \leq \sum_{t=1}^{\omega} |c_t| \left\| \big(M^T b_t + \Gamma\big)_+ \right\|$$

$$||\nabla_F \mathcal{T}(X,\Theta)||_F \leq 2\sqrt{(D_f+1)\omega} \approx 2\sqrt{D_f \omega}$$

We once again provide the norm of just the intra-layer gradient for later use. This gradient is:

$$||\nabla_F G_{t,i}|| \leq ||\big(M^T b_t + \Gamma\big)_+ e_i^T|| \lessapprox 2\sqrt{D_f}$$

Just as in the gradient with respect to M, when $b_t = \mathbf{0}_{d+1}$, the above expression becomes 0. Thus we make this explicit:

$$||\nabla_F G_{t,i}|| \lessapprox 2\sqrt{D_f} I[t \leq \omega]$$

[4] Indeed, the design decision to have non-active positions attend to the CLS token and avoid a residual connection after the attention layer were made specifically to avoid a T dependency in the gradient bounds. We could have chosen to leave the rows of $\bar{W}^{(1)}$ in $[\omega, T]$ equal to 0, and the result would have been eliminating the $O(\sqrt{T})$ dependency in the parameter norms, but we would have gained a factor of T in the perturbed gradient. We see the latter construction as preferable. Finally, we note that a the same goal could be accomplished with a residual connection after the first attention layer, and then having the MLP "clean up" the MLP output, setting $G_{t,d+1}$ to 0. This concstruction is more involved, so we prefer to "clean up" these unused activations by suing the CLS token and avoiding the first residual connection.

J.5 $||\nabla_\Gamma \mathcal{T}(X,\Theta)||_F$

$$||\nabla_\Gamma \mathcal{T}(X,\Theta)||_F \leq \sum_{t=1}^{\omega} |c_t| \Big\| diag(I\big[M^T b_t + \Gamma \geq 0\big]) F e_{d+1} \Big\|$$

$$||\nabla_\Gamma \mathcal{T}(X,\Theta)||_F \leq 8D_f \sqrt{\omega(D_f+1)} \approx 8D_f\sqrt{\omega D_f}$$

Recall that $\nabla_\Gamma G_{t,i} = diag(I\big[M^T b_t + \Gamma \geq 0\big]) F e_i$. Note that $Fe_i = \mathbf{0}_{4(D_f+1)}$ if $i \neq d+1$. And just as in the gradient with respect to M, when $b_t = \mathbf{0}_{d+1}$, the above expression becomes 0. Thus we have

$$||\nabla_\Gamma G_{t,i}|| \lessapprox 8D_f^{\frac{3}{2}} I[i = d+1 \wedge t \leq \omega]$$

J.6 $||\nabla_{V^{(1)}} \mathcal{T}(X,\Theta)||_F$

We showed that the gradient with respect to $V^{(1)}$ was a 0-matrix due to the $\left|\frac{\partial \mathcal{T}(X,\Theta)}{\partial B_{t,d+1}}\right|$ component being 0, and therefore $||\nabla_{V^{(1)}} \mathcal{T}(X,\Theta)||_F = 0$. Nonetheless, we will find it useful when calculating gradients for a perturbed construction to bound the norm of the intra-layer gradient. Recall that

$$\nabla_{V^{(1)}} B_{t,i} = X^T \phi_t e_i^T$$

The Frobenius norm of this rank-1 outer product is the product of the 2-norms:

$$\Big\| \nabla_{V^{(1)}} B_{t,i} \Big\|_F = \Big\| X^T \phi_t \Big\| \leq ||X^T||_{1,2} ||\phi_t||_1 = ||X||_{2,\infty} \lessapprox \sqrt{2}$$

where we have applied our bound on $||X||_{2,\infty}$ in 10.

J.7 $||\nabla_{W^{(1)}} \mathcal{T}(X,\Theta)||_F$

Similarly, it was already shown that the total gradient with respect to $W^{(1)}$ is the 0-matrix, and therefore $||\nabla_{W^{(1)}} \mathcal{T}(X,\Theta)||_F = 0$, we still bound the norm of the intra-layer gradient for future reference. Recall that

$$\nabla_{W_1} B_{t,i} = x_t e_i^T (V^{(1)})^T X^T \phi_t' X$$

This is once again a rank-1 outer product, and its frobenius norm is thus bounded by

$$\Big\| \nabla_{W_1} B_{t,i} \Big\|_F = \Big\| x_t \Big\|_2 \Big\| e_i^T (V^{(1)})^T X^T \phi_t' X \Big\|_2 \leq \Big\| X \Big\|_{2,\infty} \Big\| e_i^T (V^{(1)})^T X^T \phi_t' X \Big\|_2$$

Recall that the first d rows and columns of $V^{(1)}$ are all $0s$, and therefore $e_i^T (V^{(1)})^T = \mathbf{0}$ if $i \neq d+1$. Thus we have

$$\Big\| e_i^T (V^{(1)})^T X^T \phi_t' X \Big\|_2 = I[i = d+1] \Big\| e_{d+1}^T (V^{(1)})^T X^T \phi_t' X \Big\|_2$$

$$\leq 2I[i = d+1] ||X||_{2,\infty} ||XV^{(1)} e_{d+1}||_\infty,$$

where the last inequality follows from 16. Using 10, we conclude that

$$\Big\| \nabla_{W_1} B_{t,i} \Big\|_F \lessapprox 2\sqrt{2} I[i = d+1].$$

In summary, we have

$$||\nabla_\Theta \mathcal{T}(X,\Theta)||_F^2 \lessapprox 2 + 8D^2 + 64\omega D_f^3 + 4D_f\omega + 64\omega D_f^3 + 2 + 128\omega D_f^2$$

$$= 4 + 4\omega(2 + D_f + 32D_f^2 + 32D_f^3)$$

$$\in O(\omega D_f^3)$$

J.8 $||\nabla_{W^{(1)}}\mathcal{T}(X,\Theta)||_F$

In summary, we have

$$\begin{aligned}||\nabla_\Theta \mathcal{T}(X,\Theta)||_F^2 &\lessapprox 2 + 8D^2 + 64\omega D_f^3 + 4D_f\omega + 64\omega D_f^3 \\ &= 2 + 4\omega(2 + D_f + 32D_f^3) \\ &\in O(\omega D_f^3)\end{aligned}$$

□

Theorem 17. *Suppose our transformer construction with context length T is exactly representing some function f with maximum degree D_f and sparsity ω. Then the following bound on the operator norm of the hessian holds uniformly for all X.*

$$||\nabla_\Theta^2 \mathcal{T}(X,\Theta)|| \lessapprox H_u(\omega, D_f, T, d)$$

where we have defined the function $H_u(\omega, D_f, T, d) :=$

$$\begin{aligned}&8(D_f+1) + 6(D_f+1)\sqrt{\omega} + 2\sqrt{3T(D_f+1)} + \sqrt{48T(D_f+1)}D_f \\ &+\sqrt{T}\sqrt{12(D_f+1)}D_f + \sqrt{T}\sqrt{12(D_f+1)}D_f + 2\sqrt{T}\sqrt{12(D_f+1)}D_f \\ &+8\sqrt{2}\sqrt{dT}D_f + 2\sqrt{dT}D_f + 8\sqrt{2}dD_f\sqrt{\omega} + 6\sqrt{2}d\sqrt{d}D_f \\ &\in O(T^{\frac{1}{2}}D_f^{\frac{3}{2}} + \omega^{\frac{1}{2}}D_f log(T))\end{aligned}$$

Proof. **K Second Derivatives**

It is worth noting in advance of our calculation of the Hessian that, while conceivably there could be $\binom{7}{2} + 7 = 28$ terms or blocks in the Hessian, the above structure for the gradients implies that many of the second derivative terms will be 0. To understand which cross derivatives will vanish, we first recall that $\nabla_{V^{(1)}}\mathcal{T}(X,\Theta)$ and $\nabla_{W^{(1)}}\mathcal{T}(X,\Theta)$ are both 0 matrices. This removes 13 cross derivatives.

Clearly, $\nabla_{W^{(2)}}\mathcal{T}(X,\Theta)$ does not depend on $W^{(2)}$, so $\nabla^2_{W^{(2)}}\mathcal{T}(X,\Theta)$ is 0. But if we look closer at this gradient, we can see that almost all of their other cross derivatives are also 0: recall that the first d columns of G, the entire vector g_{T+1}, and the entire CLS attention vector $\phi^{(2)}_{T+1}$ (and therefore also $\phi'^{(2)}_{T+1} = diag(\phi^{(2)}_{T+1}) - \phi^{(2)}_{T+1}(\phi^{(2)}_{T+1})^T$) depend only on positional information. Thus, the gradient is 0 with respect to all other parameter matrices except for $V^{(2)}$. This eliminates 4 additional cross derivatives.

Similarly, it is clear that $\nabla_{V^{(2)}}\mathcal{T}(X,\Theta)$ does not depend on $V^{(2)}$ and therefore the second derivative with respect to $V^{(2)}$ will be 0. The matrix $G^T\phi^{(2)}_{T+1}$ does indeed depend on the MLP outputs, but this eliminates one additional cross-derivative.

Regarding cross-derivatives among the MLP matrices, note that $\nabla_F\mathcal{T}(X,\Theta)$ does not depend on F, and thus that second derivative is zero. Further, both $\nabla_\Gamma\mathcal{T}(X,\Theta)$ and $\nabla_M\mathcal{T}(X,\Theta)$ depend on the M and Γ only inside of an indicator function, which has a slope of 0 almost everywhere. Thus, 4 more (unique) cross derivatives vanish, which means that we will need to consider no more than $28 - 13 - 4 - 1 - 4 = 6$ nonzero terms in the Hessian.

1)

K.1 $\nabla^2_{W^{(2)}V^{(2)}}\mathcal{T}(X,\Theta)$:

The gradient with respect to $V^{(2)}$ is given by $\left[\nabla_{V^{(2)}}\mathcal{T}(X,\Theta)\right]_i = (G_{:,i})^T\phi^{(2)}_{T+1}$. Recall that the last element of $\phi^{(2)}_{T+1}$ representing the CLS token is 0. Since the first d columns of G are independent of the parameter matrices, and $\phi^{(2)}_{T+1}$ depends only on positional information and the matrix $W^{(2)}$, all

cross derivatives with $V^{(2)}$ except this one will all be zero whenever $i \neq d+1$. In this case however, the derivative is nonzero for $1 \leq i \leq d+1$.

$$\nabla_{W^{(2)}}\Big(\Big[\nabla_{V^{(2)}}\mathcal{T}(X,\Theta)\Big]_i\Big) = \sum_{t=1}^{T+1} G_{t,i}\nabla_{W^{(2)}}[\phi_{T+1}^{(2)}]_t$$

Note that

$$[\phi_{T+1}^{(2)}]_t = e_t^T\phi(G(W^{(2)})^T g_{T+1})$$

Taking finite differences, we have

$$e_t^T\phi(G(W^{(2)}+\Delta)^T g_{T+1}) = e_t^T\phi(G(W^{(2)})^T g_{T+1}) + e_t^T\phi'(G(W^{(2)})^T g_{T+1})G\Delta^T g_{T+1}$$

$$= e_t^T\phi(G(W^{(2)})^T g_{T+1}) + Tr\Big(g_{T+1}e_t^T\phi_{T+1}^{\prime(2)}G\Delta^T\Big)$$

From here we can conclude that $\nabla_{W^{(2)}}[\phi_{T+1}^{(2)}]_t = g_{T+1}e_t^T\phi_{T+1}^{\prime(2)}G$, and therefore

$$\nabla_{W^{(2)}}\Big(\Big[\nabla_{V^{(2)}}\mathcal{T}(X,\Theta)\Big]_i\Big) = \sum_{t=1}^{T+1} G_{t,i}\nabla_{W^{(2)}}[\phi_{T+1}^{(2)}]_t = \sum_{t=1}^{\omega} G_{t,i}g_{T+1}e_t^T\phi_{T+1}^{\prime(2)}G$$

Following the exact same arguments, we can immediately write the following 5 additional cross-derivatives involving $V^{(2)}$, noting that the cross-derivatives are all 0 for $i < d+1$.

2)

K.2 $\nabla^2_{MV^{(2)}}\mathcal{T}(X,\Theta)$:

Recall that all cross derivatives with $V^{(2)}$ will be zero whenever $i \neq d+1$. When $i = d+1$, we have

$$e_{d+1}^T G^T\phi_{T+1}^{(2)} = \sum_{t=1}^{\omega} G_{t,d+1}c_t = \mathcal{T}(X,\Theta)$$

$$\nabla_M\Big(\Big[\nabla_{V^{(2)}}\mathcal{T}(X,\Theta)\Big]_i\Big) = \begin{cases}\mathbf{0}_{(d+1)\times 4(D_f+1)}, & i \leq d \\ \nabla_M\mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega} c_t b_t e_{d+1}^T F^T diag(I\big[M^T b_t + \Gamma > 0\big]), & i = d+1\end{cases}$$

3)

K.3 $\nabla^2_{FV^{(2)}}\mathcal{T}(X,\Theta)$:

$$\nabla_F\Big(\Big[\nabla_{V^{(2)}}\mathcal{T}(X,\Theta)\Big]_i\Big) = \begin{cases}\mathbf{0}_{4(D_f+1)\times(d+1)}, & i \leq d \\ \nabla_F\mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega} c_t\big(M^T b_t + \Gamma\big)_+ e_{d+1}^T, & i = d+1\end{cases}$$

4)

K.4 $\nabla^2_{\Gamma V^{(2)}}\mathcal{T}(X,\Theta)$:

$$\nabla_\Gamma\Big(\Big[\nabla_{V^{(2)}}\mathcal{T}(X,\Theta)\Big]_i\Big) = \begin{cases}\mathbf{0}_{4(D_f+1)}, & i \leq d \\ \nabla_\Gamma\mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega} c_t e_{d+1}^T F^T diag(I\big[M^T b_t + \Gamma \geq 0\big]), & i = d+1\end{cases}$$

5)

K.5 $\nabla^2_{\Gamma F}\mathcal{T}(X,\Theta)$:

$$\nabla_F \mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega} c_t \big(M^T b_t + \Gamma\big)_+ e_{d+1}^T$$

$$e_i^T \nabla_F \mathcal{T}(X,\Theta) e_j = \sum_{t=1}^{\omega} c_t e_i^T \big(M^T b_t + \Gamma\big)_+ e_{d+1}^T e_j$$

$$\nabla_\Gamma \Big(e_i^T \nabla_F \mathcal{T}(X,\Theta) e_j\Big) = I\big[j = d+1\big] \sum_{t=1}^{\omega} c_t e_i I\big[[M^T b_t + \Gamma]_i > 0\big]$$

6)

K.6 $\nabla^2_{FM}\mathcal{T}(X,\Theta)$:

From above, we have

$$\nabla_M \mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega} c_t b_t e_{d+1}^T F^T diag(I\big[M^T b_t + \Gamma > 0\big])$$

$$\nabla_F \Big(\big[\nabla_M \mathcal{T}(X,\Theta)\big]_{i,j}\Big) = \sum_{t=1}^{\omega} c_t \nabla_F \Big(e_i^T b_t e_{d+1}^T F^T diag(I\big[M^T b_t + \Gamma > 0\big] e_j)\Big)$$

$$\nabla_F \Big(\big[\nabla_M \mathcal{T}(X,\Theta)\big]_{i,j}\Big) = \sum_{t=1}^{\omega} c_t (e_i^T b_t) \nabla_F \left(e_j^T diag\Big(I\big[M^T b_t + \Gamma > 0\big]\Big) F e_{d+1}\right)$$

$$\nabla_F \Big(\big[\nabla_M \mathcal{T}(X,\Theta)\big]_{i,j}\Big) = \sum_{t=1}^{\omega} c_t (e_i^T b_t) diag\Big(I\big[M^T b_t + \Gamma > 0\big]\Big) e_j e_{d+1}^T$$

Note that we can take the scalar $e_i^T b_t$ outside of the ∇_F operator in the third equation above because b_t is upstream of F.

L Bounding Maximum Eigenvalues

We can express the complete hessian as a $\Big(\big(3(d+1)^2 + (d+1) + (8d+6)(D_f+1)\big) \times \big(3(d+1)^2 + (d+1) + (8d+6)(D_f+1)\big)\Big)$ block matrix. To bound the maximum eigenvalue of the Hessian, we want to find $\max_{||\mathbf{v}||=1} \langle \mathbf{v}, \nabla^2_\Theta \mathcal{T}(X,\Theta)\mathbf{v}\rangle$, where

$$\mathbf{v} = concat(\mathbf{s}, \mathbf{p}, \mathbf{q}, \mathbf{r}, \eta, \gamma, \nu,),$$

$$= concat(s_1, \ldots, s_{d+1}, p_1, \ldots, p_{d+1}, q_1, \ldots, q_{d+1}, r_1, \ldots r_{d+1}, \eta_1, \ldots, \eta_{d+1}, \gamma_1, \nu_1, \ldots, \nu_{4(D_f+1)}),$$

where $s_i \in \mathbb{R}$,$p_i, q_i, r_i, \nu_i \in \mathbb{R}^{d+1}, \eta_i, \gamma_1 \in \mathbb{R}^{4(D_f+1)}$. Written out explicitly, we seek to upper bound

$$\nabla^2_\Theta \big\|\mathcal{T}(X,\Theta)\big\| = \max_{\|v\|=1} \begin{pmatrix} \mathbf{s} \\ \mathbf{p} \\ \mathbf{q} \\ \mathbf{r} \\ \boldsymbol{\eta} \\ \boldsymbol{\nu} \\ \boldsymbol{\gamma} \end{pmatrix}^T \begin{pmatrix} \nabla^2_{V^{(2)}V^{(2)}} & \nabla^2_{V^{(2)}W^{(2)}} & \cdots & \nabla^2_{V^{(2)}F} \\ \nabla^2_{W^{(2)}V^{(2)}} & \nabla^2_{W^{(2)}W^{(2)}} & \cdots & \nabla^2_{W^{(2)}F} \\ \vdots & \vdots & \ddots & \vdots \\ \nabla^2_{FV^{(2)}} & \nabla^2_{FW^{(2)}} & \cdots & \nabla^2_{FF} \end{pmatrix} \begin{pmatrix} \mathbf{s} \\ \mathbf{p} \\ \mathbf{q} \\ \mathbf{r} \\ \boldsymbol{\eta} \\ \boldsymbol{\nu} \\ \boldsymbol{\gamma} \end{pmatrix} \tag{15}$$

Therefore,

$$||\nabla^2_\Theta \mathcal{T}(X,\Theta)|| \leq$$

$$\underbrace{+2\max_{||v||=1} concat(s_1,...s_{d+1})^T\Big(\nabla^2_{W^{(2)}V^{(2)}}\mathcal{T}(X,\Theta)\Big)concat(p_1,...,p_{d+1})}_{Term_1}$$

$$\underbrace{+2\max_{||v||=1} concat(s_1,...,s_{d+1})^T\Big(\nabla^2_{MV^{(2)}}\mathcal{T}(X,\Theta)\Big)concat(\eta_1,...\eta_{d+1})}_{Term_2}$$

$$\underbrace{+2\max_{||v||=1} concat(s_1,...,s_{d+1})^T\Big(\nabla^2_{\Gamma C}\mathcal{T}(X,\Theta)\Big)\gamma_1}_{Term_3}$$

$$\underbrace{+2\max_{||v||=1} concat(s_1,...,s_{d+1})^T\Big(\nabla^2_{FV^{(2)}}\mathcal{T}(X,\Theta)\Big)concat(\nu_1,...\nu_{4(D_f+1)})}_{Term_4}$$

$$\underbrace{+2\max_{||v||=1} \gamma_1^T\Big(\nabla^2_{F\Gamma}\mathcal{T}(X,\Theta)\Big)concat(\nu_1,...\nu_{4(D_f+1)})}_{Term_5}$$

$$\underbrace{+2\max_{||v||=1} concat(\eta_1,\ldots,\eta_{d+1})^T\Big(\nabla^2_{FM}\mathcal{T}(X,\Theta)\Big)concat(\nu_1,...\nu_{4(D_f+1)})}_{Term_6}$$

L.1 $Term_1$:

We have

$$Term_1 = 2\max_{||v||=1} concat(s_1,...s_{d+1})^T\Big(\nabla^2_{W^{(2)}V^{(2)}}\mathcal{T}(X,\Theta)\Big)concat(p_1,...,p_{d+1})$$

$$= 2max_{||v||=1}\sum_{i=1}^{d+1}\sum_{t=1}^{\omega} s_i Vec\Big(G_{t,i}g_{T+1}e_t^T\phi'^{(2)}_{T+1}G\Big)\mathbf{p}$$

$$= 2max_{||v||=1}\sum_{i=1}^{d+1}\sum_{t=1}^{\omega}\sum_{k=1}^{d+1} s_i G_{t,i}G_{T+1,k}e_t^T\phi'^{(2)}_{T+1}G\mathbf{p_k}$$

$$\leq 2\sum_{i=1}^{d+1}\sum_{t=1}^{\omega}\sum_{k=1}^{d+1}|s_i|||G||^2_{1,\infty}\Big\|e_t^T\phi'^{(2)}_{T+1}G\Big\|||\mathbf{p_k}||$$

By 16, we know that $\Big\|e_t^T\phi'^{(2)}_{T+1}G\Big\| \leq 2||e_t||_\infty||G||_{2,\infty} \leq 2||G||_{2,\infty}$. By **??** we therefore have

$$Term_1 \lessapprox 4\sqrt{2}d\omega$$

L.2 $Term_2$:

$$2\max_{||v||=1} concat(s_1,...,s_{d+1})^T\Big(\nabla^2_{MV^{(2)}}\mathcal{T}(X,\Theta)\Big)concat(\eta_1,...\eta_{d+1})$$

$$= 2\max_{||v||=1} s_{d+1}\sum_{t=1}^{\omega} Vec\Big(c_t b_t e_{d+1}^T F^T diag(I\big[M^T b_t+\Gamma>0\big])\Big)\eta$$

$$= 2\max_{||v||=1} s_{d+1}\sum_{t=1}^{\omega}\sum_{k=1}^{d+1} c_t b_{tk} e_{d+1}^T F^T diag(I\big[M^T b_t+\Gamma>0\big])\eta_k$$

$$\leq 2|s_{d+1}|||B||_{1,\infty}\sum_{t=1}^{\omega}\sum_{k=1}^{d+1}|c_t|\Big\|e_{d+1}^T F^T diag(I\big[M^T b_t+\Gamma>0\big])\Big\|||\eta_k||$$

$$\leq 2|s_{d+1}|||B||_{1,\infty}\sum_{t=1}^{\omega}\sum_{k=1}^{d+1}|c_t|||F_{:,d+1}||||\eta_k||$$

The entries of the final column of F in our construction are in $[-4D_f, 4D_f]$. So $||F_{:,d+1}|| \leq 4D_f\sqrt{4(D_f+1)}$. Using 10 to bound $||B||_{2,\infty}$, we get

$$Term_2 \leq 16D_f\sqrt{2(D_f+1)\omega(d+1)}$$

L.3 $Term_3$**:**

$$Term_3 = 2\max_{||v||=1} concat(s_1,...,s_{d+1})^T\Big(\nabla^2_{\Gamma V^{(2)}}\mathcal{T}(X,\Theta)\Big)\gamma_{\mathbf{1}}$$

$$= 2\max_{||v||=1} s_{d+1}\sum_{t=1}^{\omega}c_t Vec\Big(diag(I\big[M^Tb_t+\Gamma\geq 0\big])Fe_{d+1}\Big)\gamma_{\mathbf{1}}$$

$$= 2\max_{||v||=1} s_{d+1}\sum_{t=1}^{\omega}\sum_{k=1}^{d+1}c_t e_k^T diag(I\big[M^Tb_t+\Gamma\geq 0\big])Fe_{d+1}|\gamma_{1k}|$$

$$\leq 2|s_{d+1}|\sum_{t=1}^{\omega}\sum_{k=1}^{d+1}|c_t|||e_k||||Fe_{t,d+1}|||\gamma_{1k}|$$

$$\leq 8D_f\sqrt{4(D_f+1)\omega(d+1)}$$

L.4 $Term_4$**:**

$$Term_4 = 2\max_{||v||=1} concat(s_1,...,s_{d+1})^T\Big(\nabla^2_{FV^{(2)}}\mathcal{T}(X,\Theta)\Big)concat(\nu_1,...\nu_{4(D_f+1)})$$

$$= 2\max_{||v||=1} s_{d+1}\sum_{t=1}^{\omega}c_t Vec\Big(\big(M^Tb_t+\Gamma\big)_+e_{t,d+1}^T\Big)\nu$$

$$= 2\max_{||v||=1} s_{d+1}\sum_{t=1}^{\omega}\sum_{k=1}^{4(D_f+1)}c_t\Big[\big(M^Tb_t+\Gamma\big)_+\Big]_k e_{t,d+1}^T\nu_{\mathbf{k}}$$

$$\leq 2|s_{d+1}|\sum_{t=1}^{\omega}\sum_{k=1}^{4(D_f+1)}|c_t|\Big|\Big[\big(M^Tb_t+\Gamma\big)_+\Big]_k\Big|||\nu_{\mathbf{k}}||$$

Note that $\Big[\big(M^Tb_t+\Gamma\big)_+\Big]_k$ is upper bounded by 1, and therefore we have that

$$Term_4 \leq 4\sqrt{\omega(D_f+1)}$$

L.5 $Term_5$**:**

$$2\max_{||v||=1}\gamma_1^T\Big(\nabla^2_{F\Gamma}\mathcal{T}(X,\Theta)\Big)concat(\nu_1,...,\nu_{4(D_f+1)})$$

$$= 2\max_{||v||=1} concat(\nu_1,...,\nu_{4(D_f+1)})^T\Big(\nabla^2_{\Gamma F}\mathcal{T}(X,\Theta)\Big)\gamma_1$$

$$= 2\max_{||v||=1}\sum_{i=1}^{d+1}\sum_{j=1}^{4(D_f+1)}\nu_{ij}I\big[i=d+1\big]\sum_{t=1}^{\omega}c_t Vec\Big(e_jI\big[[M^Tb_t+\Gamma]_j>0\big]\Big)\gamma_{\mathbf{1}}$$

$$= 2\max_{||v||=1}\sum_{i=1}^{d+1}\sum_{j=1}^{4(D_f+1)}\nu_{ij}I\big[i=d+1\big]\sum_{t=1}^{\omega}\sum_{k=1}^{4(D_f+1)}I\big[k=j\big]c_tI\big[[M^Tb_t+\Gamma]_j>0\big]\gamma_{1k}$$

$$= 2 \max_{||v||=1} \sum_{j=1}^{4(D_f+1)} \nu_{d+1,j} \sum_{t=1}^{\omega} c_t I\big[[M^T b_t + \Gamma]_j > 0\big] \gamma_{1j}$$

$$\leq 2\sqrt{\omega} \max_{||v||=1} \sum_{j=1}^{4(D_f+1)} |\nu_{d+1,j}\gamma_{1j}|$$

$$\leq 4\sqrt{\omega(D_f+1)}$$

In the above chain of inequalities we used Cauchy-Schwarz multiple times, as well as the fact that the 2-norm of the Fourier coefficients is 1.

L.6 $Term_6$:

$$2 \max_{||v||=1} concat(\eta_1, \ldots, \eta_{d+1})^T \Big(\nabla^2_{FM} \mathcal{T}(X, \Theta)\Big) concat(\nu_1, ...\nu_{4(D_f+1)})$$

$$= 2 \max_{||v||=1} \sum_{j=1}^{4(D_f+1)} \sum_{i=1}^{d+1} \eta_{ij} \sum_{t=1}^{\omega} Vec\Big(c_t (e_i^T b_t) diag\Big(I\big[M^T b_t + \Gamma > 0\big]\Big) e_j e_{d+1}^T\Big)\nu$$

$$= 2 \max_{||v||=1} \sum_{j=1}^{4(D_f+1)} \sum_{i=1}^{d+1} \eta_{ij} \sum_{t=1}^{\omega} c_t B_{t,i} \sum_{k=1}^{4(D_f+1)} I\big[M^T b_t + \Gamma > 0\big]_j \nu_{k,d}$$

$$\leq 2||B||_{1,\infty} \sum_{j=1}^{4(D_f+1)} \sum_{i=1}^{d+1} |\eta_{ij}| \sum_{t=1}^{\omega} |c_t| \sum_{k=1}^{4(D_f+1)} |\nu_{k,d}|$$

$$\lessapprox 8(D_f+1)\sqrt{d\omega}$$

L.7 Final Result

Putting all 10 terms together, we have

$$||\nabla^2_\Theta \mathcal{T}(X, \Theta)|| \leq \underbrace{2\sqrt{2(d+1)}}_{Term_1} + \underbrace{16 D_f \sqrt{2(D_f+1)\omega(d+1)}}_{Term_2}$$

$$+ \underbrace{16 D_f \sqrt{(D_f+1)\omega(d+1)}}_{Term_3} + \underbrace{4\sqrt{\omega(D_f+1)}}_{Term_4} + \underbrace{2\sqrt{\omega}}_{Term_5} + \underbrace{8(D_f+1)\sqrt{d\omega}}_{Term_6}$$

$$\approx \sqrt{d}(2\sqrt{2} + 4\sqrt{\omega}) + 32\sqrt{2} D_f^{\frac{3}{2}} \sqrt{\omega d} + 2\sqrt{\omega} + 8 D_f \sqrt{\omega d}$$

$$\approx 32\sqrt{2} D_f^{\frac{3}{2}} \sqrt{\omega d}$$

Finally, if we follow the more parameter and space efficient construction using random projections such that $d = O(log(T))$, we arrive at the final asymptotic result:

$$||\nabla^2_\Theta \mathcal{T}(X, \Theta)|| \in O\Big(D_f^{\frac{3}{2}} \sqrt{\omega log(T)}\Big)$$

□

M Perturbed Gradient Bounds

We now consider the gradient of a slightly perturbed network, since the bound on the average direction sharpness is only exact when the trace of the loss hessian is evaluated at some neighboring point to our actual learned minimum Θ, $\Theta + \zeta$. For convenience, denote the perturbed matrices by $\tilde{Q} = Q + \zeta_Q, \tilde{K} = K + \zeta_K, \tilde{V} = V + \zeta_V, \tilde{M} = M + \zeta_M, \tilde{F} = F + \zeta_F, \tilde{\Gamma} = \Gamma + \zeta_\Gamma$. We use nested indexing for the perturbation matrices, denoting the i, j^{th} element of the K component of the perturbations, we write $\zeta_{K_{i,j}}$. In this section we refer to these perturbations that make the Taylor's Theorem-based sharpness equation exact as the "remainder perturbations". We also reference the "original perturbations" or "original Gaussian perturbations." We use ϵ to denote these Gaussian perturbations, and index them the same way as the remainder perturbations, i.e. the i, j^{th} Gaussian perturbation in the PAC-Bayes bound is referred to as $\varepsilon_{K_{i,j}}$. By Taylor's Theorem, we can assume that the remainder perturbations are dominated in absolute value by their corresponding Gaussian perturbations.

M.1 Some Useful Lemmas

Before proving our main result, we state some useful lemmas:

Lemma 18. *Let χ be a chi-distributed variable with standard deviation σ and d degrees of freedom. Define*

$$R(\delta, d) := \sqrt{1 + 2\sqrt{\frac{log(\frac{1}{\delta})}{d}} + \frac{2log(\frac{1}{\delta})}{d}}$$

Then with probability at least $1 - \delta$, the following bound holds:

$$\chi \leq \sigma\sqrt{d}R(\delta, d) \approx \sigma\sqrt{d}$$

Proof. It is generally known that for a chi-squared distributed variable Z_d with d degrees of freedom, the following concentration bound holds:

$$P\Big(Z_d > T + 2\sqrt{dz} + 2z\Big) \leq e^{-z}$$

Noting that $\sqrt{Z_d}$ is Chi-distributed, this immediately yields a bound for any chi-distributed RV χ_d with d degrees of freedom.

$$P\Big(\chi_d > \sqrt{d + 2\sqrt{dz} + 2z}\Big) \leq e^{-z}$$

Therefore, with probability at least $1 - \delta$, the following upper bound holds:

$$||\epsilon_{V_{d+1,:}}|| \leq \sigma\sqrt{d + 2\sqrt{dlog(\frac{1}{\delta})} + 2log(\frac{1}{\delta})}$$

We now introduce a new function for notational simplicity, which increases polylogarithmically with δ and decreases polynomially with d :

$$R(\delta, d) := \sqrt{1 + 2\sqrt{\frac{log(\frac{1}{\delta})}{d}} + \frac{2log(\frac{1}{\delta})}{d}}$$

Thus we can restate our bound on$||\epsilon_{V_{d+1,:}}||$ as

$$||\epsilon_{V_{d+1,:}}|| \leq \sigma\sqrt{d}R(\delta, d) \approx \sigma\sqrt{d}$$

□

The above term approaches 1 as the number of degrees of freedom approaches infinity, which will allow us to simplify many of the expressions involving this function in the sequel. While this is clearly true when the scaling factor of δ from the union bound is constant, what happens if we are applying a union bound over T instances of R.V.s, such that the degrees of freedom is $d = O(log(T))$? In

this case we need to take care, since this polylogarithmic dependency on T has created a polynomial dependency on d, which may cause the terms involving the $log()$ function to not converge to 0 in the limit as $d \to \infty$. However, the following Lemma establishes that the variation above 1 in this term is controlled by our JLL approximation error, ϵ_p.

Lemma 19. *Suppose $\{||\epsilon_i||\}_{i\in[T]}$ is a set of T Chi-distributed variables with standard deviation σ and d degrees of freedom each, and that $d > \frac{8log(T)}{\epsilon_p^2} \iff T < e^{\frac{d\epsilon_p^2}{8}}$ By 18 and the union bound, with probability $1-\delta$, $||\epsilon_i|| \leq \sigma\sqrt{d}R(\frac{\delta}{T})$. Thus, $\lim_{T\to\infty} R(\frac{\delta}{T}, d) = 1 + O(\epsilon_p)$, and*

$$\forall i, ||\epsilon_i|| \lessapprox \sigma\sqrt{d}$$

Proof. Note that the definition of d is consistent with our dimension-reduced construction using JLL,

$$R(\frac{\delta}{T}, d) = \sqrt{1 + \frac{2}{\sqrt{d}}\sqrt{log(\frac{T}{\delta})} + \frac{2}{d}log(\frac{T}{\delta})}$$

$$\leq \sqrt{1 + 2\epsilon_p\sqrt{\frac{log(\frac{T}{\delta})}{8log(T)}} + 2\epsilon_p^2\frac{log(\frac{T}{\delta})}{8log(T)}} \leq \sqrt{1 + \frac{\epsilon_p}{\sqrt{2}}\sqrt{1 - \frac{log(\delta)}{log(T)}} + \frac{\epsilon_p^2}{4}\Big(1 - \frac{log(\delta)}{log(T)}\Big)}$$

$$\lim_{T\to\infty}\sqrt{1 + \frac{\epsilon_p}{\sqrt{2}}\sqrt{1 - \frac{log(\delta)}{log(T)}} + \frac{\epsilon_p^2}{4}\Big(1 - \frac{log(\delta)}{log(T)}\Big)}$$

$$= \sqrt{1 + \frac{\epsilon_p}{\sqrt{2}} + \frac{\epsilon_p^2}{4}} = 1 + O(\epsilon_p)$$

We note that polynomial increases in the number of variables we apply the union bound to do not change the asymptotics, and in the limit we are sure to have $R(\frac{\delta}{poly(T)}, d) \to 1 + O(\epsilon_p)$. □

Lemma 20. *Let $b_t \in \mathbb{R}^{d+1}$ be the activation right after the first attention layer, at position t. Then*

$$||\tilde{\phi}_t - \phi_t||_1 \lessapprox 3\sigma d$$

$$||\tilde{b}_t - b_t|| \lessapprox 4\sigma d$$

Proof.

$$||\tilde{b}_t - b_t|| = \Big\|(\tilde{V}^{(1)})^T X^T \phi\big(X(\tilde{W}^{(1)})^T x_t\big) - (V^{(1)})^T X^T \phi\big(X(W^{(1)})^T x_t\big)\Big\|$$

$$\leq \Big\|(\tilde{V}^{(1)})^T X^T\Big(\phi\big(X(\tilde{W}^{(1)})^T x_t\big) - \phi\big(X(W^{(1)})^T x_t\big)\Big)\Big\| + \Big\|(\tilde{V}^{(1)} - V^{(1)})^T X^T \phi\big(X(W^{(1)})^T x_t\big)\Big\|$$

$$\leq ||(\tilde{V}^{(1)})^T X^T||_{1,2}\Big\|\phi\big(X(\tilde{W}^{(1)})^T x_t\big) - \phi\big(X(W^{(1)})^T x_t\big)\Big\|_1 + \Big\|\zeta_{V^{(1)}}^T X^T\Big\|_{1,2}$$

$$\leq ||X\tilde{V}^{(1)}||_{2,\infty}\Big\|\phi\big(X(\tilde{W}^{(1)})^T x_t\big) - \phi\big(X(W^{(1)})^T x_t\big)\Big\|_1 + \Big\|X\zeta_{V^{(1)}}\Big\|_{2,\infty}$$

Where we used the triangle inequality, and the fact that for any matrix M and vector v, $||Mv||_2 \leq ||M||_{1,2}||v||_1$, and the fact that the 1-norm of the softmax is always exactly 1. We can follow a similar argument to Lemma A.6 in [12] and use the the fact that Jacobian of the softmax has a $(1,1)$-norm that is bounded by 2 to conclude (continued from above):

$$\leq 2||X\tilde{V}^{(1)}||_{2,\infty}\Big\|X\big((\tilde{W}^{(1)})^T - (W^{(1)})^T\big)x_t\Big\|_\infty + \Big\|X\zeta_V\Big\|_{2,\infty}$$

Let us take first the simpler term on the right, $\Big\|X\zeta_{V^{(1)}}\Big\|_{2,\infty}$. Recall that our dimension reduced-input matrix is $X = YJ$, where J is a $(T + 2 \times d + 1)$ projection matrix. Then the j^{th} element of the t^{th} row of $X\zeta_V$, is given by

$$\big[[X\zeta_{V^{(1)}}]_t\big]_j = \sum_{i=1}^{d} J_{t,i}\zeta_{V_{i,j}^{(1)}} + z_t\zeta_{V_{d+1,j}^{(1)}}$$

By the triangle inequality, we can thus bound the 2-norm of $[X\zeta_V]_t$ as

$$||[X\zeta_{V^{(1)}}]_t|| \leq \sqrt{\sum_{j=1}^{d+1}\Big(\sum_{i=1}^{d} J_{t,i}\zeta_{V^{(1)}_{i,j}}\Big)^2} + \sqrt{\sum_{j=1}^{d+1}(z_t\zeta_{V^{(1)}_{d+1,j}})^2}$$

$$\leq \left\| \begin{matrix} |J^T_{t,:}\zeta_{V^{(1)}_{:d,1}}| \\ \cdots \\ |J^T_{t,:}\zeta_{V^{(1)}_{:d,d+1}}| \end{matrix} \right\| + ||\zeta_{V^{(1)}_{d+1,:}}|||z_t|$$

We can now use Cauchy-Shwarz to write

$$\left\| \begin{matrix} |J^T_{t,:}\zeta_{V^{(1)}_{:d,1}}| \\ \cdots \\ |J^T_{t,:}\zeta_{V^{(1)}_{:d,d+1}}| \end{matrix} \right\| + ||\zeta_{V^{(1)}_{d+1,:}}|||z_t| \leq ||J_{t,:}|| \left\| \begin{matrix} ||\zeta_{V^{(1)}_{:d,1}}|| \\ \cdots \\ ||\zeta_{V^{(1)}_{:d,d+1}}|| \end{matrix} \right\| + ||\zeta_{V^{(1)}_{d+1,:}}||$$

Note that $||J_{t,:}||$ is a chi-distributed R.V. with standard deviation $\frac{1}{\sqrt{d}}$ and d degrees of freedom, while $||\epsilon_{V^{(1)}_{d+1,:}}||$ is a chi-squared R.V. with s.d. σ and $d+1$ degrees of freedom, and $\left\| \begin{matrix} ||\epsilon_{V^{(1)}_{:d,1}}|| \\ \cdots \\ ||\epsilon_{V^{(1)}:d,d+1}|| \end{matrix} \right\|$ is chi-distributed with s.d. σ and $d(d+1)$ degrees of freedom. Using 18 and then applying a union bound over all $T+2$ random variables potentially involved in this expression over all T positions, we can express our bound as

$$||X\zeta_{V^{(1)}}||_{2,\infty} = \max_t ||X_{t,:}\zeta_{V^{(1)}}|| \leq \max_t ||J_{t,:}|| \left\| \begin{matrix} ||\zeta_{V^{(1)}_{:d,1}}|| \\ \cdots \\ ||\zeta_{V^{(1)}_{:d,d+1}}|| \end{matrix} \right\| + ||\zeta_{V^{(1)}_{d+1,:}}||$$

$$(w.p. > 1-\delta, \forall t) \leq \sigma\sqrt{d(d+1)}R(\frac{\delta}{T+2}, d)R(\frac{\delta}{T+2}, d(d+1)) + \sigma\sqrt{d+1}R(\frac{\delta}{T+2}, d+1)$$
$$\approx \sigma d$$

Where we have used 19. Now we consider another term in our bound for $|\tilde{b_t} - b_t|$ above, $\left\|X\big((\tilde{W}^{(1)})^T - (W^{(1)})^T\big)x_t\right\|_\infty = \left\|X\zeta_{W^{(1)}}x_t\right\|_\infty$. To bound this term, we note that

$$\left\|X\zeta_{W^{(1)}}x_t\right\|_\infty \leq ||X\zeta_{W^{(1)}}||_{2,\infty}||x_t||,$$

and then bound these terms individually. Note that

$$[x_s^T\zeta^T_{W^{(1)}}]_l = \sum_{i=1}^{d} J_{s,i}\zeta_{W^{(1)}_{i,l}} + z_s\zeta_{W^{(1)}_{d+1,l}}$$

From the triangle inequality, we can factor the above expression and replace the remainder perturbations ζ with the original gaussian perturbations ϵ that dominate them in absolute value:

$$\left\|X\big((\tilde{W}^{(1)})^T - (W^{(1)})^T\big)\right\|_{2,\infty} \leq \max_s ||J_{s,:}|| \left\| \begin{matrix} ||\epsilon_{W^{(1)}_{:d,1}}|| \\ \cdots \\ ||\zeta_{W^{(1)}_{:d,d+1}}|| \end{matrix} \right\| + \left\|\zeta_{W^{(1)}_{d+1,:}}\right\||z_s|$$

note this is the exact same distribution of R.V.s that we saw when bounding $||XV^{(1)}||_{2,\infty}$ above. We can again apply the union bound over all $T+2$ R.V.s in question and apply 19 to conclude that

$$||X\zeta^T_W||_{2,\infty} \lessapprox \sigma d$$

From 10 , we know that $||x_t|| \leq ||X||_{2,\infty} \lessapprox \sqrt{2}$, which holds simultaneously $\forall t$. Therefore by applying Cauchy-Schwarz, we arrive at the following bound which holds w.p. $1-\delta$:

$$\forall t : ||X\zeta^T_{W^{(1)}}x_t||_\infty \lessapprox \sqrt{2}\sigma d$$

It follows that

$$||\tilde{\phi}_t - \phi_t||_1 \leq 2||X\zeta^T_{W^{(1)}}x_t||_\infty \lessapprox 2\sqrt{2}\sigma d \leq 3\sigma d.$$

Finally, we consider the term $||X\tilde{V}^{(1)}||_{2,\infty}$. Recall that $x_t = J^T y_t$, and so the l^{th} element of x_t is $[x_t]_l = (J^T)_{l,t}$ for $l \in [d]$, $[x_t]_{d+1} = z_t$. In addition, recall that the matrix $V^{(1)} = J^T\bar{V}^{(1)} \in \mathbb{R}^{(d+1)\times(T+2)}$ is a matrix of all 0s except in the final row and column, which has a 1.

Thus the $(t,i)^{th}$ element of $X\tilde{V}^{(1)}$ is given by:

$$[x_t^T\tilde{V}^{(1)}]_i = z_t\tilde{V}^{(1)}_{d+1,i} + \sum_{l=1}^{d} J_{t,l}\tilde{V}^{(1)}_{l,i} = z_t\zeta_{\tilde{V}^{(1)}_{d+1,i}} + I[i=d+1]\big) + \sum_{l=1}^{d} J_{t,l}\zeta_{V^{(1)}_{l,i}}$$

By the triangle inequality, the 2-norm of $x_t^T\tilde{V}^{(1)}$ is thus bounded by

$$||x_t^T\tilde{V}^{(1)}|| \leq |z_t|(||\zeta_{\tilde{V}^{(1)}_{d+1,:d}}|| + ||e_{d+1}||) + \left|\left| \begin{matrix} |J^T_{t,:}\zeta_{V^{(1)}_{:d,1}}| \\ \cdots \\ |J^T_{t,:}\zeta_{V^{(1)}_{:d,d+1}}| \end{matrix} \right|\right|$$

We can once again use Cauchy-Shwarz and apply our union bound over T+2 chi-distributed variables involved in the above expression for all t to conclude that

$$||x_t^T\tilde{V}^{(1)}|| \leq 1 + \sigma\sqrt{d}R(\frac{\delta}{T+2}, d) + \sigma\sqrt{(d(d+1)}R(\frac{\delta}{T+2}, d(d+1))R(\frac{\delta}{T+2}, d) \approx 1 + \sigma d$$

Putting this all together, we have that with probability at least $1 - 3\delta$,

$$\begin{aligned} ||\tilde{b}_t - b_t|| &\leq 2||X\tilde{V}^{(1)}||_{2,\infty}\Big|\Big|X\big((\tilde{W}^{(1)})^T - (W^{(1)})^T\big)x_t\Big|\Big|_\infty + \Big|\Big|X\zeta_V\Big|\Big|_{2,\infty} \\ &\leq 2\Big(1+\sigma d\Big)\Big(\sqrt{2}\sigma d\Big) + \sigma d \approx (2\sqrt{2}+1)\sigma d + 2\sqrt{2}\sigma^2 d^2 \end{aligned}$$

To simplify our analysis and the notation, we make the following simplification: we ignore the terms involving orders of σ of 2 and above. The reason for this is that the *additional* dependencies on ω, D_f, d that appear in terms involving e.g. σ^2 are no larger than the largest dependencies for terms involving σ. Therefore, any choice of σ that is small enough to make the first-order terms small will be small enough to make the higher order terms even smaller. While this argument is informal (we do not prove it explicitly for each and every bound), it can be easily verified, and dramatically simplifies the derivations in the sequel. One consequence of this is that, for any bounds involving terms that are the product of a perturbation term with another term that is a perturbed parameter matrix (i.e. $||\tilde{V}^T X^T||_2\Big|\Big|X\big(\tilde{W}^T - (W^{(1)})^T\big)x_t\Big|\Big|_1$ above), we can safely assume that the perturbation component will lead to an $O(\sigma^2)$ term that can be ignored. Therefore, we could have made the approximation that

$$||\tilde{V}^T X^T||_2\Big|\Big|X\big(\tilde{W}^T - (W^{(1)})^T\big)x_t\Big|\Big|_1 \approx ||(V^{(1)})^T X^T||_2\Big|\Big|X\big(\tilde{W}^T - (W^{(1)})^T\big)x_t\Big|\Big|_1,$$

and all of the perturbation terms that are first-order in σ would have been the same. We will employ such approximations in the sequel. With this in mind, we use the first-order approximation, and round our constant up to the nearest whole number:

$$||\tilde{b}_t - b_t|| \leq 4\sigma d$$

□

Lemma 21. *Let $g_t \in \mathbb{R}^{d+1}$ be the activation right after the MLP sublayer, including the residual connection, at position t. Furthermore, assume that the standard deviation of the perturbations is such that*

$$\sigma \leq \frac{1}{16dD_f}$$

Then

$$\begin{aligned} ||\tilde{g}_t - g_t|| &\lessapprox 32\sigma D_f^2 \\ &\in O(\sigma D_f^2) \end{aligned}$$

Proof. We will use the triangle inequality many times in the following arguments, starting with the observation that

$$\begin{aligned}
&\left\|g\Big((V^{(1)})^T X^T \phi\big(X(W^{(1)})^T x_t\big)\Big) - \tilde{g}\Big(\tilde{V}^T X^T \phi\big(X\tilde{W}^T x_t\big)\Big)\right\| \\
&\leq \left\|g\Big(\tilde{V}^T X^T \phi\big(X\tilde{W}^T x_t\big)\Big) - g\Big((V^{(1)})^T X^T \phi\big(X(W^{(1)})^T x_t\big)\Big)\right\| \\
&\quad + \left\|\tilde{g}\Big(\tilde{V}^T X^T \phi\big(X\tilde{W}^T x_t\big)\Big) - g\Big(\tilde{V}^T X^T \phi\big(X\tilde{W}^T x_t\big)\Big)\right\|
\end{aligned}$$

To illustrate the two main components involved in our bound, we simplify notation and write

$$\left\|g(b_t) - \tilde{g}(\tilde{b}_t)\right\| \leq \underbrace{\left\|g(\tilde{b_t}) - g(b_t)\right\|}_{\text{Attn. Perturbation}} + \underbrace{\left\|\tilde{g}(\tilde{b}_t) - g(\tilde{b_t})\right\|}_{\text{MLP Perturbation}}$$

Note that in the attention perturbation term, both terms inside the norm involve the unperturbed MLP function g, which has a slope of 0 near b_t. This term will be zero as long as none of the MLP activations changes. In reality, with some small probability, one or more of the neurons in the perturbed transformer may flip from being active to inactive. We opt to bound the probability of any such scenario occurring, effectively eschewing the need for a detailed analysis of it. Because of our choice of an MLP construction that has some width around the points in the domain for which the value is constant, this probability is low as long as σ is small (unlike some other constructions of indicators from ReLUs using e.g. triangular waves). For any neuron to change its activation, it must be the case that $|\tilde{M}_{:,i}^T \tilde{b}_t + \tilde{\Gamma}_i - (M_{:,i}^T b_t + \Gamma_i)| \geq \frac{1}{4D_F}$. In other words, for such an event to occur, the total perturbation in $M_{:,i}^T b_t + \Gamma_i$ must be at least as large as the resolution of our MLP. Now,

$$\begin{aligned}
|\tilde{M}_{:,i}^T \tilde{b}_t + \tilde{\Gamma}_i - (M_{:,i}^T b_t + \Gamma_i)| &\lessapprox |M_{:,i}^T(\tilde{b}_t - b_t)| + |(\tilde{M}_{:,i}^T - M_{:,i}^T) b_t + (\tilde{\Gamma}_i - \Gamma_i)| \\
&\leq ||\tilde{b}_t - b_t|| + ||\zeta_{M_{:,i}}|| + |\zeta_{\Gamma_i}|
\end{aligned}$$

In 20 we bounded $||\tilde{b}_t - b_t|| \lessapprox 4\sigma d$. In addition, we can bound each of the $||\zeta_{M_{:,i}}||, |\zeta_{\Gamma_i}|$ as chi-distributed variables with $d+1$ and 1 degree of freedom, respectively. Taking a union bound over all $4(D_f+1)$ instances of these variables, we have With probability $1-\delta$,

$$\forall i \in [4(D_f+1)] : |\epsilon_{\Gamma_i}| \leq \sigma R(\frac{\delta}{8(D_f+1)}, 1)$$

$$\forall i \in [4(D_f+1)] : ||\zeta_{M_{:,i}}|| \leq \sigma R(\frac{\delta}{8(D_f+1)}, d+1) \approx \sigma\sqrt{d}$$

Thus, the following bound holds for all $i \in [4(D_f+1)]$:

$$\begin{aligned}
&|\tilde{M}_{:,i}^T \tilde{b}_t + \tilde{\Gamma}_i - (M_{:,i}^T b_t + \Gamma_i)| \\
&\lessapprox 4\sigma d + \sigma R(\frac{\delta}{8(D_f+1)}, 1) + \sigma\sqrt{d} \\
&\approx 4\sigma d
\end{aligned}$$

In order for the RHS to be less than $\frac{1}{4D_f}$, we must have

$$\sigma \leq \frac{1}{16 d D_f}$$

As long as σ is small like this, then with high probability, none of the activations in our MLP will flip, and we conclude that

$$\left\|g(\tilde{b_t}) - g(b_t)\right\| = 0$$

Now, the second major term we need to bound is the perturbation in the MLP: $\left\|\tilde{g}(\tilde{b}_t) - g(\tilde{b}_t)\right\|$. Expanding out the MLP function in terms of the parameter matrices, and applying the triangle inequality, we obtain two main terms:

$$\left\|\tilde{g}(\tilde{b}_t) - g(\tilde{b}_t)\right\| = \left\|\tilde{F}^T(\tilde{M}^T \tilde{b}_t + \tilde{\Gamma})_+ - F^T(M^T \tilde{b}_t + \Gamma)_+\right\|$$

$$\leq \left\|\tilde{F}^T(\tilde{M}^T\tilde{b}_t + \tilde{\Gamma})_+ - \tilde{F}^T(M^T\tilde{b}_t + \Gamma)_+\right\| + \left\|\tilde{F}^T(M^T\tilde{b}_t + \Gamma)_+ - F^T(M^T\tilde{b}_t + \Gamma)_+\right\|$$

$$\lessapprox \underbrace{\left\|F^T(\tilde{M}^T\tilde{b}_t + \tilde{\Gamma})_+ - F^T(M^T\tilde{b}_t + \Gamma)_+\right\|}_{Term_1} + \underbrace{\left\|\tilde{F}^T(M^Tb_t + \Gamma)_+ - F^T(M^Tb_t + \Gamma)_+\right\|}_{Term_2}$$

Where we have replaced $\tilde{F}$ in the first term above with just F, since this term is already considering perturbations to M and Γ, and therefore the perturbed versions of F and b_t contribute only a second order term to the overall perturbation. Recalling that the first d columns of F are 0, we then have

$$Term_1 \leq \left|(F_{:,d+1})^T(\tilde{M}^T\tilde{b}_t + \tilde{\Gamma})_+ - (F_{:,d+1})^T(M^T\tilde{b}_t + \Gamma)_+\right|$$

$$\leq \left\|F_{:,d+1}\right\|\left\|(\tilde{M}^T\tilde{b}_t + \tilde{\Gamma})_+ - (M^T\tilde{b}_t + \Gamma)_+\right\|$$

$$\leq \left\|F_{:,d+1}\right\|\left\|(\tilde{M}^T - M^T)\tilde{b}_t + (\tilde{\Gamma} - \Gamma)\right\|$$

where we have used the fact that the $()_+$ function is 1-lipschitz. Recall from our definitions that $M \in \mathbb{R}^{(d+1)\times 4(D_f+1)}$, while $\Gamma \in \mathbb{R}^{4(D_f+1)}$, and $F \in \mathbb{R}^{4(D_f+1)\times(d+1)}$. Note that

$$\left\|F_{:,d+1}\right\| = \sqrt{\sum_{i=1}^{4(D_f+1)} (\pm 4D_f)^2} \leq 4D_f\sqrt{4(D_f+1)} \approx 8D_f^{\frac{3}{2}}$$

Now, note that the i^{th} element of $(\tilde{M}^T - M^T)\tilde{b}_t + (\tilde{\Gamma} - \Gamma)$ is given by $\zeta_{\Gamma_i} + \zeta_{M_{i,:}}\tilde{b}_t$. Therefore, by the triangle inequality

$$||(\tilde{M}^T - M^T)b_t + (\tilde{\Gamma} - \Gamma)||_2 \leq ||\zeta_\Gamma|| + ||\zeta_{M_{:,d+1}}||||B_{t,d+1}| \leq ||\zeta_\Gamma|| + ||\zeta_{M_{:,d+1}}||$$

Note that we have replaced $\tilde{b_t}$ with b_t due to the aforementioned argument, that the difference is only in the terms involving $O(\sigma^2)$, which we assumed can be safely ignored for our purposes. We also use the fact that only the $d+1$ dimension is nonzero for b_t. $||\epsilon_\Gamma||$ and $||\epsilon_{M_{:,d+1}}||$ are each chi-distributed R.V.s with $4(D_f+1)$ degrees of freedom. Thus, applying a union bound, the following two bounds hold simultaneously with probability at least $1-\delta$:

$$||\epsilon_\Gamma|| \leq \sigma\sqrt{4(D_f+1)}R(\frac{\delta}{2}, 4(D_f+1))$$

$$||\epsilon_M|| \leq \sigma\sqrt{4(D_f+1)}R(\frac{\delta}{2}, 4(D_f+1))$$

By the triangle inequality, with probability $1-\delta$,

$$||(\tilde{M}^T - M^T)b_t + (\tilde{\Gamma} - \Gamma)||_2 \leq 2\sigma\sqrt{4(D_f+1)}R(\frac{\delta}{2}, 4(D_f+1))$$

$$\lessapprox 4\sigma\sqrt{D_f}$$

We now consider

$$Term_2 = \left\|\zeta_F^T(M^T\tilde{b}_t + \Gamma)_+\right\| \leq \left\|\zeta_F\right\|_{2,\infty}\left\|M^T\tilde{b}_t + \Gamma\right\|_1$$

To bound $\left\|\zeta_F\right\|_{2,\infty}$, we use the same arguments as before for bounding chi-distributed variables with $4(D_f+1)$ degrees of freedom, using a union bound over all $d+1$ rows of ζ_F. With probability $1-\delta$:

$$||\zeta_F||_2 \leq \sigma\sqrt{4(D_f+1)}R(\frac{\delta}{d+1}, 4(D_f+1)) \approx 2\sigma\sqrt{D_f}$$

To bound $\left\|M^Tb_t + \Gamma\right\|_2$, we have

$$\left\|M^Tb_t + \Gamma\right\|_2 \approx \left\|M^Tb_t + \Gamma\right\|_2 \leq ||M|||b_t| + ||\Gamma||$$

$$\leq 2\sqrt{4(D_f+1)} \approx 4\sqrt{D_f}.$$

We now summarize the bounds on all of the ingredients to our MLP perturbation, as well as the final bound on $\underbrace{\left\|\tilde{g}(\tilde{b}_t) - g(\tilde{b_t})\right\|}_{\text{MLP Perturbation}}$.

$$\begin{aligned}
\left\|\tilde{g}(\tilde{b}_t) - g(\tilde{b_t})\right\| &\leq Term_1 + Term_2 \\
\left\|F_{:,d+1}\right\|\left\|(\tilde{M}^T - M^T)\tilde{b}_t + (\tilde{\Gamma} - \Gamma)\right\| & \\
+\left\|\zeta_F\right\|_{2,\infty}\left\|M^T\tilde{b}_t + \Gamma\right\|_1 & \\
+\left(8D_f^{\frac{3}{2}}\right)\left(4\sigma\sqrt{D_f}\right) + \left(2\sigma\sqrt{D_f}\right)\left(4\sqrt{D_f}\right) & \\
= 8\sigma D_f(4D_f + 1) \approx 32\sigma D_f^2 &
\end{aligned}$$

Finally, we combine the two results to get the perturbation to the norm of the MLP outputs: the perturbation of the MLP applied to the perturbed attention output, and the perturbation of the unperturbed MLP applied to the perturbation in the attention output:

$$\begin{aligned}
\left\|g(b_t) - \tilde{g}(\tilde{b}_t)\right\| &\leq \underbrace{\left\|g(\tilde{b_t}) - g(b_t)\right\|}_{\text{Attn. Perturbation}} + \underbrace{\left\|\tilde{g}(\tilde{b}_t) - g(\tilde{b_t})\right\|}_{\text{MLP Perturbation}} \\
&\lessapprox 32\sigma D_f^2 \\
&\in O(\sigma D_f^2)
\end{aligned}$$

□

Theorem 22. *Given a set of remainder perturbations ζ to the parameters Θ such that $|\zeta_i| \leq |\epsilon_i|$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)^n$, the error of the perturbed transformer satisfies:*

$$|f(x) - \mathcal{T}(X, \Theta + \zeta)| \leq T_p(\sigma, \omega, D_f, T)$$

where

$$T_p(\sigma, \omega, D_f, T) \lessapprox 128\sigma D_f^2 \omega log(T) \in O(\sigma D_f^2 \omega log(T))$$

Proof. We bound the magnitude of the perturbation of our transformer, $|\mathcal{T}(X, \Theta + \zeta)) - \mathcal{T}(X, \Theta))|$.

$|f(x) - \mathcal{T}(X, \Theta + \zeta)| \leq |f(x) - \mathcal{T}(X, \Theta)| + |\mathcal{T}(X, \Theta) - \mathcal{T}(X, \Theta + \zeta)| = |\mathcal{T}(X, \Theta) - \mathcal{T}(X, \Theta + \zeta)|.$

$$\begin{aligned}
\left|\mathcal{T}(X, \Theta + \zeta) - \mathcal{T}(X, \Theta)\right| &\leq \left|(\tilde{V}^{(2)})^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)} - (V^{(2)})^T G^T \phi_{T+1}^{(2)}\right| \\
\lessapprox \left|(\tilde{V}^{(2)} - V^{(2)})^T G^T \phi_{T+1}^{(2)}\right| &+ \left|(V^{(2)})^T (\tilde{G} - G)^T \phi_{T+1}^{(2)}\right| + \left|(V^{(2)})^T G^T (\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)})\right| \\
\leq \left\|G\zeta_{V^{(2)}}\right\|_\infty &+ \left\|(\tilde{G} - G)V^{(2)}\right\|_\infty + \left\|GV^{(2)}\right\|_\infty ||\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)}||_1
\end{aligned}$$

Where we have used Holder's inequality three times, and used the fact that $||\phi_{T+1}^{(2)}||_1 = 1$. Recall from 10 that $\left\|GV^{(2)}\right\|_\infty \lessapprox \sqrt{\omega}$. We showed in 25 that

$$||\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)}||_1 \lessapprox 128\sigma D_F^2 \sqrt{\omega} log(T)$$

Note that

$$\left\|G(\tilde{V}^{(2)} - V^{(2)})\right\|_\infty \leq ||G||_{2,\infty} ||\zeta_{V^{(2)}}|| \leq \sqrt{2}||\zeta_{V^{(2)}}||$$

$||\zeta_{V^{(2)}}||$ is a chi-distributed variable with $d + 1$ degrees of freedom, and this with probability $1 - \delta$,

$$||\zeta_{V^{(2)}}|| \leq \sigma\sqrt{d+1}R(\delta, d+1) \approx \sigma\sqrt{d}.$$

Therefore

$$\left\|G(\tilde{V}^{(2)} - V^{(2)})\right\|_\infty \lessapprox \sigma\sqrt{2d}$$

Now, we have

$$\left\|(\tilde{G}-G)V^{(2)}\right\|_{\infty} \leq \left\|(\tilde{G}-G)\right\|_{2,\infty}||V^{(2)}|| \leq \sqrt{\omega}\left\|(\tilde{G}-G)\right\|_{2,\infty}$$

Where we have used our bound on $D = \sum_{t=1}^{\omega}$ from 10. Combining this with our bound on $||\tilde{g}_t - g_t||_2$ in 21, this yields

$$\left\|(\tilde{G}-G)V^{(2)}\right\|_{\infty} \lessapprox 32\sqrt{\omega}\sigma D_f^2$$

Putting this all together, we have

$$\left|\mathcal{T}(X,\Theta+\zeta)-\mathcal{T}(X,\Theta)\right| \lessapprox \sigma\sqrt{2d}+32\sqrt{\omega}\sigma D_f^2+128\sigma D_f^2\omega log(T)$$

$$\approx 128\sigma D_f^2\omega log(T)$$

□

Lemma 23. *Let $g_t \in \mathbb{R}^{d+1}$ be the activation at position t after the MLP, and let $b_t \in \mathbb{R}^{d+1}$ be the activation at position t immediately after the attention sub-layer. Furthermore, assume that the standard deviation of the perturbations is such that*

$$\sigma \leq \frac{1}{16dD_f}$$

Then with probability at least $1-\delta$, the norm of the perturbation in the i^{th} column of the Jacobian of g_t with respect to b_t is given by

$$\left|\left|\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i} - \left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i}\right|\right| \lessapprox 16\sigma D_f^2$$

Proof. Our expression for g_t in terms of b_t is given by

$$g_t = F^T\Big(M^T b_t + \tilde{\Gamma}\Big)_+$$

Using basic rules of matrix calculus, the Jacobian is given by:

$$= F^T Diag\Big(I\Big[M^T b_t + \Gamma > 0\Big]\Big)M^T$$

Note that in our exact construction, only the last row of F^T and the last column of M^T are nonzero, meaning that only the bottom right entry of the Jacobian is nonzero (allowing us to make the drastic simplification in 15).

However, when we are dealing with perturbations, we must deal with the full Jacobian matrix. Consider the norm of the perturbation of the derivative of $g_t \in \mathbb{R}^{d+1}$ with respect to $B_{t,i}$:

$$\left|\left|\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i} - \left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i}\right|\right|$$

$$= \left\|\tilde{F}^T Diag\Big(I\Big[\tilde{M}^T\tilde{b}_t + \tilde{\Gamma} > 0\Big]\Big)\tilde{M}^T e_{d+1} - F^T Diag\Big(I\Big[M^T b_t + \Gamma > 0\Big]\Big)M^T e_{d+1}\right\|$$

Note that, if we can assume that none of the MLP neurons change from being positive to 0 or vice-versa under the perturbation (an assumption formally justified already in 21) , then we have $I\Big[\tilde{M}^T\tilde{b}_t + \tilde{\Gamma} > 0\Big] = I\Big[M^T\tilde{b}_t + \Gamma > 0\Big]$. From there, we can easily bound the above using the triangle inequality:

$$\lessapprox \left\|F^T Diag\Big(I\Big[M^T b_t + \Gamma > 0\Big]\Big)(M-M)^T e_{d+1}\right\| + \left\|(\tilde{F}-F)^T Diag\Big(I\Big[M^T b_t + \Gamma > 0\Big]\Big)M^T e_{d+1}\right\|$$

$$\leq \left\|F^T\right\|_{2,\infty}\left\|(M-M)^T e_{d+1}\right\| + \left\|(\tilde{F}-F)^T\right\|_{2,\infty}\left\|M^T e_{d+1}\right\|$$

The perturbation $|\tilde{g}'_t(\tilde{b}_t) - g'_t(b_t)|$ does not depend on t, and thus we do not require a union bound to make it hold $\forall t$. Recall that F is only nonzero in its final column, which has absolute values of $4D_f$. Thus $||F||_{2,\infty} \leq 8\sqrt{D_f+1}D_f \approx 8D_f^{\frac{3}{2}}$. Similarly, the only nonzero row of M is the last row, which is all $1s$, and therefore $||Me_{d+1}|| \leq 2\sqrt{D_f+1} \approx 2D_f^{\frac{1}{2}}$. $\left\|(M-M)^T e_{d+1}\right\|$ is a chi-distributed variable with $4(D_f+1)$ degrees of freedom, and therefore with probability $1-\delta$,

$$\left\|(M-M)^T e_{d+1}\right\| \leq \sigma\sqrt{4(D_f+1)}R(\delta, 4(D_f+1)) \approx 2\sigma\sqrt{D_f}$$

Finally, for $\left\|(\tilde{F}-F)^T\right\|_{2,\infty}$, we can take a union bound over $d+1$ chi-distributed variables with $4(D_f+1)$ degrees of freedom to obtain the following bound which holds with probability $1-\delta$

$$\left\|(\tilde{F}-F)^T\right\|_{2,\infty} \leq \sigma\sqrt{4(D_f+1)}R(\frac{\delta}{d+1}, 4(D_f+1)) \approx 2\sigma\sqrt{D_f}$$

Putting these pieces together, we have

$$\left\|\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i} - \left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i}\right\| \lessapprox 16\sigma D_f^2 + 4\sigma D_f \approx 16\sigma D_f^2$$

This bound only holds as long as none of the MLP activations changes, which as we show in 21, occurs with probability at least $1-\delta$ as long as

$$\sigma \leq \frac{1}{16dD_f}$$

As long as σ is small like this, then with high probability, none of the activations in our MLP will flip, and our simple bound on the norm of the perturbations of the columns of the Jacobian above holds. □

Lemma 24. *Let $g_t \in \mathbb{R}^{d+1}$ be the MLP activation at position t. Then we can bound*

$$||\nabla_{g_t}\mathcal{T}(X,\Theta)|| \leq |c_t| + 4\sqrt{\omega}log(T)I[t=T+1] + 4\omega log(T)I[t\in[\omega]]$$

Proof. The more general formula for $\nabla_{g_t}\mathcal{T}(X,\Theta)$ which considers all of the dimensions of $g_t \in \mathbb{R}^{d+1}$ is given in 15 by the following:

$$\nabla^T_{g_t}\mathcal{T}(X,\Theta) = e_t^T\phi^{(2)}_{T+1}(V^{(2)})^T + e_t^T e_{T+1}(V^{(2)})^T G^T \phi'^{(2)}_{T+1} G(W^{(2)})^T + e_t^T\phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}GW^{(2)}$$

By the triangle inequality,

$$\|\nabla^T_{g_t}\mathcal{T}(X,\Theta)\|$$
$$\leq |\frac{c_t}{D}|\left\|V^{(2)}\right\| + I\big[t=T+1\big]\left\|(V^{(2)})^T G^T\phi'^{(2)}_{T+1}G(W^{(2)})^T\right\| + \left\|e_t^T\phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}GW^{(2)}\right\|$$

Note that the only nonzero element of $V^{(2)}$ is the final dimension, which has value $D = \sum_{t=1}^{\omega} c_t \leq \sqrt{\omega}$. Therefore $|\frac{c_t}{D}|\left\|V^{(2)}\right\| = |c_t|\left\|e_{d+1}\right\| \leq |c_t|$. To bound the second term, we use 16, which tells us that

$$\left\|(V^{(2)})^T G^T\phi'^{(2)}_{T+1}G(W^{(2)})^T\right\| \leq 2\|GV^{(2)}\|_\infty \left\|G(W^{(2)})^T\right\|_{2,\infty}$$

We have already shown in 10 that $\|GV^{(2)}\|_\infty \leq \sqrt{\omega}$. It is shown in 10 that $\|G(W^{(2)})^T\|_{2,\infty} \lessapprox 2log(T)$. Thus

$$\left\|(V^{(2)})^T G^T\phi'^{(2)}_{T+1}G(W^{(2)})^T\right\| \leq 4\sqrt{\omega}log(T)$$

Now we bound

$$\left\|e_t^T\phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}GW^{(2)}\right\|$$
$$\left\|e_t^T\phi'^{(2)}_{T+1}\right\|\left\|GV^{(2)}\right\|\left\|e^T_{T+1}GW^{(2)}\right\|$$

We can use 16 to bound $\left\|e_t^T\phi'^{(2)}_{T+1}\right\| \leq 2I[t\in[\omega]]$. Note that we can condition on $t\in[\omega]$ because the unperturbed softmax Jacobian matrix $\phi'^{(2)}_{T+1}$ is a (symmetric) outer-product matrix with nonzero

entries only in the first ω rows and columns. In 10 we showed that $\left\|GV^{(2)}\right\| \lessapprox \sqrt{\omega}$, and that $\|(W^{(2)})^T g_{T+1}\| \lessapprox 2\sqrt{\omega} log(T)$.

Therefore

$$\left\|e_t^T \phi_{T+1}'^{(2)} GV^{(2)} e_{T+1}^T GW^{(2)}\right\| \leq 4\omega log(T) I[t \in [\omega]]$$

In summary, we have

$$\|\nabla_{g_t}^T \mathcal{T}(X,\Theta)\| \leq |c_t| + 4\sqrt{\omega} log(T) I[t = T+1] + 4\omega log(T) I[t \in [\omega]]$$

□

Lemma 25. *Let $\phi_{T+1}^{(2)} \in \mathbb{R}^{T+1}$ be the softmax score for position $T+1$ in the second attention layer of our transformer. Then*

$$\|\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)}\|_\infty \leq \|\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)}\|_1 \lessapprox 128\sigma D_f^2 \sqrt{\omega} log(T)$$

Proof. By construction,

$$\phi_{T+1}^{(2)} = \phi(G(W^{(2)})^T g_{T+1}.$$

We seek an upper bound on the 1-norm of the perturbed final activation, which also serves as an upper bound on the $\infty-$norm:

$$\|\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)}\|_\infty \leq \|\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)}\|_1$$

By Lemma A.6 in [12], the 2-boundedness of the (1,1)-norm of the softmax's Jacobian matrix implies that

$$\|\phi\big(\tilde{G}(\tilde{W}^{(2)})^T \tilde{g}_{T+1}\big) - \phi\big(G(W^{(2)})^T g_{T+1}\big)\|_1 \leq 2\|\tilde{G}(\tilde{W}^{(2)})^T \tilde{g}_{T+1} - G(W^{(2)})^T g_{T+1}\|_\infty$$

$$\lessapprox 2\|(\tilde{G}-G)(W^{(2)})^T g_{T+1}\|_\infty + 2\|G(\tilde{W}^{(2)} - W^{(2)})^T g_{T+1}\|_\infty + 2||G(W^{(2)})^T(\tilde{g}_{T+1} - g_{T+1})||_\infty$$

$$\leq 2||\tilde{G}-G||_{2,\infty}||(W^{(2)})^T g_{T+1}|| + 2||G(\tilde{W}^{(2)} - W^{(2)})^T||_{2,\infty}||g_{T+1}|| + 2||G(W^{(2)})^T||_{2,\infty}||(\tilde{g}_{T+1} - g_{T+1})||$$

Note that $||\tilde{G} - G||_{2,\infty}$ was already bounded in 21. In 10, we show that $||(W^{(2)})^T g_{T+1}|| \leq 2\sqrt{\omega} log(T)$. To bound the term $||G(\tilde{W}^{(2)} - W^{(2)})^T||_{2,\infty}$, we follow similar arguments to the proof of 20 to arrive at the bound

$$\left\|G\big((\tilde{W}^{(2)})^T - (W^{(2)})^T\big)\right\|_{2,\infty} \leq \max_s ||J_{s,:}|| \left\| \begin{matrix} ||\epsilon_{W_{:d,1}^{(2)}}|| \\ \cdots \\ ||\zeta_{W_{:d,d+1}^{(2)}}|| \end{matrix} \right\| + \left\|\zeta_{W_{d+1,:}^{(2)}}\right\| |g_s|$$

Continuing to follow that proof, we can again apply the union bound over all $T+2$ R.V.s in question and apply 19 to conclude that

$$||G\zeta_W^T||_{2,\infty} \lessapprox \sigma d$$

We have already bounded $||g_{T+1}||$ in 10 as $\lessapprox 1$. Clearly, $||(\tilde{g}_{T+1} - g_{T+1})|| \leq ||\tilde{G} - G||_{2,\infty}$. Finally, we show in 10 that $||G(W^{(2)})^T||_{2,\infty} \lessapprox 2log(T)$.

Putting all of these bounds together, we have

$$||\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)}||_\infty \leq ||\tilde{\phi}_{T+1}^{(2)} - \phi_{T+1}^{(2)}||_1$$

$$\lessapprox 128\sigma D_f^2 \sqrt{\omega} log(T) + 2\sigma d + 128\sigma D_f^2 log(T)$$

$$\approx 128\sigma D_f^2 \sqrt{\omega} log(T)$$

□

Lemma 26. *Let $g_t, \tilde{g}_t \in \mathbb{R}^{d+1}$ be the unperturbed and perturbed (respectively) post-mlp activations of our transformer. Then the 2-norm of the difference of the gradients is given by*

$$|\langle \nabla_{\tilde{g}_t} \mathcal{T}(X,\tilde{\Theta}) - \nabla_{g_t} \mathcal{T}(X,\Theta), e_i \rangle|$$

$$\lessapprox 128\sigma D_f^2 \omega log(T) I[i = d+1] + 768\sigma D_f^2 \omega^{\frac{3}{2}} log^2(T) I[i < d+1]$$

Proof. Recall that the more general formula for $\nabla_{\tilde{g}_t}\mathcal{T}(X,\tilde{\Theta})$ which considers all of the coordinates of $\tilde{g}_t$ is given in 15 by the following:

$$\nabla^T_{\tilde{g}_t}\mathcal{T}(X,\tilde{\Theta}) = \underbrace{e_t^T\tilde{\phi}^{(2)}_{T+1}(\tilde{V}^{(2)})^T}_{Term_I} + \underbrace{e_t^T\tilde{\phi}'^{(2)}_{T+1}\tilde{G}\tilde{V}^{(2)}e^T_{T+1}\tilde{G}\tilde{W}^{(2)}}_{Term_{II}} + \underbrace{e_t^Te_{T+1}(\tilde{V}^{(2)})^T\tilde{G}^T\tilde{\phi}'^{(2)}_{T+1}\tilde{G}(\tilde{W}^{(2)})^T}_{Term_{III}}$$

We can decompose the perturbation to this gradient in terms of the perturbations to these three terms :

$$|\langle\nabla_{\tilde{g}_t}\mathcal{T}(X,\tilde{\Theta}) - \nabla_{g_t}\mathcal{T}(X,\Theta), e_i\rangle|$$
$$= |(\tilde{Term}_I - Term_I)e_i| + |(\tilde{Term}_{II} - Term_{II})e_i| + |(\tilde{Term}_{III} - Term_{III})e_i|$$

We begin by decomposing the perturbation in $Term_I$ into two components, using the triangle inequality.

$$|(\tilde{Term}_I - Term_I)e_i| \leq \left|e_t^T(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})(V^{(2)})^Te_i\right| + \left|e_t^T\phi^{(2)}_{T+1}(\tilde{V}^{(2)} - V^{(2)})^Te_i\right|$$
$$\leq \sqrt{\omega}\left\|\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}\right\|_\infty I[i = d+1] + |c_t|\left|(\tilde{V}^{(2)} - V^{(2)})^Te_i\right|$$

We show in 25 that

$$\left\|\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}\right\|_\infty \lessapprox 128\sigma D_f^2\sqrt{\omega}log(T)$$

In addition, we note that $\left|(\tilde{V}^{(2)} - V^{(2)})^Te_i\right| \leq \sigma R(\delta, 1)$. Thus we have

$$|(\tilde{Term}_I - Term_I)e_i| \lessapprox 128\sigma D_f^2\omega log(T)I[i = d+1] + \sigma R(\delta, 1)$$

Moving on to the perturbation of $Term_{II}$, we once again apply the triangle inequality:

$$|(\tilde{Term}_{II} - Term_{II})e_i| \leq \left|e_t^T\tilde{\phi}'^{(2)}_{T+1}\tilde{G}\tilde{V}^{(2)}e^T_{T+1}\tilde{G}\tilde{W}^{(2)}e_i - e_t^T\phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}GW^{(2)}e_i\right|$$

By the triangle inequality, this is bounded by

$$\left|e_t^T\left(\tilde{\phi}'^{(2)}_{T+1}\tilde{G}\tilde{V}^{(2)}e^T_{T+1}\tilde{G} - \phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}G\right)W^{(2)}e_i\right| - \left|e_t^T\phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}G(\tilde{W}^{(2)} - W^{(2)})e_i\right|$$

Note that the first term above terms vanishes when $i = d+1$: the term with only the unperturbed $W^{(2)}$ matrix vanishes, because the final, $(d+1)^{th}$ row and column are all all $0s$. In the general case where $i < d+1$,

$$|(\tilde{Term}_{II} - Term_{II})e_i| \leq \left|e_t^T\tilde{\phi}'^{(2)}_{T+1}\tilde{G}\tilde{V}^{(2)}e^T_{T+1}\tilde{G}\tilde{W}^{(2)}e_i - e_t^T\phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}GW^{(2)}e_i\right|$$
$$\lessapprox \underbrace{\left\|e_t^T(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})GV^{(2)}e^T_{T+1}GW^{(2)}\right\|}_{Term_A} I[i < d+1]$$
$$+ \underbrace{\left\|e_t^T\phi'^{(2)}_{T+1}(\tilde{G}\tilde{V}^{(2)} - GV^{(2)})e^T_{T+1}GW^{(2)}\right\|}_{Term_B} I[i < d+1]$$
$$+ \underbrace{\left\|e_t^T\phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}G(\tilde{W}^{(2)} - W^{(2)})\right\|}_{Term_C}$$
$$+ \underbrace{\left\|e_t^T\phi'^{(2)}_{T+1}GV^{(2)}e^T_{T+1}(\tilde{G} - G)W^{(2)}\right\|}_{Term_D} I[i < d+1]$$

<need to complete this>

$$Term_A = \left\|e_t^T(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})GV^{(2)}e^T_{T+1}GW^{(2)}\right\| \leq \left\|(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})e_t\right\|_1\left\|GV^{(2)}\right\|_\infty\left\|e^T_{T+1}GW^{(2)}\right\|,$$

where we have used the symmetry of $\tilde{\phi}'_t - \phi'_t$ and Holder's inequality. Note that by 10 , $\left\|GV^{(2)}\right\|_\infty \leq \sqrt{\omega}$, and $\left\|e^T_{T+1}GW^{(2)}\right\| = ||(W^{(2)})^Tg_{T+1}|| \lessapprox 2\sqrt{\omega}log(T)$. We showed in 27 that $\left\|(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})e_t\right\|_1 \leq 384\sigma D_f^2\sqrt{\omega}log(T)$. Plugging this bound in, we have that

$$Term_A = \left\|e_t^T(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})GV^{(2)}e^T_{T+1}GW^{(2)}\right\| \lessapprox 768\sigma D_f^2\omega^{\frac{3}{2}}log^2(T)$$

For $Term_B$, we can use Holder's inequality:

$$Term_B = \left\| e_t^T \phi_{T+1}^{\prime(2)} (\tilde{G}\tilde{V}^{(2)} - GV^{(2)}) e_{T+1}^T G W^{(2)} \right\| \leq \left\| e_t^T \phi_{T+1}^{\prime(2)} \right\|_1 \left\| (\tilde{G}\tilde{V}^{(2)} - GV^{(2)}) \right\|_\infty \left\| e_{T+1}^T G W^{(2)} \right\|$$

Note that $\left\| e_t^T \phi_{T+1}^{\prime(2)} \right\|_1 \leq \left\| \phi_{T+1}^{\prime(2)} \right\|_{1,1}$ is the maximum absolute row sum of the softmax Jacobian. Again from Lemma A.6 in [12], we know that this is bounded by 2. To bound $\left\| (\tilde{G}\tilde{V}^{(2)} - GV^{(2)}) \right\|_\infty$ we start with the triangle inequality:

$$\left\| (\tilde{G}\tilde{V}^{(2)} - GV^{(2)}) \right\|_\infty \lessapprox \left\| (G(\tilde{V}^{(2)} - V^{(2)}) \right\|_\infty + \left\| (\tilde{G} - G)V^{(2)} \right\|_\infty$$

$$\left\| G(\tilde{V}^{(2)} - V^{(2)}) \right\|_\infty \leq ||G||_{2,\infty} ||\zeta_{V^{(2)}}|| \leq \sqrt{2} ||\zeta_{V^{(2)}}||$$

$||\zeta_{V^{(2)}}||$ is a chi-distributed variable with $d+1$ degrees of freedom, and this with probability $1-\delta$,

$$||\zeta_{V^{(2)}}|| \leq \sigma\sqrt{d+1} R(\delta, d+1) \approx \sigma\sqrt{d}.$$

Therefore

$$\left\| G(\tilde{V}^{(2)} - V^{(2)}) \right\|_\infty \lessapprox \sigma\sqrt{2d}$$

Now, we have

$$\left\| (\tilde{G} - G)V^{(2)} \right\|_\infty \leq \left\| (\tilde{G} - G) \right\|_{2,\infty} ||V^{(2)}|| \leq D \leq \sqrt{\omega} \left\| (\tilde{G} - G) \right\|_{2,\infty}$$

Where we have used our bound on $D = \sum_{t=1}^{\omega}$ from 10. Combining this with our bound on $||\tilde{g}_t - g_t||_2$ in 21, this yields

$$\left\| (\tilde{G} - G)V^{(2)} \right\|_\infty \lessapprox 32\sqrt{\omega}\sigma D_f^2$$

Thus

$$\left\| (\tilde{G}\tilde{V}^{(2)} - GV^{(2)}) \right\|_\infty \lessapprox \sigma\sqrt{2d} + 32\sqrt{\omega}\sigma D_f^2 \approx 32\sqrt{\omega}\sigma D_f^2$$

We showed in 10 that $||(W^{(2)})^T g_{T+1}|| \lessapprox 2\sqrt{\omega} log(T)$. Combining these results together, we conclude that

$$Term_B \leq 64\sigma\omega D_f^2 log(T)$$

Moving on to $Term_C$, we have

$$Term_C = ||e_t^T \phi_{T+1}^{\prime(2)} G V^{(2)} e_{T+1}^T (G(\tilde{W}^{(2)} - W^{(2)})||$$

$$\leq \left\| e_t^T \tilde{\phi}_{T+1}^{\prime(2)} \right\|_1 \left\| GV^{(2)} \right\|_\infty \left\| e_{T+1}^T (G(\tilde{W}^{(2)} - W^{(2)}) \right\|.$$

The first two terms have already been bounded, so we focus on the third term. To bound $\left\| e_{T+1}^T G(\tilde{W}^{(2)} - W^{(2)}) \right\|$, we follow a similar argument to that of **??**

$$\left\| g_{T+1}^T \big(\tilde{W}^{(2)} - W^{(2)}\big) \right\|_2 \leq \left\| G\big(\tilde{W}^{(2)} - W^{(2)}\big) \right\|_{2,\infty}$$

Note that the perturbation matrix $\zeta_{W^{(2)}} = \tilde{W}^{(2)} - W^{(2)}$ has the same exact distribution if it is transposed. Therefore we can use the same bound from 25 for $\left\| G\big(\tilde{W}^{(2)} - W^{(2)}\big)^T \right\|_{2,\infty}$ to conclude that

$$\left\| G\big(\tilde{W}^{(2)} - W^{(2)}\big) \right\|_{2,\infty} \lessapprox \sigma d$$

Therefore

$$Term_C \leq 2\sigma\sqrt{\omega} d$$

Now we consider $Term_D$. First, note that

$$\left\| (\tilde{G} - G)W^{(2)} \right\|_{2,\infty} \leq ||\tilde{G} - G||_{2,\infty} ||W^{(2)}||_2$$

It is shown in 10 that $||W^{(2)}||_2 \lessapprox 2\sqrt{\omega}log(T)$. We have already bounded $||\tilde{G} - G||_{2,\infty} \lessapprox 32\sigma D_f^2$. Therefore,

$$Term_D \lessapprox 64\sigma\sqrt{\omega}dD_f^2log(T)$$

Finally, we can bound

$$||\tilde{Term}_{II} - Term_{II}|| \leq \left\|e_t^T \tilde{\phi}_{T+1}^{\prime(2)} \tilde{G}\tilde{V}^{(2)} e_{T+1}^T \tilde{G}\tilde{W}^{(2)}\right\|$$

$$\lessapprox 768\sigma D_f^2 \omega^{\frac{3}{2}} log^2(T) I[i < d+1] + 64\sigma\omega D_f^2 log(T) I[i < d+1]$$
$$+2\sigma\sqrt{\omega}d + 64\sigma\sqrt{\omega}dD_f^2 log(T) I[i < d+1]$$
$$\approx 768\sigma D_f^2 \omega^{\frac{3}{2}} log^2(T) I[i < d+1] + 2\sigma\sqrt{\omega}d$$

For the perturbation in $Term_{III}$, we have

$$|(\tilde{Term}_{III} - Term_{III})e_i| \leq I[t = T+1]\left|\left((\tilde{V}^{(2)})^T \tilde{G}^T \tilde{\phi}_{T+1}^{\prime(2)} \tilde{G}(\tilde{W}^{(2)})^T - (V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} G(W^{(2)})^T\right) e_i\right|$$

By the triangle equality,

$$\left\|(\tilde{V}^{(2)})^T \tilde{G}^T \tilde{\phi}_{T+1}^{\prime(2)} \tilde{G}(\tilde{W}^{(2)})^T - (V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} G(W^{(2)})^T\right\|$$
$$\leq \underbrace{\left\|(\tilde{G}\tilde{V}^{(2)} - GV^{(2)})^T \phi_{T+1}^{\prime(2)} G(W^{(2)})^T\right\|}_{Term_A} I[i < d+1]$$
$$+ \underbrace{\left\|(V^{(2)})^T G^T (\tilde{\phi_{T+1}^{\prime(2)}} - \phi_{T+1}^{\prime(2)}) G(W^{(2)})^T\right\|}_{Term_B} I[i < d+1]$$
$$\underbrace{\left\|(V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} G(\tilde{W}^{(2)} - W^{(2)})^T)\right\|}_{Term_C}$$
$$\underbrace{\left\|(V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} (\tilde{G} - G)(W^{(2)})^T)\right\|}_{Term_C} I[i < d+1]$$

First we handle $Term_A$. By 16, we have $\left\|(\tilde{G}\tilde{V}^{(2)} - GV^{(2)})^T \phi_{T+1}^{\prime(2)} G(W^{(2)})^T\right\| \leq 2\left\|(\tilde{G}\tilde{V}^{(2)} - GV^{(2)})\right\|_\infty \left\|G(W^{(2)})^T\right\|_{2,\infty}$ In our above bound on the perturbation to $Term_{II}$, we showed that

$$\left\|(\tilde{G}\tilde{V}^{(2)} - GV^{(2)})\right\|_\infty \lessapprox 32\sqrt{\omega}\sigma D_f^2$$

In addition, we have from 10 that $\left\|G(W^{(2)})^T\right\|_{2,\infty} \lessapprox 2log(T)$ Thus $Term_A \lessapprox 128\sqrt{\omega}\sigma D_f^2 log(T)$ Next, we turn to $Term_B$.

$$Term_B \leq \left\|(V^{(2)})^T G^T (\tilde{\phi_{T+1}^{\prime(2)}} - \phi_{T+1}^{\prime(2)})\right\|_1 \left\|(G(W^{(2)})^T)^T\right\|_{1,2}$$
$$\leq \left\|(V^{(2)})^T G^T (\tilde{\phi_{T+1}^{\prime(2)}} - \phi_{T+1}^{\prime(2)})\right\|_1 \left\|G(W^{(2)})^T\right\|_{2,\infty}$$

In order to bound the term $\left\|(V^{(2)})^T G^T (\tilde{\phi_{T+1}^{\prime(2)}} - \phi_{T+1}^{\prime(2)})\right\|_1$, we can use the same overall argument as above for bounding $\left\|(\tilde{\phi}_{T+1}^{\prime(2)} - \phi_{T+1}^{\prime(2)})e_t\right\|_1$. The only difference in the argument is that each of the three main terms will have a factor of $|(V^{(2)})^T G^T \phi_{T+1}^{(2)}|$ out front (instead of $|e_t^T \phi_{T+1}^{(2)}|$). When we apply Holder's inequality to this, we gain a factor of $||GV^{(2)})||_\infty$, which we have already bounded as $\lessapprox \sqrt{\omega}$. Therefore we conclude that

$$\left\|(V^{(2)})^T G^T (\tilde{\phi_{T+1}^{\prime(2)}} - \phi_{T+1}^{\prime(2)})\right\|_1 \lessapprox 192\sigma D_F(d + 2D_f)\omega log(T)$$

Using our bound on $\left\|G(W^{(2)})^T\right\|_{2,\infty}$ from 10, it follows that

$$Term_B \lessapprox 768\sigma D_f^2 \omega log^2(T)$$

Finally, for $Term_C$ we have by 16 that

$$\left\|(V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} G(\tilde{W}^{(2)} - W^{(2)})^T\right\| \leq 2\left\|GV^{(2)}\right\|_\infty \left\|G(\tilde{W}^{(2)} - W^{(2)})^T\right\|_{2,\infty}$$

We have already bounded $\left\|G(\tilde{W}^{(2)} - W^{(2)})^T\right\|_{2,\infty} \lessapprox \sigma d$ in the proof of 25. Thus, we have

$$Term_C \lessapprox 2\sigma\sqrt{\omega} d$$

For $Term_D$, we similarly have

$$\left\|(V^{(2)})^T G^T \phi_{T+1}^{\prime(2)} (\tilde{G} - G)(W^{(2)})^T\right\| \leq 2\left\|GV^{(2)}\right\|_\infty \left\|(\tilde{G} - G)(W^{(2)})^T\right\|_{2,\infty}$$

Using standard matrix norm inequalities, we have

$$\left\|(\tilde{G} - G)(W^{(2)})^T\right\|_{2,\infty} \leq \left\|\tilde{G} - G\right\|_{2,\infty} \left\|W^{(2)}\right\|_2$$

We show in 10 that $\left\|W^{(2)}\right\|_2 \lessapprox 2\sqrt{\omega} log(T)$. Therefore we have

$$\left\|(\tilde{G} - G)(W^{(2)})^T\right\|_{2,\infty} \lessapprox 64\sigma\sqrt{\omega} D_f^2 log(T)$$

Therefore

$$Term_D \lessapprox 128\sigma\omega D_f^2 log(T)$$

Combining $Term_A, Term_B, Term_C, Term_D$ we have

$$||\tilde{Term}_{III} - Term_{III}||$$

$$\lessapprox 128\sigma\sqrt{\omega} D_f^2 log(T) I[i < d+1] + 768\sigma D_f^2 \omega log^2(T) I[i < d+1]$$

$$+2\sigma\sqrt{\omega} d + 128\sigma\omega D_f^2 log(T) I[i < d+1]$$

$$\approx 768\sigma D_f^2 \omega log^2(T) I[i < d+1] + 2\sigma\sqrt{\omega} d$$

Finally, we have

$$|(\nabla_{\tilde{g}_t} \mathcal{T}(X, \tilde{\Theta}) - \nabla_{g_t} \mathcal{T}(X, \Theta)) e_{d+1}|$$

$$\leq |(\tilde{Term}_I - Term_I) e_{d+1}| + |(\tilde{Term}_{II} - Term_{II}) e_{d+1}| + |(\tilde{Term}_{III} - Term_{III}) e_{d+1}|$$

$$\lessapprox 128\sigma D_f^2 \omega log(T) I[i = d+1] + \sigma R(\delta, 1) d$$

$$+768\sigma D_f^2 \omega^{\frac{3}{2}} log^2(T) I[i < d+1] + 2\sigma\sqrt{\omega} d$$

$$+768\sigma D_f^2 \omega log^2(T) I[i < d+1] + 2\sigma\sqrt{\omega} d$$

$$\approx 128\sigma D_f^2 \omega log(T) I[i = d+1] + 768\sigma D_f^2 \omega^{\frac{3}{2}} log^2(T) I[i < d+1]$$

□

Theorem 27. *Let $\tilde{\phi}_{T+1}^{\prime(2)}, \phi_{T+1}^{\prime(2)}$ be the perturbed an unperturbed Jacobian of the softmax scores for the CLS position in the second attention layer in our transformer. Then the following bound holds for the maximum row/column 2-norm (noting that both $\tilde{\phi}_{T+1}^{\prime(2)}, \phi_{T+1}^{\prime(2)}$ are symmetric) of $\tilde{\phi}_{T+1}^{\prime(2)} - \phi_{T+1}^{\prime(2)}$:*

$$\left\|(\tilde{\phi}_{T+1}^{\prime(2)} - \phi_{T+1}^{\prime(2)}) e_t\right\| \leq \left\|(\tilde{\phi}_{T+1}^{\prime(2)} - \phi_{T+1}^{\prime(2)}) e_t\right\|_1 \leq 384\sigma D_f^2 \sqrt{\omega} log(T)$$

Proof. We use the well-known Jacobian of the softmax to obtain

$$(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1}) = diag(\tilde{\phi}^{(2)}_f) - \tilde{\phi}^{(2)}_{T+1}(\tilde{\phi}^{(2)}_{T+1})^T - diag(\phi^{(2)}_{T+1}) + \phi^{(2)}_{T+1}(\phi^{(2)}_{T+1})^T$$
$$= diag(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}) + \tilde{\phi}^{(2)}_{T+1}(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})^T + \phi^{(2)}_{T+1}(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})^T$$

Taking the 1-norm of $(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})e_t$ and then applying the triangle inequality, we have

$$\left\|(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})e_t\right\|$$
$$\leq \left\|diag(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})e_t\right\|_1 + \left\|\tilde{\phi}^{(2)}_{T+1}(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})^T e_t\right\|_1 + \left\|\phi^{(2)}_{T+1}(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})^T e_t\right\|_1$$

The 1-norm of $diag(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})e_t$ is upper bounded by $||\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}||_\infty \leq ||\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}||_1$. Meanwhile,

$$\left\|e_t^T \tilde{\phi}^{(2)}_{T+1}\left(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}\right)^T\right\|_1 \approx \left\|e_t^T \phi^{(2)}_{T+1}\left(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}\right)^T\right\|_1$$
$$= |e_t^T \phi^{(2)}_{T+1}|\left\|(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})^T)\right\|_1$$
$$\leq \left\|\phi^{(2)}_{T+1}\right\|_1\left\|(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})^T)\right\|_1 = \left\|(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})^T)\right\|_1$$

Putting this together, we have

$$\left\|(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})e_t\right\| \leq 3\left\|(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})^T)\right\|_1$$

$\left\|\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}\right\|_1$ was already bounded with high probability in 25. Plugging in this bound, we have

$$\left\|(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1})e_t\right\| \leq 384\sigma D_f^2\sqrt{\omega}log(T)$$

□

Theorem 28. *Given a set of remainder perturbations ζ to the parameters Θ such that $|\zeta_i| \leq |\epsilon_i|$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)^n$, the norm of the perturbed gradients satisfies:*

$$||\nabla\mathcal{T}(X, \Theta + \zeta) - \nabla\mathcal{T}(X, \Theta)|| \leq G_p(\sigma, \omega, D_f, T)$$

where

$$G_p(\omega, D_f, T, d) \in o(\sigma D_f^{\frac{7}{2}}\omega^{\frac{3}{2}}log(T)^2)$$

Proof. **M.2** $||\nabla_{\tilde{V}^{(2)}}\mathcal{T}(X, \tilde{\Theta}) - \nabla_{V^{(2)}}\mathcal{T}(X, \Theta)||$

Consider the gradient $\nabla_{V^{(2)}}\mathcal{T}(X, \Theta) = G^T\phi^{(2)}_{T+1}$, the norm of which we upper bounded with $||\nabla_{V^{(2)}}\mathcal{T}(X, \Theta)|| \leq \sqrt{2}$ above. If we perturb the network weights slightly, we have the following perturbed output:

$$\nabla_{\tilde{V}^{(2)}}\mathcal{T}(X, \tilde{\Theta}) = \tilde{G}^T\tilde{\phi}^{(2)}_{T+1}$$

Taking the norm of the perturbation, we have

$$\left\|\nabla_{\tilde{V}^{(2)}}\mathcal{T}(X, \tilde{\Theta}) - \nabla_{V^{(2)}}\mathcal{T}(X, \Theta)\right\|$$
$$= \left\|\tilde{G}^T\tilde{\phi}^{(2)}_{T+1} - G^T\phi^{(2)}_{T+1}\right\| \leq \left\|(\tilde{G} - G)^T\phi^{(2)}_{T+1}\right\| + \left\|G^T(\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1})\right\|$$
$$\leq ||(\tilde{G} - G)^T||_{1,2}||\phi^{(2)}_{T+1}||_1 + ||G^T||_{1,2}||\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}||_1$$
$$\leq ||\tilde{G} - G||_{2,\infty} + ||G||_{2,\infty}||\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}||_1$$

We just proved in 21 that

$$\forall t : ||\tilde{g}_t - g_t|| \lessapprox 32\sigma D_f^2,$$

In addition, we showed in 10 that $||G||_{2,\infty} \lessapprox \sqrt{2}$, and in 25 we showed that

$$||\tilde{\phi}^{(2)}_{T+1} - \phi^{(2)}_{T+1}||_1 \lessapprox 128\sigma D_f^2\sqrt{\omega}log(T)$$

Thus

$$||\nabla_{\tilde{V}^{(2)}}\mathcal{T}(X, \Theta + \zeta) - \nabla_{\tilde{V}^{(2)}}\mathcal{T}(X, \Theta)|| \lessapprox 32\sigma D_f^2 + 128\sqrt{2}\sigma D_f^2\sqrt{\omega}log(T)$$
$$\approx 128\sqrt{2}\sigma D_f^2\sqrt{\omega}log(T)$$

M.3 $||\nabla_{\tilde{W}^{(2)}}\mathcal{T}(X,\tilde{\Theta}) - \nabla_{W^{(2)}}\mathcal{T}(X,\Theta)||$

Plugging our perturbed matrices into the formula for the unperturbed gradient of our transformer with respect to $W^{(1)}$,

$$\nabla_{W^{(2)}}\mathcal{T}(X,\Theta) = g_{T+1}(V^{(2)})^T G^T \phi'^{(2)}_{T+1} G,$$

Noting that the RHS is a rank-1 outer product, we have

$$\left\|\nabla_{W^{(2)}}\mathcal{T}(X,\Theta)\right\|_F = ||g_{T+1}||\left\|(V^{(2)})^T G^T \phi'^{(2)}_{T+1} G\right\|,$$

By the triangle inequality, we have

$$||\nabla_{\tilde{W}^{(2)}}\mathcal{T}(X,\Theta+\zeta)|| \leq ||\nabla_{W^{(2)}}\mathcal{T}(X,\Theta)|| + ||\nabla_{\tilde{W}^{(2)}}\mathcal{T}(X,\Theta+\zeta) - \nabla_{W^{(2)}}\mathcal{T}(X,\Theta)||$$

We can further decompose the perturbation term using the triangle inequality:

$$||\nabla_{\tilde{W}^{(2)}}\mathcal{T}(X,\Theta+\zeta) - \nabla_{W^{(2)}}\mathcal{T}(X,\Theta)|| = \left\|(\tilde{V}^{(2)})^T \tilde{G}^T \phi'^{(2)}_{T+1}\tilde{G} - (V^{(2)})^T G^T \phi'^{(2)}_{T+1} G\right\|$$

$$\left\|(\tilde{V}^{(2)} - V^{(2)})^T \tilde{G}^T \phi'^{(2)}_{T+1}\tilde{G}\right\| + \left\|(V^{(2)})^T (\tilde{G}-G)^T \phi'^{(2)}_{T+1}\tilde{G}\right\| + \left\|(V^{(2)})^T G^T \phi'^{(2)}_{T+1}(\tilde{G}-G)\right\|$$

From 16, and only keeping terms that are first order in σ, we have

$$\left\|(\tilde{V}^{(2)} - V^{(2)})^T \tilde{G}^T \phi'^{(2)}_{T+1}\tilde{G}\right\| \leq 2\left\|\tilde{G}(\tilde{V}^{(2)} - V^{(2)})\right\|_\infty ||\tilde{G}||_{2,\infty} \approx 2\left\|G(\tilde{V}^{(2)} - V^{(2)})\right\|_\infty ||G||_{2,\infty},$$

$$\left\|(V^{(2)})^T (\tilde{G}-G)^T \phi'^{(2)}_{T+1}\tilde{G}\right\| \leq 2\left\|(\tilde{G}-G)V^{(2)}\right\|_\infty ||\tilde{G}||_{2,\infty} \approx 2\left\|(\tilde{G}-G)V^{(2)}\right\|_\infty ||G||_{2,\infty}$$

$$\left\|(V^{(2)})^T G^T \phi'^{(2)}_{T+1}(\tilde{G}-G)\right\| \leq 2\left\|GV^{(2)}\right\|_\infty ||\tilde{G}-G||_{2,\infty}$$

We have already bounded $||\tilde{G}-G||_{2,\infty}$ in 21 and $||G||_{2,\infty}$, $\left\|GV^{(2)}\right\|_\infty$ in 10 . We show in 22 that the following two bounds hold:

$$\left\|G(\tilde{V}^{(2)} - V^{(2)})\right\|_\infty \lessapprox \sigma\sqrt{2d}$$

$$\left\|(\tilde{G}-G)V^{(2)}\right\|_\infty \lessapprox 32\sqrt{\omega}\sigma D_f^2$$

Putting this all together, we have

$$\begin{aligned}||\nabla_{\tilde{W}^{(2)}}\mathcal{T}(X,\Theta+\zeta) - \nabla_{W^{(2)}}\mathcal{T}(X,\Theta)|| &\lessapprox 4\sigma\sqrt{d} + 64\sqrt{2}\sigma\sqrt{\omega}D_f^2 + 64\sqrt{\omega}\sigma D_f^2 \\ &\leq 192\sigma\sqrt{\omega}D_f^2\end{aligned}$$

M.4 $||\nabla_{\tilde{W}^{(1)}}\mathcal{T}(X,\tilde{\Theta}) - \nabla_{W^{(1)}}\mathcal{T}(X,\Theta)||$

We showed in 15 that

$$\nabla_{W^{(1)}}\mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega} c_t g'_t x_t e_{d+1}^T V_1^T X^T \phi'_t X$$

However, this formula relied on the assumption that the dependency graph in our exact construction only involves the final dimensions of g_t and b_t. While true for our exact construction, this is no longer true when the parameters are perturbed. In the case of the perturbed construction, we need to use the more general formula provided in 15, applied to the perturbed transformer:

$$\nabla_{\tilde{W}^{(1)}}\mathcal{T}(X,\tilde{\Theta}) = \sum_{t=1}^{T+1}\sum_{i=1}^{d+1} \nabla^T_{\tilde{g}_t}\mathcal{T}(X,\tilde{\Theta})\left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i} \nabla_{\tilde{W}^{(1)}}\tilde{B}_{t,i}$$

Note that the sums in these gradients now go all the way up to $T+1$ rather than stopping at ω. This is because in a perturbed transformer, the MLP outputs for "empty" positions beyond the ω required to compute our function may no longer all be 0, and may contribute to the transformers output and gradients.

In order to bound the perturbed gradient, we first bound the norm of the perturbation, using the triangle inequality:

$$\begin{aligned}
&\left\|\nabla_{\tilde{W}^{(1)}}\mathcal{T}(X,\Theta+\zeta)-\nabla_{\tilde{W}^{(1)}}\mathcal{T}(X,\Theta)\right\|_F \\
&\lessapprox \underbrace{\sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\left(\nabla_{\tilde{g}_t}^T\mathcal{T}(X,\tilde{\Theta})-\nabla_{g_t}^T\mathcal{T}(X,\Theta)\right)\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i}\right|\left\|\nabla_{W^{(1)}}B_{t,i}\right\|_F}_{Term_1} \\
&+\underbrace{\sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\nabla_{g_t}^T\mathcal{T}(X,\Theta)\left(\left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i}-\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i}\right)\right|\left\|\nabla_{W^{(1)}}B_{t,i}\right\|_F}_{Term_2} \\
&+\underbrace{\sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\nabla_{g_t}^T\mathcal{T}(X,\Theta)\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i}\right|\left\|\nabla_{\tilde{W}^{(1)}}\tilde{B}_{t,i}-\nabla_{W^{(1)}}B_{t,i}\right\|_F}_{Term_3}
\end{aligned}$$

Recall that the only nonzero element of the unperturbed Jacobian $\mathcal{J}_{g_t}(b_t)$ is the last row and column, and therefore we can remove the sum over i in $Term_1$ and $Term_3$. For the final column of the Jacobian, we have $\left[\mathcal{J}_{g_t}(b_t)\right]_{:,d+1}=e_{d+1}\left|\frac{\partial G_{t,d+1}}{\partial B_{t,d+1}}\right|$. Recall that in our construction, the slope of the MLP output with respect to small changes around the points in the grid of possible values for $B_{t,d+1}$ for is 0, and thus we have $\left|\frac{\partial G_{t,d+1}}{\partial B_{t,d+1}}\right|=0$. Therefore, only $Term_2$ survives. We show in **??** that

$$\left|\left|\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i}-\left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i}\right|\right|\lessapprox 16\sigma D_f^2$$

In 24, we show that

$$||\nabla_{g_t}^T\mathcal{T}(X,\Theta)||\leq|c_t|+4\sqrt{\omega}log(T)I[t=T+1]+4\omega log(T)I[t\in[\omega]]$$

By Cauchy-Schwarz,

$$\begin{aligned}
\left|\nabla_{g_t}^T\mathcal{T}(X,\Theta)\left(\left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i}-\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i}\right)\right|&\leq\left\|\nabla_{g_t}\mathcal{T}(X,\Theta)\right\|\left|\left|\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i}-\left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i}\right|\right| \\
&\lessapprox 16\sigma D_f^2(|c_t|+4\sqrt{\omega}log(T)I[t=T+1]+4\omega log(T)I[t\in[\omega]])
\end{aligned}$$

Taking the sum over t and applying the indicator functions, we have the following:

$$\begin{aligned}
&\sum_{t=1}^{T+1}\left\|\nabla_{g_t}\mathcal{T}(X,\Theta)\right\|\left|\left|\left[\mathcal{J}_{g_t}(b_t)\right]_{:,i}-\left[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\right]_{:,i}\right|\right|\left\|\nabla_{W^{(1)}}B_{t,i}\right\|_F \\
&=16\sigma D_f^2\sum_{t=1}^{T+1}\left(|c_t|+4\sqrt{\omega}log(T)I[t=T+1]+4\omega log(T)I[t\in[\omega]]\right)\left\|\nabla_{W^{(1)}}B_{t,i}\right\|_F \\
&=16\sigma D_f^2\left(\sqrt{\omega}+4\sqrt{\omega}log(T)+4\omega^2log(T)\right)\left\|\nabla_{W^{(1)}}B_{t,i}\right\|_F \\
&\approx 64D_f^2\omega^2log(T)\left\|\nabla_{W^{(1)}}B_{t,i}\right\|_F
\end{aligned}$$

Finally, we handle the $\left\|\nabla_{W^{(1)}}B_{t,i}\right\|_F$ term. It was shown in 15 that

$$\left\|\nabla_{W_1}B_{t,i}\right\|_F\lessapprox 2\sqrt{2}I[i=d+1].$$

Thus we have

$$\begin{aligned}
\left\|\nabla_{\tilde{W}^{(1)}}\mathcal{T}(X,\Theta+\zeta)-\nabla_{\tilde{W}^{(1)}}\mathcal{T}(X,\Theta)\right\|_F&\lessapprox\sum_{i=1}^{d+1}64D_f^2\omega^2log(T)\left\|\nabla_{W^{(1)}}B_{t,i}\right\|_F \\
&\leq 128\sqrt{2}D_f^2\omega^2log(T)
\end{aligned}$$

M.5 $||\nabla_{\tilde{V}^{(1)}}\mathcal{T}(X,\tilde{\Theta}) - \nabla_{V^{(1)}}\mathcal{T}(X,\Theta)||$

We once again use the more general expression for our gradient with respect to $V^{(1)}$ that uses the full Jacobian.

$$\nabla_{\tilde{V}^{(1)}}\mathcal{T}(X,\tilde{\Theta}) = \sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\nabla_{\tilde{g}_t}^T\mathcal{T}(X,\tilde{\Theta})\Big[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\Big]_{:,i}\nabla_{\tilde{V}^{(1)}}\tilde{B}_{t,i}$$

After canceling out the terms involving the unperturbed Jacobian of g_t with respect to b_t and keeping only the term involving the perturbation of the Jacobian, we have:

$$\Big\|\nabla_{\tilde{V}^{(1)}}\mathcal{T}(X,\Theta+\zeta) - \nabla_{\tilde{V}^{(1)}}\mathcal{T}(X,\Theta)\Big\|_F$$

$$\lessapprox \sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\big|\nabla_{g_t}^T\mathcal{T}(X,\Theta)\big(\big[\mathcal{J}_{\tilde{g}_t}(\tilde{b}_t)\big]_{:,i} - \big[\mathcal{J}_{g_t}(b_t)\big]_{:,i}\big)\big|\Big\|\nabla_{V^{(1)}}B_{t,i}\Big\|_F$$

$$\approx 64D_f^2\omega^2 log(T)\sum_{i=1}^{d+1}\Big\|\nabla_{V^{(1)}}B_{t,i}\Big\|_F$$

It was shown in 15 that

$$\Big\|\nabla_{V^{(1)}}B_{t,i}\Big\|_F \lessapprox \sqrt{2}$$

Thus

$$\Big\|\nabla_{\tilde{V}^{(1)}}\mathcal{T}(X,\Theta+\zeta) - \nabla_{\tilde{V}^{(1)}}\mathcal{T}(X,\Theta)\Big\|_F \lessapprox 64\sqrt{2}dD_f^2\omega^2 log(T)$$

M.6 $||\nabla_{\tilde{M}}\mathcal{T}(X,\tilde{\Theta}) - \nabla_M\mathcal{T}(X,\Theta)||$

It was shown in 15 that

$$\nabla_M\mathcal{T}(X,\Theta) = \sum_{t=1}^{\omega}c_t b_t e_{d+1}^T F^T diag(I\big[M^T b_t + \Gamma > 0\big])$$

In the case of the perturbed construction, we once again need to use a slightly more general formula that accounts for the possibility of the transformer output depending on the first d dimensions of g_t.

$$\nabla_{\tilde{M}}\mathcal{T}(X,\tilde{\Theta}) = \sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\frac{\partial\mathcal{T}(X,\tilde{\Theta})}{\partial\tilde{G}_{t,i}}\nabla_{\tilde{M}}\tilde{G}_{t,i}$$

$$\Big\|\nabla_{\tilde{M}}\mathcal{T}(X,\Theta+\zeta) - \nabla_M\mathcal{T}(X,\Theta)\Big\|_F$$

$$\lessapprox \sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\frac{\partial\mathcal{T}(X,\tilde{\Theta})}{\partial\tilde{G}_{t,i}} - \frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\Big\|\nabla_M G_{t,i}\Big\|_F + \left|\frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\Big\|\nabla_{\tilde{M}}\tilde{G}_{t,i} - \nabla_M G_{t,i}\Big\|_F$$

It was shown in 15 that

$$\nabla_M G_{t,i} = b_t e_i^T F^T diag(I\big[M^T b_t + \Gamma > 0\big]),$$

and that the norm of this intra-layer gradient term can be bounded as:

$$\Big\|\nabla_M G_{t,i}\Big\|_F \lessapprox 8D_f^{\frac{3}{2}}I[i = d+1 \wedge t \leq \omega]$$

In 26, it was shown that

$$|\langle\nabla_{\tilde{g}_t}\mathcal{T}(X,\tilde{\Theta}) - \nabla_{g_t}\mathcal{T}(X,\Theta), e_{d+1}\rangle| \lessapprox 128\sigma D_f^2\sqrt{\omega}log(T)$$

Note that, unlike the unperturbed gradient with respect to M, we no longer have the unperturbed derivative with respect to $G_{t,d+1}$, $\frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,d+1}} = c_t$, to cancel out the gradient for the "inactive" positions after the first attention layer (with $t > \omega$). However, since the unperturbed intra-layer gradient

$\nabla_M G_{t,i}$ is also 0 for $t > \omega$, we can still avoid the T dependency in our gradient perturbation. Thus we have

$$\sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\frac{\partial \mathcal{T}(X,\tilde{\Theta})}{\partial \tilde{G}_{t,i}} - \frac{\partial \mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\left\|\nabla_M G_{t,i}\right\|_F$$
$$= 8D_f^{\frac{3}{2}}\sum_{t=1}^{\omega}\left|\frac{\partial \mathcal{T}(X,\tilde{\Theta})}{\partial \tilde{G}_{t,d+1}} - \frac{\partial \mathcal{T}(X,\Theta)}{\partial G_{t,d+1}}\right|$$
$$\leq 1024\sigma D_f^{\frac{7}{2}}\omega^{\frac{3}{2}}log(T)$$

In addition, note that by Cauchy-Schwarz,

$$\left|\frac{\partial \mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right| = \left|\nabla_{g_t}^T \mathcal{T}(X,\Theta)e_i\right| \leq \left\|\nabla_{g_t}\mathcal{T}(X,\Theta)\right\|$$

We have already shown in 28 that

$$\sum_{t=1}^{T+1}||\nabla_{g_t}\mathcal{T}(X,\Theta)|| \lessapprox \sqrt{\omega} + 4\sqrt{\omega}log(T) + 4\omega^2 log(T) \approx 4\omega^2 log(T)$$

We now consider the term $\left\|\nabla_{\tilde{M}}\tilde{G}_{t,i} - \nabla_M G_{t,i}\right\|_F$. Starting from our formula for the unperturbed gradient, we have

$$\left\|\nabla_{\tilde{M}}\tilde{G}_{t,i} - \nabla_M G_{t,i}\right\|_F \leq \left\|\tilde{b}_t e_i^T \tilde{F}^T diag(I\big[\tilde{M}^T\tilde{b}_t + \tilde{\Gamma} > 0\big]) - b_t e_i^T F^T diag(I\big[M^T b_t + \Gamma > 0\big])\right\|$$
$$\lessapprox ||b_t||\left\|e_i^T F^T\Big(diag(I\big[\tilde{M}^T\tilde{b}_t + \tilde{\Gamma} > 0\big]) - diag(I\big[M^T b_t + \Gamma > 0\big])\Big)\right\|$$
$$+||b_t||\left\|e_i^T(\tilde{F} - F)^T diag(I\big[M^T b_t + \Gamma > 0\big])\right\|$$
$$+||\tilde{b}_t - b_t||\left\|e_i^T F^T diag(I\big[M^T b_t + \Gamma > 0\big])\right\|$$

Just as we saw in 21 and then again in 23, with some small probability, some of the neurons in the perturbed transformer may flip from being active to inactive. We use the same assumption that $\sigma \leq \frac{1}{16dD_f}$ to that this will happen with probability at most δ, and thus $diag(I\big[\tilde{M}^T\tilde{b}_t + \tilde{\Gamma} > 0\big]) - diag(I\big[M^T b_t + \Gamma > 0\big]) = \mathbf{0}_{4(D_f+1)\times 4(D_f+1)}$, and thus the first term will be 0.

Moving to the second additive term, we have with probability $1-\delta$

$$\left\|e_i^T(\tilde{F} - F)^T diag(I\big[M^T b_t + \Gamma > 0\big])\right\|$$
$$\leq ||\zeta_F^T||_{2,\infty} \leq ||\zeta_F||_{1,2} \leq ||\epsilon_F||_{1,2} \leq \sigma\sqrt{4(D_f+1)}R(\frac{\delta}{d+1}, 4(D_f+1))$$
$$\approx 2\sigma D_f^{\frac{1}{2}}$$

For the last term, recall that all but the last column of F is 0, we have

$$\left\|e_i^T F^T diag(I\big[b_t M + \Gamma^T > 0\big])\right\| \leq ||F||_{1,2} = 4D_f\sqrt{4(D_f+1)}I[i = d+1] \approx 8D_f^{\frac{3}{2}}I[i = d+1]$$

Using this, and combining our bound for $||\tilde{b}_t - b_t|| \lessapprox 4\sigma d$ from 20, we have

$$\left\|\nabla_{\tilde{M}}\tilde{G}_{t,i} - \nabla_M G_{t,i}\right\|_F \lessapprox 2\sigma D_f^{\frac{1}{2}} + 32\sigma d D_f^{\frac{3}{2}} I[i = d+1]$$

Therefore

$$\sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\frac{\partial \mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\left\|\nabla_{\tilde{M}}\tilde{G}_{t,i} - \nabla_M G_{t,i}\right\|_F$$
$$\leq 4\omega^2 log(T)\sum_{i=1}^{d+1}||b_t||\left\|e_i^T(\tilde{F} - F)^T diag(I\big[M^T b_t + \Gamma > 0\big])\right\|$$

$$+4\omega^2 log(T)\sum_{i=1}^{d+1}||\tilde{b}_t - b_t||\Big\|e_i^T F^T diag(I\big[M^T b_t + \Gamma > 0\big])\Big\|$$

$$\leq 4\omega^{@} log(T)\sum_{i=1}^{d+1}\left(2\sigma D_f^{\frac{1}{2}}\right)$$

$$+4\omega^2 log(T)(4\sigma d)(8D_f^{\frac{3}{2}})$$

$$\leq 8\sigma d\omega^2 D_f^{\frac{1}{2}} log(T) + 128\sigma d\sqrt{\omega} log(T) D_f^{\frac{3}{2}}$$

$$\lessapprox 128\sigma d\omega^2 log(T) D_f^{\frac{3}{2}}$$

M.7 $||\nabla_{\tilde{\Gamma}}\mathcal{T}(X,\tilde{\Theta}) - \nabla_{\Gamma}\mathcal{T}(X,\Theta)||$

We have for the perturbed transformer the following expression for the gradient:

$$\nabla_{\tilde{\Gamma}}\mathcal{T}(X,\tilde{\Theta}) = \sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\frac{\partial\mathcal{T}(X,\tilde{\Theta})}{\partial\tilde{G}_{t,i}}\nabla_{\tilde{\Gamma}}\tilde{G}_{t,i}$$

We again decompose the perturbation in the norm of the gradient using the triangle inequality:

$$\Big\|\nabla_{\tilde{\Gamma}}\mathcal{T}(X,\Theta+\zeta) - \nabla_{\Gamma}\mathcal{T}(X,\Theta)\Big\|_F$$

$$\lessapprox \sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\frac{\partial\mathcal{T}(X,\tilde{\Theta})}{\partial\tilde{G}_{t,i}} - \frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\Big\|\nabla_{\Gamma}G_{t,i}\Big\|_F + \left|\frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\Big\|\nabla_{\tilde{\Gamma}}\tilde{G}_{t,i} - \nabla_{\Gamma}G_{t,i}\Big\|_F$$

We just upper bounded both of the inter-layer derivatives, $\left|\frac{\partial\mathcal{T}(X,\tilde{\Theta})}{\partial\tilde{G}_{t,i}} - \frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|$ and $\left|\frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|$. It remains to bound the perturbation in the intra-layer gradient. We show in 15 that

$$||\nabla_{\Gamma}G_{t,i}|| \lessapprox 8D_f^{\frac{3}{2}} I[i = d+1 \wedge t \leq \omega]$$

The first sum becomes

$$\sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\frac{\partial\mathcal{T}(X,\tilde{\Theta})}{\partial\tilde{G}_{t,i}} - \frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\Big\|\nabla_{\Gamma}G_{t,i}\Big\|_F$$

$$\leq 8D_f^{\frac{3}{2}}\sum_{t=1}^{\omega}128\sigma D_f^2\omega log(T)$$

$$= 1024\sigma D_f^{\frac{7}{2}}\omega^2 log(T)$$

Now, turning to the second sum, we have

$$||\nabla_{\tilde{\Gamma}}\tilde{G}_{t,i} - \nabla_{\Gamma}G_{t,i}||$$

$$\leq \Big\|diag(I\big[\tilde{M}^T\tilde{b}_t + \tilde{\Gamma} \geq 0\big])\tilde{F}e_i - diag(I\big[M^T b_t + \Gamma \geq 0\big])Fe_i\Big\|$$

We again apply the triangle inequality, and apply our first-order (in σ) approximation to the perturbation, noting that the term involve the difference of indicators vanishes, assuming that σ is small in the sense of 21. Thus we obtain

$$||\nabla_{\tilde{\Gamma}}\tilde{G}_{t,i} - \nabla_{\Gamma}G_{t,i}|| \lessapprox \Big\|diag(I\big[M^T b_t + \Gamma > 0\big])(\tilde{F} - F)e_{d+1}\Big\|$$

$$\leq ||\zeta_{\Gamma}||_{1,2} \leq ||\epsilon_{\Gamma}||_{1,2} \leq \sigma\sqrt{4(D_f+1)}R(\frac{\delta}{d+1}, 4(D_f+1)) \approx 2\sigma\sqrt{D_f}$$

Putting this together, for the second sum we have

$$\sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\Big\|\nabla_{\tilde{\Gamma}}\tilde{G}_{t,i} - \nabla_{\Gamma}G_{t,i}\Big\|_F$$

$$\leq 4\sqrt{\omega}log(T)\sum_{i=1}^{d+1}\left\|\nabla_{\tilde{\Gamma}}\tilde{G}_{t,i}-\nabla_{\Gamma}G_{t,i}\right\|_F$$
$$\leq 4\sqrt{\omega}log(T)\Big(2\sigma\sqrt{D_f}\Big)(d+1)\approx 8\sigma D_f^{\frac{1}{2}}\sqrt{\omega}log(T)d$$

Putting the two sums together, we have

$$\left\|\nabla_{\tilde{\Gamma}}\mathcal{T}(X,\Theta+\zeta)-\nabla_{\Gamma}\mathcal{T}(X,\Theta)\right\|_F$$
$$\leq 1024\sigma D_f^{\frac{7}{2}}\omega^{\frac{3}{2}}log(T)+8\sigma D_f^{\frac{1}{2}}\sqrt{\omega}log(T)d$$

M.8 $||\nabla_{\tilde{F}}\mathcal{T}(X,\tilde{\Theta})-\nabla_F\mathcal{T}(X,\Theta)||$

We follow the same overall flow as the previous two gradient calculations. We showed in 15 that

$$||\nabla_F G_{t,i}||\leq||\big(M^Tb_t+\Gamma\big)_+e_{t,i}^T||\lessapprox 2\sqrt{D_f}I[t\leq\omega]$$

In addition, we note that

$$||\nabla_{\tilde{F}}\tilde{G}_{t,i}-\nabla_F G_{t,i}||_F$$
$$\leq\left\|\big(\tilde{M}^T\tilde{b}_t+\tilde{\Gamma}\big)_+e_{t,i}^T-\big(M^Tb_t+\Gamma\big)_+e_{t,i}^T\right\|_F=\left\|\big(\tilde{M}^T\tilde{b}_t+\tilde{\Gamma}\big)_+-\big(M^Tb_t+\Gamma\big)_+\right\|$$

Using the triangle inequality, we have

$$\lessapprox\left\|\big(\tilde{M}^Tb_t+\tilde{\Gamma}\big)-\big(M^Tb_t+\Gamma\big)\right\|+\left\|\big(M^T(\tilde{b}_t-b_t)\right\|$$

Where we have used the $1-$lipschitzness of the max function. It was shown in 21 that

$$||(\tilde{M}^T-M^T)b_t+(\tilde{\Gamma}-\Gamma)||_2\lessapprox 4\sigma D_f^{\frac{1}{2}}$$

and it is easy to see that $\left\|M^T(\tilde{b}_t-b_t)\right\|\leq\sqrt{4(D_f+1)}||\tilde{b}_t-b_t||\lessapprox 8\sigma D_f^{\frac{1}{2}}d$. Thus we obtain

$$\left\|\nabla_{\tilde{F}}\tilde{G}_{t,i}-\nabla_F G_{t,i}\right\|_F\lessapprox 4\sigma D_f^{\frac{1}{2}}+8\sigma D_f^{\frac{1}{2}}d\approx 8\sigma D_f^{\frac{1}{2}}d$$

Putting this all together, we get

$$\left\|\nabla_{\tilde{F}}\mathcal{T}(X,\Theta+\zeta)-\nabla_F\mathcal{T}(X,\Theta)\right\|_F$$
$$\lessapprox\sum_{t=1}^{T+1}\sum_{i=1}^{d+1}\left|\frac{\partial\mathcal{T}(X,\tilde{\Theta})}{\partial\tilde{G}_{t,i}}-\frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\left\|\nabla_F G_{t,i}\right\|_F+\left|\frac{\partial\mathcal{T}(X,\Theta)}{\partial G_{t,i}}\right|\left\|\nabla_{\tilde{F}}\tilde{G}_{t,i}-\nabla_F G_{t,i}\right\|_F$$
$$\leq\Big(128\sigma D_f^2\omega log(T)\Big)\Big(2\sqrt{D_f}\Big)(d+1)\omega+4\sqrt{\omega}log(T)\Big(4\sigma\sqrt{D_f}+8\sigma D_f^{\frac{1}{2}}d\Big)(d+1)$$
$$\approx 256\sigma D_f^{\frac{5}{2}}\omega^2d+32\sigma D_f^{\frac{1}{2}}\omega^2d^2$$

M.9 Final Result for Perturbed Gradient Norm Bounds

Putting this all together, we have that

$$||\nabla_{\tilde{\Theta}}\mathcal{T}(X,\tilde{\Theta})-\nabla_{\Theta}\mathcal{T}(X,\Theta)||$$
$$\lessapprox\underbrace{128\sqrt{2}\sigma D_f^2\sqrt{\omega}log(T)}_{||\nabla_{\tilde{V}^{(2)}}-\nabla_{V^{(2)}}||}+\underbrace{192\sigma\sqrt{\omega}D_f^2}_{||\nabla_{\tilde{W}^{(2)}}-\nabla_{W^{(2)}}||}+\underbrace{128\sqrt{2}D_f^2\omega^2log(T)}_{||\nabla_{\tilde{W}^{(1)}}-\nabla_{W^{(1)}}||}+\underbrace{64\sqrt{2}dD_f^2\omega^2log(T)}_{||\nabla_{\tilde{V}^{(1)}}-\nabla_{V^{(1)}}||}$$
$$+\underbrace{128\sigma d\omega^2log(T)D_f^{\frac{3}{2}}}_{||\nabla_{\tilde{M}}-\nabla_M||}+\underbrace{1024\sigma D_f^{\frac{7}{2}}\omega^{\frac{3}{2}}log(T)+8\sigma D_f^{\frac{1}{2}}\sqrt{\omega}log(T)d}_{||\nabla_{\tilde{\Gamma}}-\nabla_{\Gamma}||}+\underbrace{256\sigma D_f^{\frac{5}{2}}\omega^2d+32\sigma D_f^{\frac{1}{2}}\omega^2d^2}_{||\nabla_{\tilde{F}}-\nabla_F||}$$
$$\in o(\sigma D_f^{\frac{7}{2}}\omega^{\frac{3}{2}}log(T)^2)$$

□

N Perturbed Hessian Norm Bounds

Theorem 29. *Given our transformer construction, a set of $\sigma-$ normal perturbations ϵ, and a corresponding set of remainder perturbations ζ (as defined in **??**), the oeprator norm of the perturbation to the hessian Hessian obeys:*

$$||\nabla^2_\Theta \mathcal{T}(X, \Theta + \zeta) - \nabla^2_\Theta \mathcal{T}(X, \Theta)|| \lessapprox H_p(\sigma, \omega, D_f, T, d)$$

where

$$H_p(\sigma, \omega, D_f, T) \in o(\sigma D_f^{\frac{7}{2}} \omega^{\frac{5}{2}} log(T)^{\frac{7}{2}})$$

Combined with our bound on the operator norm of the exact hessian in 17, it follows that

$$\begin{aligned} ||\nabla^2_\Theta \mathcal{T}(X, \Theta + \zeta)|| &\leq ||\nabla^2_\Theta \mathcal{T}(X, \Theta)|| + ||\nabla^2_\Theta \mathcal{T}(X, \Theta + \zeta) - \nabla^2_\Theta \mathcal{T}(X, \Theta)|| \\ &\lessapprox H_u(\omega, D_f, T, d) + H_p(\sigma, \omega, D_f, T, d) \end{aligned}$$

Proof. We calculate the perturbations in the Hessian norm by calculating the perturbation in each of the 6 non-zero terms contributing to the Hessian norm, as presented in the previous subsection. We illustrate the overall approach with an example. Suppose we want to upper bound the perturbation in the hessian norm contributed by the perturbed second derivative $\nabla^2_{\tilde{M}\tilde{V}^{(2)}} \mathcal{T}(X, \Theta)$ The perturbation in the hessian is given by

$$\begin{aligned} &\nabla_{\tilde{M}} \Big(\big[\nabla_{\tilde{V}^{(2)}} \mathcal{T}(X, \Theta) \big]_i \Big) - \nabla_M \Big(\big[\nabla_{V^{(2)}} \mathcal{T}(X, \Theta) \big]_i \Big) \\ &= \sum_{t=1}^{T+1} \sum_{j=1}^{d+1} \frac{\Big[\nabla_{\tilde{V}^{(2)}} \mathcal{T}(X, \tilde{\Theta}) \Big]_i}{\partial \tilde{G}_{t,j}} \nabla_{\tilde{M}} \tilde{G}_{t,i} - \frac{\Big[\nabla_{V^{(2)}} \mathcal{T}(X, \Theta) \Big]_i}{\partial G_{t,j}} \nabla_M G_{t,i} \end{aligned}$$

Recall that the i^{th} row of $\nabla^2_{\tilde{M}\tilde{V}^{(2)}} \mathcal{T}(X, \tilde{\Theta}) - \nabla^2_{MV^{(2)}} \mathcal{T}(X, \Theta)$ is equal to

$$Vec\left(\nabla_{\tilde{M}} \Big(\big[\nabla_{\tilde{V}^{(2)}} \mathcal{T}(X, \Theta) \big]_i \Big) - \nabla_M \Big(\big[\nabla_{V^{(2)}} \mathcal{T}(X, \Theta) \big]_i \Big) \right)$$

for a fixed i. So the contribution to the overall operator norm of the perturbation in $\nabla^2_{\tilde{M}\tilde{V}^{(2)}} \mathcal{T}(X, \tilde{\Theta})$ is given by

$$\begin{aligned} &2 \max_{||v||=1} concat(s_1, ...s_{d+1})^T \Big(\nabla^2_{W^{(2)}V^{(2)}} \mathcal{T}(X, \Theta) - \nabla^2_{\tilde{W}^{(2)}\tilde{V}^{(2)}} \mathcal{T}(X, \tilde{\Theta}) \Big) concat(p_1, ..., p_{d+1}) \\ &\leq 2 \sum_{i=1}^{d+1} |s_i| \Big\| Vec\Big(\nabla_{\tilde{M}} \Big(\big[\nabla_{\tilde{V}^{(2)}} \mathcal{T}(X, \Theta) \big]_i \Big) - \nabla_M \Big(\big[\nabla_{V^{(2)}} \mathcal{T}(X, \Theta) \big]_i \Big) \Big) \Big\| ||\mathbf{p}|| \\ &\leq 2 \sum_{i=1}^{d+1} \sum_{k=1}^{d+1} |s_i| \Big\| e_k^T \nabla_{\tilde{M}} \Big(\big[\nabla_{\tilde{V}^{(2)}} \mathcal{T}(X, \Theta) \big]_i \Big) - e_k^T \nabla_M \Big(\big[\nabla_{V^{(2)}} \mathcal{T}(X, \Theta) \big]_i \Big) \Big\| ||\mathbf{p_k}|| \end{aligned}$$

Note that the s_i and p_j here are different from the s_i and p_i involved in calculating the unperturbed hessian, and even different than the singular vectors involved in the calculation of the overall hessian perturbation (as opposed to the operator norm of a sub-matrix within the overall Hessian). However, since the only thing we care about is that these vectors and all projections onto different subspaces of coordinates) have a 2-norm that is upper bounded by 1, we use the same notation. We refer to the contribution of the perturbation of a given sub-matrix to the operator of the overall Hessian perturbation as Δ_{Term_i}, where $Term_i$ is the term corresponding to the given sub-matrix in 17.

Ultimately, we will use the triangle inequality to conclude that

$$\begin{aligned} \Big\| \nabla^2_{\tilde{\Theta}} \mathcal{T}(X, \tilde{\Theta}) \Big\| &\leq \Big\| \nabla^2_\Theta \mathcal{T}(X, \Theta) \Big\| + \Big\| \nabla^2_{\tilde{\Theta}} \mathcal{T}(X, \tilde{\Theta}) - \nabla^2_\Theta \mathcal{T}(X, \Theta) \Big\| \\ &\leq \Big\| \nabla^2_\Theta \mathcal{T}(X, \Theta) \Big\| + \sum_{i=1}^{6} \Delta_{Term_i} \end{aligned}$$

N.1 Δ_{Term_1}

Recall from 17 that

$$Term_1 = 2 \max_{||v||=1} concat(s_1, ...s_{d+1})^T \Big(\nabla^2_{W^{(2)}V^{(2)}} \mathcal{T}(X, \Theta)\Big) concat(p_1, ..., p_{d+1}).$$

Recall our formula for $\Big[\nabla_{V^{(2)}} \mathcal{T}(X, \Theta)\Big]_i = (G_{:,i})^T \phi^{(2)}_{T+1}$ from 17. Whereas in the exact hessian case, $G_{:,i}$ is independent of all parameter matrices (since the first d columns of G, like X, encode purely positional information), in the perturbed case $G_{:,i}$ may in fact depend on the perturbations of some of the parameter matrices. To handle this more general case, we will use the product rule:

$$\nabla_{\tilde{W}^{(2)}} \Big(\Big[\nabla_{\tilde{V}^{(2)}} \mathcal{T}(X, \tilde{\Theta})\Big]_i\Big) = \nabla_{\tilde{W}^{(2)}} \Big((\tilde{G}_{:,i})^T \tilde{\phi}^{(2)}_{T+1}\Big) = \sum_{t=1}^{T+1} \nabla_{\tilde{W}^{(2)}} \big(\big[\tilde{\phi}^{(2)}_{T+1}\big]_t\big) \tilde{G}_{t,i} + \nabla_{\tilde{W}^{(2)}} \big(\tilde{G}_{t,i}\big) \big[\tilde{\phi}^{(2)}_{T+1}\big]_t$$

In this case, the perturbed MLP output $\tilde{G}_{:,i}$ does not depend on the (downstream) parameters $\tilde{W}^{(2)}$, and $\nabla_{\tilde{W}^{(2)}} \big(\tilde{G}_{:,i}\big) \tilde{\phi}^{(2)}_{T+1} = 0$. It remains to calculate $\nabla_{\tilde{W}^{(2)}} \big(\tilde{\phi}^{(2)}_{T+1}\big) \tilde{G}_{:,i}$. Recall from 17 (specifically, in the derivation of $\nabla^2_{W^{(2)}V^{(2)}} \mathcal{T}(X, \Theta)$) that $\nabla_{W^{(2)}} [\phi^{(2)}_{T+1}]_t = g_{T+1} e_t^T \phi'^{(2)}_{T+1} G$, and thus we can write the perturbed Hessian as:

$$\nabla_{\tilde{W}^{(2)}} \Big(\Big[\nabla_{\tilde{V}^{(2)}} \mathcal{T}(X, \tilde{\Theta})\Big]_i\Big) = \sum_{t=1}^{T+1} \nabla_{\tilde{W}^{(2)}} \big(\big[\tilde{\phi}^{(2)}_{T+1}\big]_t\big) \tilde{G}_{t,i} = \sum_{t=1}^{T+1} \tilde{G}_{t,i} \tilde{g}_{T+1} e_t^T \tilde{\phi}'^{(2)}_{T+1} \tilde{G}$$

$$\nabla_{\tilde{W}^{(2)}} \Big(\Big[\nabla_{\tilde{V}^{(2)}} \mathcal{T}(X, \tilde{\Theta})\Big]_i\Big) - \nabla_{W^{(2)}} \Big(\Big[\nabla_{V^{(2)}} \mathcal{T}(X, \Theta)\Big]_i\Big) = \sum_{t=1}^{T+1} \tilde{G}_{t,i} \tilde{g}_{T+1} e_t^T \tilde{\phi}'^{(2)}_{T+1} \tilde{G} - G_{t,i} g_{T+1} e_t^T \phi'^{(2)}_{T+1} G$$

$$\begin{aligned}
&\approx \sum_{t=1}^{T+1} (\tilde{G}_{t,i} \tilde{g}_{T+1} - G_{t,i} g_{T+1}) e_t^T \phi'^{(2)}_{T+1} G \\
&+ \sum_{t=1}^{T+1} G_{t,i} g_{T+1} e_t^T \Big(\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1}\Big) G \\
&+ \sum_{t=1}^{T+1} G_{t,i} g_{T+1} e_t^T \phi'^{(2)}_{T+1} (\tilde{G} - G)
\end{aligned}$$

We can decompose the contribution to the perturbation of the overall hessian using the triangle inequality:

$$\Delta_{Term_1} \leq \underbrace{2 \sum_{i=1}^{d+1} \sum_{t=1}^{\omega} \sum_{k=1}^{d+1} |s_i| \Big|G_{t,i} G_{T+1,k}\Big| \Big\| e_t^T \phi'^{(2)}_{T+1} (\tilde{G} - G) \Big\| ||\mathbf{p_k}||}_{Term_I}$$

$$+ \underbrace{2 \sum_{i=1}^{d+1} \sum_{t=1}^{\omega} \sum_{k=1}^{d+1} |s_i| \Big|G_{t,i} G_{T+1,k}\Big| \Big\| e_t^T (\tilde{\phi}'^{(2)}_{T+1} - \phi'^{(2)}_{T+1}) G \Big\| ||\mathbf{p_k}||}_{TermII}$$

$$+ \underbrace{2 \sum_{i=1}^{d+1} \sum_{t=1}^{\omega} \sum_{k=1}^{d+1} |s_i| \Big|\tilde{G}_{t,d+1} \tilde{G}_{T+1,d+1} - G_{t,d+1} G_{T+1,d+1}\Big| \Big\| e_t^T \phi'^{(2)}_{T+1} G \Big\| ||\mathbf{p_k}||}_{Term_{III}}$$

First, we bound $Term_I$. By 16, we have $\Big\| e_t^T \phi'^{(2)}_{T+1} (\tilde{G} - G) \Big\| \leq 2 ||\tilde{G} - G||_{2,\infty} \lessapprox 8\sigma d$. Thus

$$Term_I \lessapprox 8\sigma d ||G||^2_{1,\infty} \sum_{i=1}^{d+1} \sum_{t=1}^{\omega} \sum_{k=1}^{d+1} |s_i| ||\mathbf{p_k}||$$

$$\lessapprox 8\sigma d^2\omega$$

For $Term_{II}$, we note that $\Big\|e_t^T(\tilde{\phi}_{T+1}^{\prime(2)} - \phi_{T+1}^{\prime(2)})G\Big\| \leq \Big\|G^T(\tilde{\phi}_{T+1}^{\prime(2)} - \phi_{T+1}^{\prime(2)})e_t\Big\| \leq ||G^T||_{1,2}\Big\|(\tilde{\phi}_{T+1}^{\prime(2)} - \phi_{T+1}^{\prime(2)})e_t\Big\|_1 = ||G||_{2,\infty}\Big\|(\tilde{\phi}_{T+1}^{\prime(2)} - \phi_{T+1}^{\prime(2)})e_t\Big\|_1$

$$\lessapprox 384\sqrt{2}\sigma D_F^2\sqrt{\omega}log(T)$$

It follows that

$$Term_{II} \lessapprox 768\sqrt{2}\sigma D_f^2\omega^{\frac{3}{2}}dlog(T)$$

Finally, for $Term_{III}$, note that

$$\Big|\tilde{G}_{t,d+1}\tilde{G}_{T+1,d+1} - G_{t,d+1}G_{T+1,d+1}\Big|$$

$$\lessapprox \Big|G_{t,d+1}\big(\tilde{G}_{T+1,d+1} - G_{T+1,d+1}\big)\Big| + \Big|\big(\tilde{G}_{t,d+1} - G_{t,d+1}\big)G_{T+1,d+1}\Big|$$

$$\leq 2||\tilde{G} - G||_{2,\infty} \leq 64\sigma D_f^2$$

Therefore

$$Term_{III} = 2\sum_{i=1}^{d+1}\sum_{t=1}^{\omega}\sum_{k=1}^{d+1}|s_i|\Big|\tilde{G}_{t,d+1}\tilde{G}_{T+1,d+1} - G_{t,d+1}G_{T+1,d+1}\Big|\Big\|e_t^T\phi_{T+1}^{\prime(2)}G\Big\|||\mathbf{p_k}||$$

$$\leq 128\sqrt{2}\sigma d\omega D_f^2$$

Putting this all together, we have

$$\Delta_{Term_1} \leq 8\sigma d^2\omega + 768\sqrt{2}\sigma D_f^2\omega^{\frac{3}{2}}dlog(T) + 128\sqrt{2}\sigma d\omega D_f^2$$

$$\approx 768\sqrt{2}\sigma D_f^2\omega^{\frac{3}{2}}dlog(T)$$

N.2 Δ_{Term_2}

$$Term_2 = 2\max_{||v||=1} concat(s_1, ..., s_{d+1})^T\Big(\nabla^2_{MV^{(2)}}\mathcal{T}(X,\Theta)\Big)concat(\eta_1, ...\eta_{d+1})$$

Recall that

$$\Big[\nabla_{\tilde{V}^{(2)}}\mathcal{T}(X,\tilde{\Theta})\Big]_i = e_i^T\tilde{G}^T\tilde{\phi}_{T+1}^{(2)}$$

And note that this expression is very similar to $\mathcal{T}(X,\tilde{\Theta}) = (V^{(2)})^T\tilde{G}^T\tilde{\phi}_{T+1}^{(2)}$, except that the $V^{(2)}$ has been replaced by e_i^T. We can therefore follow the derivation of $\nabla_G\mathcal{T}(X,\Theta)$ in 15 to conclude that

$$\frac{\partial\Big(e_i^T\tilde{G}^T\tilde{\phi}_{T+1}^{(2)}\Big)}{\partial\tilde{G}_{t,j}} = e_t^T\tilde{\phi}_{T+1}^{(2)}e_i^Te_j + e_t^Te_{T+1}e_i^T\tilde{G}^T\tilde{\phi}_{T+1}^{\prime(2)}\tilde{G}(\tilde{W}^{(2)})^Te_j$$

$$= e_t^T\tilde{\phi}_{T+1}^{(2)}I[i=j] + e_i^T\tilde{G}^T\tilde{\phi}_{T+1}^{\prime(2)}\tilde{G}(\tilde{W}^{(2)})^Te_jI[t=T+1]$$

We have from 15 that

$$\nabla_M G_{t,i} = b_te_i^TF^Tdiag(I\big[M^Tb_t + \Gamma > 0\big])$$

Therefore

$$\nabla_{\tilde{M}}\Big(\Big[\nabla_{\tilde{V}^{(2)}}\mathcal{T}(X,\tilde{\Theta})\Big]_i\Big) = \nabla_{\tilde{M}}\Big(e_i^T\tilde{G}^T\tilde{\phi}_{T+1}^{(2)}\Big)$$

$$= \sum_{t=1}^{T+1}\sum_{j=1}^{d+1}\frac{\partial\Big(e_i^T\tilde{G}^T\tilde{\phi}_{T+1}^{(2)}\Big)}{\partial\tilde{G}_{t,j}}\nabla_{\tilde{M}}\tilde{G}_{t,i}$$

We can decompose the perturbation as:

$$\nabla_{\tilde{M}}\Big(\Big[\nabla_{\tilde{V}^{(2)}}\mathcal{T}(X,\tilde{\Theta})\Big]_i\Big) - \nabla_M\Big(\Big[\nabla_{V^{(2)}}\mathcal{T}(X,\Theta)\Big]_i\Big)$$

$$= \sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}\Big)}{\partial \tilde{G}_{t,j}} \nabla_{\tilde{M}} \tilde{G}_{t,i} - \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}} \nabla_M G_{t,i}$$

$$\approx \sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \Big(\frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}\Big)}{\partial \tilde{G}_{t,j}} - \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}} \Big) \nabla_M G_{t,i} + \sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}} \Big(\nabla_{\tilde{M}} \tilde{G}_{t,i} - \nabla_M G_{t,i} \Big)$$

Thus

$$\Delta_{Term_2} = \max_{||v||=1} \sum_{i=1}^{d+1} s_i Vec\Big(\sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}\Big)}{\partial \tilde{G}_{t,j}} \nabla_{\tilde{M}} \tilde{G}_{t,i} - \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}} \nabla_M G_{t,i} \Big) \eta$$

$$\leq \sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \left\| \frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}\Big)}{\partial \tilde{G}_{t,j}} \nabla_{\tilde{M}} \tilde{G}_{t,i} - \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}} \nabla_M G_{t,i} \right\| ||\eta_{\mathbf{k}}||$$

$$\leq \underbrace{\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \left| \frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}\Big)}{\partial \tilde{G}_{t,j}} - \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}} \right| \Big\| \nabla_M G_{t,i} \Big\| ||\eta_{\mathbf{k}}||}_{Term_A}$$

$$+ \underbrace{\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \left| \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}} \right| \Big\| \nabla_{\tilde{M}} \tilde{G}_{t,i} - \nabla_M G_{t,i} \Big\| ||\eta_{\mathbf{k}}||}_{Term_B}$$

In the above series of inequalities, we used Cauchy-Schwarz and the triangle inequality multiple times. Now, we derive the ingredients needed to evaluate the RHS above. It was shown in 15 that

$$\nabla_M G_{t,i} = b_t e_i^T F^T diag(I\big[M^T b_t + \Gamma > 0\big]),$$

and that the norm of this intra-layer gradient term can be bounded as:

$$\Big\| \nabla_M G_{t,i} \Big\| \lessapprox 8 D_f^{\frac{3}{2}} I[i = d+1 \wedge t \leq \omega]$$

In 26, it was shown that

$$\left| \frac{\partial\Big(\mathcal{T}(X, \tilde{\Theta})\Big)}{\partial \tilde{G}_{t,d+1}} - \frac{\partial\Big(\mathcal{T}(X, \Theta)\Big)}{\partial G_{t,d+1}} \right| = |\langle \nabla_{\tilde{g}_t} \mathcal{T}(X, \tilde{\Theta}) - \nabla_{g_t} \mathcal{T}(X, \Theta), e_{d+1} \rangle| \lessapprox 128 \sigma D_f^2 \omega log(T)$$

We can follow similar arguments as 26 but with $V^{(2)}$ replaced with e_i, and conclude that

$$\left| \frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}\Big)}{\partial \tilde{G}_{t,d+1}} - \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,d+1}} \right| \lessapprox 128 \sigma D_f^2 \sqrt{\omega} log(T)$$

Note that this differs from the perturbation in the gradient in 28 only by a factor of $\sqrt{\omega}$, as expected, since $V^{(2)}$ was carrying this factor. Thus

$$Term_A = \sum_{k=1}^{d+1}\sum_{i=1}^{d+1}\sum_{t=1}^{\omega} |s_i| \Big(128 \sigma D_f^2 \sqrt{\omega} log(T) \Big) \Big(8 D_f^{\frac{3}{2}} \Big) ||\eta_{\mathbf{k}}|| = 1024 \sigma (d+1) D_f^{\frac{7}{2}} \omega^{\frac{3}{2}} log(T)$$

We now move on to $Term_B$. In 28 we showed that

$$\Big\| \nabla_{\tilde{M}} \tilde{G}_{t,i} - \nabla_M G_{t,i} \Big\|_F \lessapprox 2 \sigma D_f^{\frac{1}{2}} + 32 \sigma d D_f^{\frac{3}{2}} I[i = d+1]$$

Finally, we can make a similar argument to 24, but once again replacing $V^{(2)}$ with e_i, to conclude that

$$||\nabla_{g_t} e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}|| \leq \frac{|c_t|}{\sqrt{\omega}} + 4 log(T) I[t = T+1] + 4 \omega^{\frac{1}{2}} log(T) I[t \leq \omega]$$

Summing over $t \in [T+1]$, we have

$$\sum_{t=1}^{T+1} ||\nabla_{g_t} e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}|| \leq 1 + 4log(T) + 4\omega^{\frac{3}{2}} log(T) \approx 4\omega^{\frac{3}{2}} log(T)$$

$$Term_B \leq 2\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{j=1}^{d+1} \Big(4\omega^{\frac{3}{2}} log(T)\Big)\Big(2\sigma D_f^{\frac{1}{2}} + 32\sigma d D_f^{\frac{3}{2}} I[i=d+1]\Big)||\eta_{\mathbf{k}}||$$

$$= 8\omega^{\frac{3}{2}} log(T) \sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{j=1}^{d+1} \Big(2\sigma D_f^{\frac{1}{2}} + 32\sigma d D_f^{\frac{3}{2}} I[i=d+1]\Big)||\eta_{\mathbf{k}}||$$

$$= 8\omega^{\frac{3}{2}} log(T) \sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1} \Big(2(d+1)\sigma D_f^{\frac{1}{2}} + 32\sigma d D_f^{\frac{3}{2}}\Big)||\eta_{\mathbf{k}}||$$

$$\leq 8\omega^{\frac{3}{2}} log(T)(d+1)\Big(2(d+1)\sigma D_f^{\frac{1}{2}} + 32\sigma d D_f^{\frac{3}{2}}\Big)$$

$$\approx 256\sigma\omega^{\frac{3}{2}} d^2 D_f^{\frac{3}{2}} log(T)$$

Putting $Term_A$ and $Term_B$ together, we have

$$\Delta_{Term_2} \lessapprox 1024\sigma d D_f^{\frac{7}{2}} \omega^{\frac{3}{2}} log(T) + 256\sigma\omega^{\frac{3}{2}} d^2 D_f^{\frac{3}{2}} log(T)$$

N.3 Δ_{Term_3}

$$Term_3 = 2 \max_{||v||=1} concat(s_1, ..., s_{d+1})^T \Big(\nabla^2_{\Gamma V^{(2)}} \mathcal{T}(X, \Theta)\Big)\gamma_{\mathbf{1}}$$

We follow the same approach as $\tilde{Term_2}$. We have

$$\tilde{Term_3} - Term_3 \lessapprox \underbrace{2\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \left|\frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}\Big)}{\partial \tilde{G}_{t,j}} - \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}}\right| \Big|\Big|\nabla_\Gamma G_{t,i}\Big|\Big| ||\eta_{\mathbf{k}}||}_{Term_A}$$

$$\underbrace{+ 2\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \left|\frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}}\right| \Big|\Big|\nabla_{\tilde{\Gamma}} \tilde{G}_{t,i} - \nabla_\Gamma G_{t,i}\Big|\Big| ||\eta_{\mathbf{k}}||}_{Term_B}$$

In 15, we showed that

$$||\nabla_\Gamma G_{t,i}|| \lessapprox 8 D_f^{\frac{3}{2}} I[i = d+1 \wedge t \leq \omega]$$

In addition, 28, tells us that

$$||\nabla_{\tilde{\Gamma}} \tilde{G}_{t,i} - \nabla_\Gamma G_{t,i}|| \leq 2\sigma\sqrt{D_f}$$

Combining this with our previous bounds on $\left|\frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}_{T+1}^{(2)}\Big)}{\partial \tilde{G}_{t,j}} - \frac{\partial\Big(e_i^T G^T \phi_{T+1}^{(2)}\Big)}{\partial G_{t,j}}\right| \Big|\Big|\nabla_\Gamma G_{t,i}\Big|\Big|$ earlier in this section, we have

$$Term_A \leq 2|s_{d+1}| \sum_{k=1}^{d+1}\sum_{t=1}^{\omega}\sum_{j=1}^{d+1} \Big(128\sigma D_f^2 \sqrt{\omega} log(T)\Big)\Big(8 D_f^{\frac{3}{2}}\Big)||\eta_{\mathbf{k}}||$$

$$\leq 1024\sigma D_f^{\frac{7}{2}} \omega^{\frac{3}{2}} log(T)$$

$$Term_B \leq 4\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{j=1}^{d+1} \Big(1 + 4log(T) + 4\omega^{\frac{3}{2}} log(T)\Big)\Big(2\sigma D_f^{\frac{1}{2}}\Big)||\eta_{\mathbf{k}}||$$

$$\approx 32\sigma\omega^{\frac{3}{2}} D_f^{\frac{1}{2}} d log(T)$$

Putting $Term_A$ and $Term_B$ together, we have

$$\Delta_{Term_3} \lessapprox 1024\sigma D_f^{\frac{7}{2}} \omega^{\frac{3}{2}} log(T)$$

N.4 Δ_{Term_4}

$$Term_4 = 2 \max_{||v||=1} concat(s_1, ..., s_{d+1})^T \Big(\nabla^2_{FV^{(2)}} \mathcal{T}(X, \Theta)\Big) concat(\nu_1, ...\nu_{4(D_f+1)})$$

In 15, we showed that

$$||\nabla_F G_{t,i}|| \lessapprox 2\sqrt{D_f} I[t \leq \omega]$$

In addition, 28, tells us that

$$||\nabla_{\tilde{F}} \tilde{G}_{t,i} - \nabla_F G_{t,i}|| \leq 8\sigma D_f^{\frac{1}{2}} d$$

$$\Delta_{Term} \lessapprox \underbrace{2\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \left| \frac{\partial\Big(e_i^T \tilde{G}^T \tilde{\phi}^{(2)}_{T+1}\Big)}{\partial \tilde{G}_{t,j}} - \frac{\partial\Big(e_i^T G^T \phi^{(2)}_{T+1}\Big)}{\partial G_{t,j}} \right| \Big|\Big| Vec\big(\nabla_F G_{t,i}\big) \Big|\Big| ||\eta_{\mathbf{k}}||}_{Term_A}$$

$$+ \underbrace{2\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{t=1}^{T+1}\sum_{j=1}^{d+1} \left| \frac{\partial\Big(e_i^T G^T \phi^{(2)}_{T+1}\Big)}{\partial G_{t,j}} \right| \Big|\Big| Vec\big(\nabla_{\tilde{F}} \tilde{G}_{t,i} - \nabla_F G_{t,i}\big) \Big|\Big| ||\eta_{\mathbf{k}}||}_{Term_B}$$

$$Term_A \leq 2\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{t=1}^{\omega}\sum_{j=1}^{d+1} \Big(128\sigma D_f^2 \sqrt{\omega} log(T)\Big)\big(2\sqrt{D_f}\big)||\eta_{\mathbf{k}}||$$

$$\leq 512\sigma d^2 D_f^{\frac{5}{2}} \omega^{\frac{3}{2}} log(T)$$

$$Term_B \leq 2\sum_{i=1}^{d+1} |s_i| \sum_{k=1}^{d+1}\sum_{j=1}^{d+1} \Big(1 + 4log(T) + 4\omega^{\frac{3}{2}} log(T)\Big)\Big(8\sigma D_f^{\frac{1}{2}} d\Big)||\eta_{\mathbf{k}}||$$

$$\lessapprox 64\sigma D_f^{\frac{1}{2}} log(T) d^2$$

Thus

$$\Delta_{Term_4} \lessapprox 512\sigma d^2 D_f^{\frac{5}{2}} \omega^{\frac{3}{2}} log(T) + 64\sigma\omega^{\frac{3}{2}} D_f^{\frac{1}{2}} log(T) d^2$$

$$\approx 512\sigma D_f^{\frac{5}{2}} \omega^{\frac{3}{2}} d^2 log(T)$$

N.5 Δ_{Term_5}

$$Term_5 = 2 \max_{||v||=1} \gamma_1^T \Big(\nabla^2_{F\Gamma} \mathcal{T}(X, \Theta)\Big) concat(\nu_1, ..., \nu_{4(D_f+1)})$$

$$= 2 \max_{||v||=1} concat(\nu_1, ..., \nu_{4(D_f+1)})^T \Big(\nabla^2_{\Gamma F} \mathcal{T}(X, \Theta)\Big) \gamma_1$$

Recall that

$$\Big[\nabla_F \mathcal{T}(X, \Theta)\Big]_{i,j} = \sum_{k=1}^{d+1}\sum_{t=1}^{\omega} c_t e_i^T \big(M^T b_t + \Gamma\big)_+ e_k^T e_j$$

To obtain $\nabla^2_{\tilde{F}\tilde{\Gamma}} \mathcal{T}(X, \tilde{\Theta})$ in the perturbed case, we once again modify the unperturbed hessian to account for the fact that the final derivative in the perturbed case depends on more than just the last dimension. We start with a generalization of the perturbed gradient, noting that in the unperturbed case, we had $\frac{\partial \mathcal{T}(X,\tilde{\Theta})}{\partial \tilde{G}_{t,d+1}} = c_t$.

$$\Big[\nabla_{\tilde{F}} \mathcal{T}(X, \tilde{\Theta})\Big]_{i,j} = \sum_{k=1}^{d+1}\sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} e_i^T \nabla_{\tilde{F}} \tilde{G}_{t,k} e_j$$

$$\Big[\nabla_{\tilde{F}} \mathcal{T}(X, \tilde{\Theta})\Big]_{i,j} = \sum_{k=1}^{d+1}\sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} e_i^T \big(\tilde{M}^T \tilde{b}_t + \tilde{\Gamma}\big)_+ e_k^T e_j$$

$$= \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,j}} \left([(\tilde{M}_{:,i})^T \tilde{b}_t + \tilde{\Gamma}_i] \right)_+$$

Taking the derivative with respect to $\tilde{\Gamma}$ yields

$$\nabla_{\tilde{\Gamma}} \Big(\big[\nabla_{\tilde{F}} \mathcal{T}(X, \tilde{\Theta}) \big]_{i,j} \Big) = \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,j}} e_i I\big[\tilde{M}_i^T \tilde{b}_t + \tilde{\Gamma}_i > 0\big]$$

We have used the fact that $\frac{\partial \mathcal{T}(X,\tilde{\Theta})}{\partial \tilde{G}_{t,k}}$ is determined by the portion of the transformer downstream of the MLP; namely, the second attention layer. Therefore it is independent of the MLP and constant relative to Γ, and can be taken outside of the derivative. We then upper bound the difference as

$$\Delta_{Term_5} \leq$$

$$\sum_{i=1}^{4(D_f+1)} \sum_{j=1}^{d+1} \sum_{t=1}^{\omega} |\nu_{ij}| \left| Vec\Big(\frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,j}} e_i I\big[\tilde{M}_i^T \tilde{b}_t + \tilde{\Gamma}_i > 0\big] - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,j}} e_i I\big[M_i^T b_t + \Gamma_i > 0\big] \Big) \right| |\gamma_1|$$

$$= \sum_{i=1}^{4(D_f+1)} \sum_{j=1}^{d+1} \sum_{t=1}^{\omega} \sum_{k=1}^{d+1} |\nu_{ij}| I[i=k] \left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,j}} I\big[\tilde{M}_i^T \tilde{b}_t + \tilde{\Gamma}_i > 0\big] - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,j}} I\big[M_i^T b_t + \Gamma_i > 0\big] \right| |\gamma_{1k}|$$

$$= \sum_{i=1}^{4(D_f+1)} \sum_{j=1}^{d+1} \sum_{t=1}^{\omega} |\nu_{ij}| \left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,j}} I\big[\tilde{M}_i^T \tilde{b}_t + \tilde{\Gamma}_i > 0\big] - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,j}} I\big[M_i^T b_t + \Gamma_i > 0\big] \right| |\gamma_{1i}|$$

$$\lessapprox \underbrace{\sum_{i=1}^{4(D_f+1)} \sum_{j=1}^{d+1} \sum_{t=1}^{\omega} |\nu_{ij}| \left| \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,j}} \Big(I\big[\tilde{M}_i^T \tilde{b}_t + \tilde{\Gamma}_i > 0\big] - I\big[M_i^T b_t + \Gamma_i > 0\big] \Big) \right| |\gamma_{1i}|}_{Term_A}$$

$$+ \underbrace{\sum_{i=1}^{4(D_f+1)} \sum_{j=1}^{d+1} \sum_{t=1}^{\omega} |\nu_{ij}| \left| \Big(\frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,j}} - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,j}} \Big) I\big[M_i^T b_t + \Gamma_i > 0\big] \right| |\gamma_{1i}|}_{Term_B}$$

Where the last inequality follows from the triangle inequality, and the fact that we are approximating the perturbation using only terms that are first-order in σ. Once again, we use the fact that with high probability, none of the neurons in the MLP will flip with high probability to conclude that $Term_A = 0$. Clearly, $\left| I\big[M^T b_t + \Gamma \geq 0\big]_i \right| \leq 1$, and thus

$$Term_B \leq \sum_{i=1}^{4(D_f+1)} \sum_{j=1}^{d+1} \sum_{t=1}^{\omega} |\nu_{ij}| \left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,j}} - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,j}} \right| |\gamma_{1i}|$$

In 26 we show that

$$\left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,i}} - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,i}} \right| \lessapprox 128 \sigma D_f^2 \omega log(T) I[i = d+1] + 768 \sigma D_f^2 \omega^{\frac{3}{2}} log^2(T) I[i < d+1].$$

Therefore,

$$\Delta_{Term_5} = Term_B \leq$$

$$\sum_{i=1}^{4(D_f+1)} \sum_{t=1}^{\omega} |\nu_{i,d+1}| \left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,d+1}} - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,d+1}} \right| |\gamma_{1i}| + \sum_{i=1}^{4(D_f+1)} \sum_{j=1}^{d} \sum_{t=1}^{\omega} |\nu_{ij}| \left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,j}} - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,j}} \right| |\gamma_{1i}|$$

$$256 \sigma D_f^{\frac{7}{2}} \omega^2 log(T) + 1536 \sigma D_f^{\frac{7}{2}} \omega^{\frac{5}{2}} d^{\frac{1}{2}} log^2(T)$$

$$\approx 1536 \sigma D_f^{\frac{7}{2}} \omega^{\frac{5}{2}} d^{\frac{1}{2}} log^2(T)$$

N.6 Δ_{Term_6}

Recall that

$$Term_6 = 2 \max_{||v||=1} concat(\eta_1, \ldots, \eta_{d+1})^T \Big(\nabla^2_{FM} \mathcal{T}(X, \Theta)\Big) concat(\nu_1, ..., \nu_{4(D_f+1)})$$

We once again re-derive the hessian under the assumption that the output depends on all dimensions of g_t. By the chain rule, we have the following for the perturbed gradient with respect to M :

$$\nabla_{\tilde{M}} \mathcal{T}(X, \tilde{\Theta}) = \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} \tilde{b}_t e_k^T \tilde{F}^T diag(I\big[\tilde{M}^T \tilde{b}_t + \tilde{\Gamma} > 0\big])$$

$$\implies \Big[\nabla_{\tilde{M}} \mathcal{T}(X, \tilde{\Theta})\Big]_{ij} = \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} e_i^T \tilde{b}_t e_k^T \tilde{F}^T diag(I\big[\tilde{M}^T \tilde{b}_t + \tilde{\Gamma} > 0\big]) e_j$$

where $e_i, e_k \in \mathbb{R}^{d+1}$, $e_j \in \mathbb{R}^{4(D_f+1)}$. The partial derivative $\frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}}$ is independent of F, and therefore the gradient with resect to F is given by:

$$\nabla_{\tilde{F}} \Big(\Big[\nabla_{\tilde{M}} \mathcal{T}(X, \tilde{\Theta})\Big]_{ij} \Big) = \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} \nabla_{\tilde{F}} \Big(e_i^T \tilde{b}_t e_k^T \tilde{F}^T diag(I\big[\tilde{M}^T \tilde{b}_t + \tilde{\Gamma} > 0\big]) e_j \Big)$$

We again leverage the rule of matrix calculus that says $\frac{d}{dM}\Big(x^T M y\Big) = xy^T$, we have

$$\nabla_{\tilde{F}} \Big(\Big[\nabla_{\tilde{M}} \mathcal{T}(X, \tilde{\Theta})\Big]_{ij} \Big) = \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} e_i^T \tilde{b}_t e_k e_j^T diag(I\big[\tilde{M}^T \tilde{b}_t + \tilde{\Gamma} > 0\big])$$

We can decompose this perturbation as follows:

$$\nabla_{\tilde{F}} \Big(\Big[\nabla_{\tilde{M}} \mathcal{T}(X, \tilde{\Theta})\Big]_{ij} \Big) - \nabla_F \Big(\Big[\nabla_M \mathcal{T}(X, \Theta)\Big]_{ij} \Big)$$

$$\lessapprox \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} \Big(\frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} - \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} \Big) e_i^T b_t e_k e_j^T diag(I\big[M^T b_t + \Gamma > 0\big])$$

$$+ \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} e_i^T (\tilde{b}_t - b_t) e_k e_j^T diag(I\big[M^T b_t + \Gamma > 0\big])$$

$$= \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} \Big(e_i^T b_t e_k e_j^T diag(I\big[\tilde{M}^T \tilde{b}_t + \tilde{\Gamma} > 0\big]) - diag(I\big[M^T b_t + \Gamma > 0\big]) \Big)$$

Note that once again we argue that for σ small in the sense described in 21, the final sum will vanish. In the first sum, note that since the first d dimensions of the unperturbed attention outputs b_t are all 0, the summand vanishes when $i \neq d+1$. To bound the norm of the perturbation, we apply the triangle inequality:

$$\Delta_{Term_6} \lessapprox \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} |\eta_{d+1,j}| \left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} - \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} \right| ||\nu_{\mathbf{k}}||$$

$$+ \sum_{i=1}^{d+1} \sum_{j=1}^{d+1} \sum_{k=1}^{d+1} \sum_{t=1}^{\omega} |\eta_{\mathbf{ij}}| \left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,k}} \right| \left\| \tilde{b}_t - b_t \right\| ||\nu_{\mathbf{k}}||$$

We once again use 26:

$$\left| \frac{\partial \mathcal{T}(X, \tilde{\Theta})}{\partial \tilde{G}_{t,i}} - \frac{\partial \mathcal{T}(X, \Theta)}{\partial G_{t,i}} \right| \lessapprox 128 \sigma D_f^2 \omega log(T) I[i = d+1] + 768 \sigma D_f^2 \omega^{\frac{3}{2}} log^2(T) I[i < d+1].$$

we also leverage our bound from 24 that says that

$$||\nabla_{g_t} \mathcal{T}(X, \Theta)|| \leq |c_t| + 4\sqrt{\omega} log(T) I[t = T+1] + 4\omega log(T) I[t \in [\omega]]$$

Plugging these bounds in, we get

$$\Delta_{Term_6} \lessapprox 768\sigma D_f^2\omega^{\frac{3}{2}}log^2(T)\sum_{j=1}^{d+1}\sum_{k=1}^{d}\sum_{t=1}^{\omega}|\eta_{d+1,j}|||\nu_{\mathbf{k}}||+128\sigma D_f^2\omega log(T)\sum_{j=1}^{d+1}\sum_{t=1}^{\omega}|\eta_{d+1,j}|||\nu_{\mathbf{d+1}}||$$

$$+4\omega^2 log(T)\sum_{i=1}^{d+1}\sum_{j=1}^{d+1}\sum_{k=1}^{d+1}|\eta_{\mathbf{ij}}|\left\|\tilde{b}_t - b_t\right\|||\nu_{\mathbf{k}}||$$

$$\leq 768\sigma D_f^2\omega^{\frac{5}{2}}dlog^2(T) + 128\sigma D_f^2\omega^2 d^{\frac{1}{2}}log(T) + 16\sigma\omega^2 d^{\frac{5}{2}}log(T)$$

$$\approx 768\sigma D_f^2\omega^{\frac{5}{2}}dlog^2(T) + 16\sigma\omega^2 d^{\frac{5}{2}}log(T)$$

N.7 Final Result for Perturbed Hessian Norm Bounds

Putting these together, we have

$$\underbrace{768\sqrt{2}\sigma D_F^2\omega^{\frac{3}{2}}dlog(T)}_{\Delta_{Term_1}} + \underbrace{256\sigma\omega^{\frac{3}{2}}d^2D_f^{\frac{3}{2}}log(T)}_{\Delta_{Term_2}} + \underbrace{1024\sigma D_f^{\frac{7}{2}}\omega^{\frac{3}{2}}log(T)}_{\Delta_{Term_3}}$$

$$+\underbrace{512\sigma D_f^{\frac{5}{2}}\omega^{\frac{3}{2}}d^2log(T)}_{\Delta_{Term_4}} + \underbrace{1536\sigma D_f^{\frac{7}{2}}\omega^{\frac{5}{2}}d^{\frac{1}{2}}log^2(T)}_{\Delta_{Term_5}} + \underbrace{768\sigma D_f^2\omega^{\frac{5}{2}}dlog^2(T) + 16\sigma\omega^2 d^{\frac{5}{2}}log(T)}_{\Delta_{Term_6}}$$

$$\in o(\sigma D_f^{\frac{7}{2}}\omega^{\frac{5}{2}}log(T)^{\frac{7}{2}})$$

□

O Chain Of Thought

Theorem 30. *Let $T \in \mathbb{N}$. Let $f_1, \ldots, f_T$ where $f_i(x_1 \ldots x_T f_1 \ldots f_{i-1}) = \bigoplus_{j\leq i} x_j$. Let $\widehat{\Theta_{CoT}}(\alpha, \beta)$ be the solution returned by the general learning procedure on the training set of size m jointly consisting of these functions using regularization parameters α, β, and let $\delta_{CoT}(\alpha, \beta)$ be the expected final error for a chain of T auto-regressive steps using the transformer parametrized by $\widehat{\Theta_{CoT}}$ to solve the Parity Task of length T. Then there exist regularization parameters $\alpha_{CoT}, \beta_{CoT}$ such that*

$$\delta_{CoT}(\alpha_{CoT}, \beta_{CoT}) \leq B_{CoT}(\sigma_1),$$

where

$$B_{CoT}(\sigma_1) := Te^{-\frac{m}{8\Sigma^2}}e^{\frac{m\sigma_1^2}{4\Sigma^2}\left(2G_u(1,2)+P(\sigma_1,1,2,T)\right)^2+\frac{L(1,2,T)}{2\sigma_1^2}},$$

and σ_1 is the σ that minimizes the generalization bound.

In contrast, Let $\widehat{\Theta_{OP}}(\alpha, \beta)$ be the solution returned by the general learning procedure on the training set of size m consisting of only the inputs and outputs of the complete Parity task of length T, i.e. the functions $f(x_1 \ldots x_T) = \bigoplus_{j\leq T} x_j$). Let $\delta_{OP}(\alpha, \beta)$ be the expected error for a single pass of a transformer parameterized by $\widehat{\Theta_{OP}}(\alpha, \beta)$ on the parity task of length T. Then there exist regularization parameters α_{OP}, β_{OP}) such that

$$\delta_{OP} \leq B_{OP}(\sigma_2),$$

where

$$B_{OP}(\sigma_2) := e^{-\frac{m}{8\Sigma_2^2}}e^{\frac{m\sigma^2}{4\Sigma^2}\left(2G_u(1,T)+P(\sigma_2,1,T,T)\right)+\frac{L(1,T,T)}{2\sigma_2^2}}$$

It follows that

$$B_{OP} \geq \frac{B_{CoT}^T e^{\frac{m}{8\Sigma^2}(T-1)}}{T}$$

In other words, our bound on the error for Parity increases exponentially with length when using the one-pass approach, whereas using the CoT approach, the error increases only linearly with T.

Proof. First, we show the existence of a construction which recognizes PARITY perfectly using a transformer that has learned a degree-2 boolean function, and therefore has low sharpness and parameter norms.

Let $\mathcal{T}_{\mathcal{C}\prime\mathcal{T}}$ be a transformer with context length $2T$ matching our main construction, except that the positional encodings (and their dimension-reduced random projections) are cyclic with period T. For example, positions 1 and $T+1$ have the same positional encoding. Note that the (pre-random-projection) input vectors $y_t \in \mathbb{R}^{T+2}$, and $\bar{W}^{(1)} \in \mathbb{R}^{(T+2)\times(T+2)}$ still, despite there being a longer context window now. Like our main construction, the matrix $\bar{W}^{(1)}$ has all zeros in the final row and column, so that attention patterns are purely position-aware. The pre-random-projection matrix $\bar{W}$ however looks like one with just a single component of degree 1, containing only the current position.

We begin the CoT by passing the model the input of length T along with a CLS token, which has the $d+1^{th}$ dimension for bit values set to 0 as described previously in our main construction. Thus our input on the first step is $x_1 \dots x_T CLS$. Being that the $T+1^{th}$ and 1^{st} positions are the only ones with this positional encoding, the positional-identity map in $W^{(1)}$ creates an attention pattern for position $T+1$ which is uniform across those two positions, just as if we had a component of degree 2 across those positions when using $T+1$ absolute positional encodings (in fact, we will use this simulation argument later in the proof). from there, the transformer is exactly identical to our construction, using $D_f = 2$ and $\omega = 1$. We read the final scalar output from the $T+1^{th}$ position, and if the output is closer to 0 than it is to 1, output 0; otherwise output 1. We then write this output to position $T+2$, and then repeat the process, considering the new current position to be $T+2$. In other words, the input to step 2 of the COT is $x_1 \dots x_T CLS f_1$, where $f_1 = 0 \bigoplus x_1 = x_1$. We then use our transformer to calculate $f_2 = f_1 \bigoplus x_2 = x_1 \bigoplus x_2$, and so on. Applying this autoregressively, we will have

$$f_i(x_1 \dots x_T CLS f_1 \dots f_{i-1}) = \bigoplus_{j \le i} x_j$$

We have shown that our transformer correctly outputs the parity of the input string using COT. it remains to show that our bounds on the norm and sharpness for our construction for $D_f = 2, \omega = 1$ still hold in this slightly modified construction. Note that our matrix $\bar{W}^{(1)}$ actually resembles our construction with $D_f = 1$ in the sense relevant to the parameter norm: the first column of $\bar{W}^{(1)}$ has only a single nonzero entry. Thus the norm of the CoT construction will actually be smaller than that of our main construction using $D_f = 2$. As for the sharpness, we might expect this modification to the $W^{(1)}$ matrix to cause the perturbed gradients and Hessian to differ from the fixed-positional-encoding, degree-2 construction. However, we note that all derivatives in our construction depend on $W^{(1)}$ only indirectly, through the weights of the attention matrix, ϕ_t. Since we can simulate the exact attention patterns of our transformer at timestep i with a transformer with $T+i+1$ fixed positional encodings and our usual degree-2 construction, these two constructions will have the exact same Hessian and gradients.

Therefore, there are regularization parameters $\alpha_{CoT}, \beta_{CoT} \ge 0$ such that the general learning procedure returns a solution $\hat{\Theta}_S$ with the following properties:

$$||\hat{\Theta}_S|| \le L(1,2,T)$$
$$Tr\Big(\nabla^2\big[\hat{L}(f_{\Theta_S+\zeta})\big]\Big) \le 2G_u(1,2) + P(\sigma)$$

where $L(\omega, D_f, T) \in O\Big(D_f^3 + log(T)^2\omega^2 D_f\Big)$ is an upper bound on the parameter norm, and $G_u(\omega, D_f) + P(\sigma) \in O(\omega D_f^3) + o\Big(\sigma D_f^5 \omega^2 \big(log^{\frac{7}{2}}(T) + D_f log^{\frac{5}{2}}(T)\big)\Big)$ is an upper bound on the operator norm of the loss Hessian. According to our generalization bound, each time step has a generalization error bounded by

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[L(f_{\Theta_S+\epsilon})] \le \sigma^2\Big(G_u(1,2) + \frac{1}{2}P(\sigma,1,2,T)\Big) + 2\sqrt{\frac{\Sigma^2}{2m}\Big(\frac{L(1,2,T)}{2\sigma^2} + ln\frac{1}{\delta}\Big)}$$

To have an incorrect step in CoT after taking the argmax, the error in the output of our transformer on that step must be at least 0.5. Setting the error to be at least 0.5 and solving for δ, we have

$$\delta_S \le e^{-\frac{m}{8\Sigma^2}\left(1-\sigma^2\left(2G_u(1,2)+P(\sigma,1,2,T)\right)\right)^2 + \frac{L(1,2,T)}{2\sigma^2}}$$

When our bound is non-vacuous, we have $\sigma^2\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)<1$, and we have

$$-\frac{m}{8\Sigma^2}\Big(1-\sigma^2\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)\Big)^2+\frac{L(1,2,T)}{2\sigma^2}$$
$$=-\frac{m}{8\Sigma^2}+\frac{2m\sigma^2}{8\Sigma^2}\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)-\frac{m\sigma^4}{8\Sigma^2}\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)^2$$
$$\leq-\frac{m}{8\Sigma^2}+\frac{2m\sigma^2}{8\Sigma^2}\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)$$

Thus, we can use a simpler form of the upper bound which approximates the above:

$$\delta_S\leq e^{-\frac{m}{8\Sigma^2}}e^{\frac{m\sigma^2}{4\Sigma^2}\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)+\frac{L(1,2,T)}{2\sigma^2}}$$

to bound the probability of the entire CoT having an error, we can use a union bound. Thus, the probability of the final answer being incorrect is upper bounded by

$$\delta_{CoT}\leq B_{CoT}:=Te^{-\frac{m}{8\Sigma^2}}e^{\frac{m\sigma^2}{4\Sigma^2}\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)+\frac{L(1,2,T)}{2\sigma^2}}$$

We compare this to our generalization bound for the solution learned by a regularized learner for calculating parity of the full string of length T all at once. In that scenario, there exist (different) λ_1,λ_2 a generalization gap bounded by:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[L(f_{\Theta_{OP}+\epsilon})]\leq\sigma^2\Big(G_u(1,T)+\frac{1}{2}P(\sigma,1,T,T)\Big)+2\sqrt{\frac{\Sigma^2}{2m}\Big(\frac{L(1,T,T)}{2\sigma^2}+ln\frac{1}{\delta}\Big)}$$

The probability of the final answer being correct in this one-pass approach is bounded by

$$\delta_{OP}\leq B_{OP}:=e^{-\frac{m}{8\Sigma^2}}e^{\frac{m\sigma^2}{4\Sigma^2}\Big(2G_u(1,T)+P(\sigma,1,T,T)\Big)+\frac{L(1,T,T)}{2\sigma^2}}$$

Note that in this scenario where the output is a binary, an upper bound on the probability of an error is an upper bound on the expected error. If we compare our bound for CoT with this bound, we can see that when using CoT, the expected error only scales with T, whereas the expected error in the one-pass scenario will scale exponentially with T. We know that G_u is cubic in $D_f=T$, and $P()$ carries a factor of $D_f^{\frac{7}{2}}$. Meanwhile, the norm term $L()$ carries a factor of T. Therefore, assuming m,Σ,σ are all held constant, increasing the degree by a factor of T will exponentially increase the upper bound on the expected error compared to δ_S, our upper bound on the error for a single step.

Mathematically, we have that for fixed σ,

$$B_{OP}e^{\frac{m}{8\Sigma^2}}=e^{\frac{m\sigma^2}{4\Sigma^2}\Big(2G_u(1,T)+P(\sigma,1,T,T)\Big)+\frac{L(1,T,T)}{2\sigma^2}}$$
$$\geq e^{\frac{m\sigma^2T}{8\Sigma^2}\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)+\frac{TL(1,2,T)}{2\sigma^2}}=\frac{1}{T}(B_{CoT}e^{\frac{m}{8\Sigma^2}})^T$$

Thus

$$B_{OP}\geq\frac{B_{CoT}^Te^{\frac{m}{8\Sigma^2}(T-1)}}{T}$$

So far, we have left out one detail, which is that all three of the scenarios – a single CoT step, a full CoT of length T, and a one-pass scenario – will have different optimal values for σ. Suppose we let σ_1 be the σ that minimizes our generalization bound for CoT (and therefore also δ_{CoT}), and let σ_2 be the σ that minimizes δ_{OP}. We write $B_{OP}(\sigma_2)$ and $B(\sigma_1)$ for our bounds with distinct optimal σ. Then we can write

$$B_{OP}(\sigma_2)e^{\frac{m}{8\Sigma^2}}=e^{\frac{m\sigma_2^2}{4\Sigma^2}\Big(2G_u(1,T)+P(\sigma_2,1,T,T)\Big)+\frac{L(1,T,T)}{2\sigma_2^2}}\geq\Big(e^{\frac{m\sigma_2^2}{8\Sigma^2}\Big(2G_u(1,2)+P(\sigma_2,1,2,T)\Big)+\frac{L(1,2,T)}{2\sigma_2^2}}\Big)^T$$

Now, note that the σ that minimizes the expression $\frac{m\sigma^2}{8\Sigma^2}\Big(2G_u(1,2)+P(\sigma,1,2,T)\Big)+\frac{L(1,2,T)}{2\sigma^2}$ is exactly σ_1. Therefore we have

$$B_{OP}(\sigma_2)e^{\frac{m}{8\Sigma^2}}=\Big(e^{\frac{m\sigma_2^2}{8\Sigma^2}\Big(2G_u(1,2)+P(\sigma_2,1,2,T)\Big)+\frac{L(1,2,T)}{2\sigma_2^2}}\Big)^T$$

$$\geq \left(e^{\frac{m\sigma_1^2}{8\Sigma^2}\left(2G_u(1,2)+P(\sigma_2,1,2,T)\right)+\frac{L(1,2,T)}{2\sigma_1^2}}\right)^T = \frac{1}{T}(B_{CoT}(\sigma_1)e^{\frac{m}{8\Sigma^2}})^T$$

Thus,

$$B_{OP}(\sigma_2) \geq \frac{B_{CoT}(\sigma_1)^T e^{\frac{m}{8\Sigma^2}(T-1)}}{T}$$

□

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We claim to provide a new generalization bound that is non-vacuous under reasonable conditions. As we show in 1, our bound is indeed less than 1 for m large. We define non-vacuous to be not only that the bound is <1, but that this occurs for a sample complexity that is better than what would be required for perfect function memorization, which is 2^T. We show clearly that as long as $T \geq 50$, the sample complexity is less than 2^T. In addition, we claim to provide a new framework for bounding the generalization gap of transformers on specific tasks using 1) an existence result by way of a specific construction, 2) derivation of bounds on the sharpness and norm for said construction 3) derivation of bounds on the sharpness for a perturbed version of the construction 4) a feasibility result showing with mechanistic interpretability that this the construction is also feasibly learned through SGD, and 5) the assumption of a general learner that finds a low-sharpness interpolator. We provided each of these ingredients in our paper, and they can all be applied in other settings/constructions. Finally, we claim to provide an explanation of why CoT works better than one-pass inference on the parity task. Our proof of the error bounds shows that CoT's error is upper bounded by a function that is exponentially smaller than our upper bound for the one-pass approach. While our bounds may not be tight, the fact that they are non-vacuous supports the idea that an exponential separation in the upper bound corresponds to an exponential separation in the actual error.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes, our limitations section describes two major limitations of our work, which include the fact that our bounds on the perturbation term are likely not tight, and the fact that our approach to bounding the trace of the hessian is not very scalable to larger networks. We also suggest possible paths to address these limitations in the future.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All of our theoretical results have associated proofs in the appendix. These proofs reference other theorems and lemmas appropriately.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: In our experimental details section, we describe the setup needed to reproduce our experiments, including the exact architecture which is very close to a standard transformer, as well as the libraries for defining the model, and training it in parallel across 8 GPUs. We also describe clearly how to calculate the sub-gaussian constant for our bound. And finally, we describe the minor differences in architecture between our mechanistic interpretability experiment and our main experiment, so that this can also be reproduced.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: There is no dataset for our experiments, as we simply generated the bitstrings used for training data on the fly, since it is so cheap. As of now, our code is not in a state that it is shareable, as it contains fragments from several different versions of our construction and architecture that still need to be cleaned up. But we do commit to releasing the code for the final submission. Part of our reason is concern that somehow the code would not be properly anonymized,

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In our experimental details section in the appendix, we give a full description of the hyperparameters used in our experiments. However these hyperparameters are not necessary to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The only experiment with quantitative results in our paper is the empirical generalization gap. For that experiment, we only had computational resources to train five functions in each complexity class. We do include error bars in 7, which represent the $1-\sigma$ variations, where σ is the standard deviation of the empirical generalization gap for each complexity class. Due to the very small sample size, we we do not have meaningful information on statistical significance. If given more time to run more experiments including more functions per complexity class, we will include statistical significance numbers as well.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the experimental implementation details section, we describe the exact compute resources required for our experiments, and indicate that they took roughly 1 day to complete on 8 A100 GPUs

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: As a theory paper, our research does not rely on any human subjects or personally identifiable information. In addition, since we are not providing a model or a piece of code as our main result, there is no concern of using our paper for harm.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the broader impacts in our Impacts section. Again, as a theory paper dedicated purely to understanding transformers and their generalization properties better, there is no concern about our results having a negative impact on society.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not pose any high risk for misuse. we do not provide a new model for any purposes other than as a toy model representing the properties of traditional transformer models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: We do not use existing assets in our paper other than open source code libraries that are standard in any machine learning task, such as pytorch.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Our paper does not release any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: We do not use any crowdsourcing of any kind.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not use any crowdsourcing of any kind.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: We did not use LLMs in the development of any core methods for this paper. The only LLM usage was for basic search on the internet and minor math-related queries that were not central to the proofs or main results.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.