The Lipschitz Constant of Self-Attention

Hyunjik Kim [1] **George Papamakarios** [1] **Andriy Mnih** [1]

Abstract

Lipschitz constants of neural networks have been explored in various contexts in deep learning, such as provable adversarial robustness, estimating Wasserstein distance, stabilising training of GANs, and formulating invertible neural networks. Such works have focused on bounding the Lipschitz constant of fully connected or convolutional networks, composed of linear maps and pointwise non-linearities. In this paper, we investigate the Lipschitz constant of self-attention, a non-linear neural network module widely used in sequence modelling. We prove that the standard dot-product self-attention is *not* Lipschitz, and propose an alternative L2 self-attention that *is* Lipschitz. We derive an upper bound on the Lipschitz constant of L2 self-attention and provide empirical evidence for its asymptotic tightness. To demonstrate the practical relevance of the theory, we formulate invertible self-attention and use it in a Transformer-based architecture for a character-level language modelling task.

1. Introduction

Lipschitz continuity is a strong form of continuity for functions. Roughly, a function is *Lipschitz continuous* if changing its input by a certain amount cannot change its output by more than K times that amount. The constant K is a constraint on how fast the function can vary, and the smallest such K is known as the function's *Lipschitz constant*.

In deep learning, we often use Lipschitz continuity as a constraint, to control how much a network's output can change relative to its input. Such Lipschitz constraints are useful in several contexts: adversarial robustness (Cisse et al., 2017; Tsuzuku et al., 2018; Anil et al., 2019), generalisation bounds (Sokolić et al., 2017), estimating Wasserstein distances (Peyré & Cuturi, 2019), stabilising training for GANs (Miyato et al., 2018) and constructing invertible models and normalizing flows (Behrmann et al., 2019; Chen et al., 2019; 2018; Grathwohl et al., 2019). However, designing Lipschitz-continuous neural networks and computing (or upper-bounding) their Lipschitz constant is hard. Previous work has focused on fully-connected and convolutional networks, not only because they are common in deep learning, but also because they are simpler to analyze, as compositions of linear maps and pointwise non-linearities (Virmaux & Scaman, 2018; Fazlyab et al., 2019; Latorre et al., 2020).

Recently, *self-attention* (Vaswani et al., 2017) has become a popular alternative to recurrent neural networks. Self-attention is a key component of the Transformer (Vaswani et al., 2017), which has found success in models of various data modalities, such as natural-language processing (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020), computer vision (Zhang et al., 2019; Parmar et al., 2019), audio generation (Huang et al., 2019), and reinforcement learning (Parisotto et al., 2020). However, no previous work has analyzed the Lipschitz properties of self-attention, and thus it has been unclear whether self-attention is a viable option in applications that require Lipschitz constraints.

In this work, we address this gap in the theory of self-attention by providing a thorough analysis of its Lipschitz properties. We make the following contributions:

- We prove that the widely used *dot-product self-attention* is *not* Lipschitz, and therefore not suitable to use in applications requiring Lipschitz constraints.
- We formulate *L2 self-attention* as an alternative, and show that it *is* Lipschitz by deriving a theoretical upper bound on its Lipschitz constant.
- We use this bound to formulate an invertible variant of self-attention, and explore its use in a Transformer architecture for a character-level language modelling task.

The full-sized version of this paper can be found at: `https://arxiv.org/abs/2006.04710`

[1]DeepMind, London, UK. Correspondence to: Hyunjik Kim <hyunjikk@google.com>.

Second workshop on *Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models* (ICML 2020), Virtual Conference

2. Lipschitz Constant of Fully-Connected/Convolutional Layers

We first define the notion of Lipschitz continuity, and proceed to define the Lipschitz constant.

Definition 2.1. Given two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, a function $f : \mathcal{X} \to \mathcal{Y}$ is called *Lipschitz continuous* (or *K-Lipschitz*) if there exists a constant $K \geq 0$ such that $d_{\mathcal{Y}}(f(x), f(x')) \leq K d_{\mathcal{X}}(x, x')$ for all $x, x' \in \mathcal{X}$. The smallest such K is the *Lipschitz constant* of f, denoted $\mathrm{Lip}(f)$.

In this paper, we focus on the common case where $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \mathbb{R}^m$, and $d_{\mathcal{X}}, d_{\mathcal{Y}}$ are induced by a p-norm $\|x\|_p := (\sum_i |x_i|^p)^{1/p}$. We will primarily consider the cases $p = 2$ and $p = \infty$, where $\|x\|_\infty := \max_i |x_i|$. To emphasise the dependence of the Lipschitz constant on the choice of p-norm, we will often denote it by $\mathrm{Lip}_p(f)$.

Next, we outline some basic results that are useful for estimating Lipschitz constants of fully-connected networks (`FCN`) and convolutional neural networks (`CNN`), also covered in related works (Virmaux & Scaman, 2018; Behrmann et al., 2019). To begin with, the following theorem suggests a way to bound $\mathrm{Lip}_p(f)$ for a differentiable Lipschitz function f:

Theorem 2.1 (Federer, 1969). *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable and Lipschitz continuous under a choice of p-norm $\|\cdot\|_p$. Let $J_f(x)$ denote its total derivative (Jacobian) at x. Then $\mathrm{Lip}_p(f) = \sup_{x \in \mathbb{R}^n} \|J_f(x)\|_p$ where $\|J_f(x)\|_p$ is the induced operator norm on $J_f(x)$.*

Hence if f is a linear map represented by a matrix W then

$$\mathrm{Lip}_p(f) = \|W\|_p = \begin{cases} \sigma_{\max}(W), & \text{if } p = 2 \\ \max_i \sum_j |W_{ij}| & \text{if } p = \infty \end{cases}$$

where $\|W\|_p$ is the operator norm on matrices induced by the vector p-norm, and $\sigma_{\max}(W)$ is the largest singular value of W. Under this choice of norm, many common non-linearities (including `relu`, `sigmoid`, `tanh`, `elu`) are 1-Lipschitz. $\|W\|_2 = \sigma_{\max}(W)$ is usually estimated via *power iteration*; we provide details on how this is done in Appendix A.

Since we now know the Lipschitz constants of the components of both `FCN` and `CNN`, we can bound their Lipschitz constants by applying the following lemma:

Lemma 2.1 (Federer, 1969). *Let g, h be two composable Lipschitz functions. Then $g \circ h$ is also Lipschitz with $\mathrm{Lip}(g \circ h) \leq \mathrm{Lip}(g)\,\mathrm{Lip}(h)$.*

For `FCN` and `CNN`, this gives us the following upper bound on their Lipschitz constants:

Corollary 2.1. For a fully-connected (`FCN`) or a convolutional neural network (`CNN`) $f = W_K \circ \rho_{K-1} \circ W_{K-1} \circ \ldots \circ \rho_1 \circ W_1$, we have $\mathrm{Lip}_p(f) \leq \prod_k \|W_k\|_p$ under a choice of p-norm with 1-Lipschitz non-linearities ρ_k.

The above bound is not necessarily tight; there are various works that compute tighter bounds for `FCN` and `CNN` (e.g. Virmaux & Scaman, 2018; Fazlyab et al., 2019; Latorre et al., 2020).

3. Lipschitz Constant of Self-Attention

3.1. Dot-product self-attention is *not* Lipschitz

Moving on, we investigate whether self-attention is Lipschitz. We first consider the widely used *(scaled) dot-product multihead self-attention* as formulated by Vaswani et al. (2017). Let $x_1, \ldots, x_N$ be a sequence of N elements, where $x_i \in \mathbb{R}^D$ for $i = 1, \ldots, N$. We represent this sequence as a matrix $X \in \mathbb{R}^{N \times D}$ such that the ith row of X is the ith element of the sequence, i.e. $X_{i:} = x_i^\top$. Dot-product multihead self-attention (`DP-MHA`) is a map from $\mathbb{R}^{N \times D}$ to $\mathbb{R}^{N \times D}$ consisting of H 'heads', where H is chosen to divide D. Each head is a map from $\mathbb{R}^{N \times D}$ to $\mathbb{R}^{N \times D/H}$ defined by

$$DP(X) := \mathrm{softmax}\left(\frac{XW^Q(XW^K)^\top}{\sqrt{D/H}}\right)XW^V, \quad (1)$$

where $W^Q, W^K, W^V \in \mathbb{R}^{D \times D/H}$ are learnable parameters specific to each head. The input to the softmax is an $N \times N$ matrix of dot products (hence *dot-product* self-attention), and the softmax is applied to each row of this matrix. Finally, the outputs of all heads are concatenated into an $N \times D$ matrix and are right multiplied by $W^O \in \mathbb{R}^{D \times D}$, thus `DP-MHA` is defined by

$$MHA(X) := \left[DP^1(X), \ldots, DP^H(X)\right] W^O. \quad (2)$$

We first prove that MHA as defined above is *not* Lipschitz, assuming that the MHA map is non-trivial, i.e. $W^Q, W^K, W^V, W^O \neq 0$.

Theorem 3.1. `DP-MHA` *is not Lipschitz for any vector p-norm $\|\cdot\|_p$ with $p \in [1, \infty]$.*

Proof. See Appendix C. □

The implications of this result are the following. (1) There can be undesirable behaviour (e.g. training instabilities) for the Transformer when some inputs are close to zero. (2) Dot-product self-attention (and hence the standard Transformer) is not a suitable choice when we require a Lipschitz neural network, such as for formulating invertible residual networks (Behrmann et al., 2019).

3.2. L2 self-attention: a Lipschitz formulation of self-attention

The pathology in dot-product self-attention arises because the softmax probabilities $P_{i:}$ are constant with respect to $x_{\neq i}$ when $x_i = 0$ (c.f. Appendix C). This behaviour can be undesirable as we may want P_{ij} to be high for x_j close to x_i, and low for x_j far from x_i, regardless of whether x_i is large or small. Hence we propose an alternative form of self-attention based on L2 distance:

$$P_{ij} \propto \exp\left(-\frac{\left\|x_i^\top W^Q - x_j^\top W^K\right\|_2^2}{\sqrt{D/H}}\right), \qquad (3)$$

with the normalisation constant ensuring that $\sum_j P_{ij} = 1$. We will refer to it as *L2 self-attention*. It is reminiscent of the standard squared-exponential kernel, but with softmax normalisation that ensures each row of the kernel matrix sums to 1. Normalisation is usually necessary to deal with inputs of varying length N (Wang et al., 2018), hence we keep the softmax for L2 self-attention. Similarly to dot-product self-attention, L2 self-attention can be computed efficiently with matrix operations; see Appendix B for details.

We first state the mathematical formulation of L2 multihead self-attention (`L2-MHA`) before proving the main result — the upper bound of its Lipschitz constant with respect to $\|\cdot\|_p$ for $p = 2, \infty$. The full `L2-MHA` map $F : \mathbb{R}^{N\times D} \to \mathbb{R}^{N\times D}$ is defined as

$$F(X) := \left[f^1(X)W^{V,1}, \dots, f^H(X)W^{V,H}\right] W^O \quad \text{where} \quad f^h(X) := P^h X A_h.$$

In the above, $W^{V,h} \in \mathbb{R}^{D\times D/H}$, $W^O \in \mathbb{R}^{D\times D}$, P^h is defined as in Equation (3) with $W^{Q,h} = W^{K,h} \in \mathbb{R}^{D\times D/H}$, and $A_h := W^{Q,h}W^{Q,h^\top}/\sqrt{D/H} \in \mathbb{R}^{D\times D}$. There are two changes from the usual form of multihead self-attention:

(1) We require $W^{Q,h} = W^{K,h}$ for each head $f^h(X)$ to be Lipschitz. In Lemma D.1 of Appendix D we show that `L2-MHA` is *not* Lipschitz for arbitrary $W^{Q,h}$, $W^{K,h}$, and that tying $W^{Q,h} = W^{K,h}$ is sufficient for `L2-MHA` to be Lipschitz.

(2) In each head $f^h(X)$, right multiplication by A_h has been included for the theorem below to hold (details are in the proof). In practice, there is little harm done by this extra linear transformation, since when the heads are combined together in F, each $f^h(X)$ is additionally transformed by $W^{V,h}$, a free parameter.

The second main result of the paper is the following:

Theorem 3.2. `L2-MHA` *is Lipschitz, with the following bound on* $\mathrm{Lip}_\infty(F)$*:*

$$\mathrm{Lip}_\infty(F) \leq \left(4\phi^{-1}(N-1) + \frac{1}{\sqrt{D/H}}\right)\|W^{O^\top}\|_\infty \max_h \|W^{Q,h}\|_\infty \|W^{Q,h^\top}\|_\infty \max_h \|W^{V,h^\top}\|_\infty$$

and the following bound on $\mathrm{Lip}_2(F)$*:*

$$\mathrm{Lip}_2(F) \leq \frac{\sqrt{N}}{\sqrt{D/H}}\left(4\phi^{-1}(N-1)+1\right)\left(\sqrt{\textstyle\sum_h \|W^{Q,h}\|_2^2\,\|W^{V,h}\|_2^2}\right)\|W^O\|_2$$

where $\phi(x) := x\exp(x+1)$ *is an invertible univariate function on* $x > 0$*, and* N *is the input sequence length.*

Specifically, $\phi^{-1}(N-1) = W_0(\frac{N}{e})$ *where* W_0 *is the Lambert* W*-function, which grows sub-logarithmically as* $O(\log N - \log\log N)$ *(Corless et al., 1996). Hence the above bounds can be simplified to* $O(\log N)$ *for* $p = \infty$ *and* $O(\sqrt{N}\log N)$ *for* $p = 2$*.*

Proof. See Appendix D. Also, Appendix E shows how to modify the proof for the case with masking. □

Appendix F provides empirical evidence of the asymptotic tightness of the bound on $\mathrm{Lip}_\infty(F)$.

4. Application: Invertible Self-Attention

4.1. Invertible residual network

Consider the residual function $g(x) := x + f(x)$. Behrmann et al. (2019) give the following sufficient condition for its invertibility: if f is a *contraction* with respect to some metric, i.e. if $\mathrm{Lip}(f) < 1$, and the metric space on which f is defined is complete, then g is invertible. (A Euclidean space with a metric induced by a p-norm $\|\cdot\|_p$ for $p \in [1, \infty]$ is always complete.) The inverse $g^{-1}(y)$ is the unique fixed point of the recursion $x^{i+1} := y - f(x^i)$, since by the definition of the inverse $y = g^{-1}(y) + f(g^{-1}(y))$. Because f is a contraction, *Banach's Fixed Point Theorem* guarantees that this fixed point exists and is unique for all y, and that the recursion converges for all initial values x^0 (often set to y in practice) exponentially fast.

A composition of such invertible residual blocks is also invertible. Behrmann et al. (2019) use this observation to design invertible ResNets: they take f to be a `CNN` normalised by an upper bound on $\mathrm{Lip}(f)$ given by Corollary 2.1, making the resulting function *contractive*. For the 2-norm $\|\cdot\|_2$, a hyperparameter $c < 1$ is chosen and each linear map (convolution) W in the `CNN` is multiplied by $c/\|W\|_2$ if $c < \|W\|_2$ where $\|W\|_2$ is estimated by power iteration (c.f. Appendix A). This multiplicative factor determines the scale of the resulting Lipschitz constant.

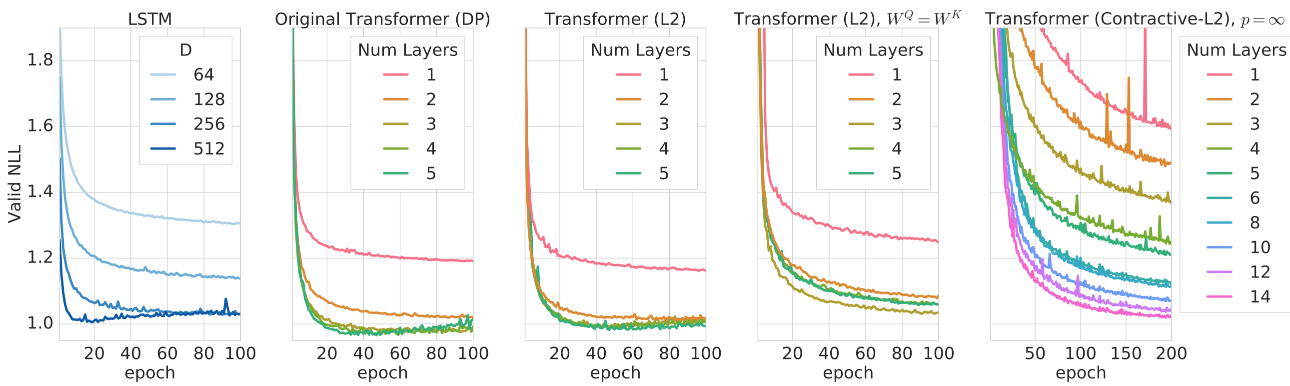

Figure 1. Validation NLL curves during training for various LSTM/Transformer models.

4.2. Invertible self-attention

The standard use case of self-attention is with a residual connection inside the Transformer. A Transformer block is composed of residual blocks of multihead self-attention (`MHA`) and fully-connected (`FCN`) layers (Figure 2). Hence similarly to invertible ResNets, we can normalise `L2-MHA` by the upper bounds given in Theorem 3.2 to obtain `Contractive-L2-MHA` f, with which we can obtain invertible self-attention $g(x) = x + f(x)$. Since `Dropout` is also part of the residual branch along with `Contractive-L2-MHA`, we should check that it is also contractive. At test time, `Dropout` multiplies inputs by the dropout keep probability $p < 1$, so it is a contraction with Lipschitz constant p at evaluation time. At training time, `Dropout` amounts to setting some inputs to zero, while keeping other inputs constant. This can be expressed as right multiplication by a diagonal binary matrix M, and for such matrices we can verify $\|M\|_p := \sup_{\|x\|_p=1} \|Mx\|_p \leq 1$.

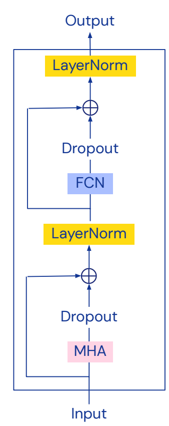

Figure 2. Transformer

In the next section, we investigate the properties of invertible self-attention and how it compares with the standard dot-product self-attention, by replacing `DP-MHA` in the Transformer with `Contractive-L2-MHA`, hence replacing the residual self-attention module with invertible self-attention. Also, in Appendix H we numerically check the invertibility of `Contractive-L2-MHA`, and show how `DP-MHA` fails to yield an invertible residual block.

5. Experimental Results

A natural question to ask is: how does the expressiveness of `L2-MHA` and `Contractive-L2-MHA` (that leads to invertible self-attention with the residual connection) compare with the original `DP-MHA`? We investigate this question by comparing the performance of the original Transformer and the Transformer with invertible self-attention (c.f. Figure 2) at character-level language modelling on the Penn Treebank dataset (Marcus et al., 1993). We compare the validation negative log-likelihood (NLL) of a baseline LSTM, the original Transformer (`DP-MHA`), and a series of models between the original Transformer and the Transformer with invertible self-attention (`Contractive-L2-MHA`), making one change at a time. For `Contractive-L2-MHA`, we normalise `L2-MHA` by the bound on $\mathrm{Lip}_\infty(F)$ as it is tighter than the bound on $\mathrm{Lip}_2(F)$. See Appendix G for experimental details.

The results are shown in Figure 1. The first plot shows the best performing LSTM reaching a validation NLL of around 1.0, and the second plot shows the best performing Transformer reaching a slightly improved performance for 3–5 layers of Transformer blocks. We observe instabilities in training for a higher number of layers, requiring careful tuning of the learning rate schedule for stability at the cost of performance, a commonly observed phenomenon in the literature of deep Transformer architectures (Bapna et al., 2018; Parisotto et al., 2020). The third plot shows results for the Transformer with `DP-MHA` replaced with `L2-MHA` but without tying W^Q and W^K, and we observe a very similar validation performance. The fourth plot shows the change when we further tie the query and key weights (making $W^Q = W^K$); we see that there is a small degradation in performance. Here the number of trainable parameters has been reduced, which is partly responsible for the degradation in performance, but the performance is still reasonable. We note that performance saturates at around 5 layers for each Transformer model so far. On the rightmost plot we show results when further dividing self-attention in each block by the upper bound on $\mathrm{Lip}_\infty(F)$, to obtain invertible self-attention. This does give reduced performance for the same number of layers, but we can attain similar performance with more layers, no longer saturating at 5 layers.

Thus we conclude the following. (1) Replacing the dot-product with the L2 distance incurs hardly any loss in expressiveness. (2) Tying the query and key weights to obtain Lipschitz self-attention incurs a small loss in expressiveness due to reducing the number of trainable parameters. (3) Dividing by the upper bound on $\mathrm{Lip}_\infty(F)$ to obtain invertible self-attention incurs a noticeable loss in expressiveness, but also has a stabilization effect on the optimisation of the Transformer, allowing one to compensate for the apparent loss in expressiveness by increasing the number of layers.

Acknowledgements

We would like to thank Adam Kosiorek, Arnaud Doucet, Yee Whye Teh, Michalis Titsias, Emilien Dupont and Theophane Weber for helpful discussion and feedback.

References

Anil, C., Lucas, J., and Grosse, R. Sorting out Lipschitz function approximation. In *International Conference on Machine Learning*, pp. 291–301, 2019.

Bapna, A., Chen, M. X., Firat, O., Cao, Y., and Wu, Y. Training deeper neural machine translation models with transparent attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3028–3033, 2018.

Behrmann, J., Grathwohl, W., Chen, R. T. Q., Duvenaud, D., and Jacobsen, J.-H. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pp. 6571–6583, 2018.

Chen, R. T. Q., Behrmann, J., Duvenaud, D., and Jacobsen, J.-H. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, 2019.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 854–863, 2017.

Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E. On the Lambert W function. *Advances in Computational mathematics*, 5(1):329–359, 1996.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 11423–11434, 2019.

Federer, H. *Geometric Measure Theory*. Classics in Mathematics. Springer Berlin Heidelberg, 1969. ISBN 9783642620102.

Grathwohl, W., Chen, R. T. Q., Betterncourt, J., Sutskever, I., and Duvenaud, D. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., and Eck, D. Music Transformer. In *International Conference on Learning Representations*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, 2015.

Latorre, F., Rolland, P., and Cevher, V. Lipschitz constant estimation of neural networks via sparse polynomial optimization. In *International Conference on Learning Representations*, 2020.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

Mises, R. and Pollaczek-Geiringer, H. Praktische verfahren der gleichungsauflösung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 9(2):152–164, 1929.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*, 2018.

Parisotto, E., Song, H. F., Rae, J. W., Pascanu, R., Gulcehre, C., Jayakumar, S. M., Jaderberg, M., Kaufman, R. L., Clark, A., Noury, S., Botvinick, M. M., Heess, N., and Hadsell, R. Stabilizing Transformers for reinforcement learning. In *International Conference on Machine Learning*, 2020.

Parmar, N., Ramachandran, P., Vaswani, A., Bello, I., Levskaya, A., and Shlens, J. Stand-alone self-attention in vision models. In *Advances in Neural Information Processing Systems*, pp. 68–80, 2019.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6): 355–607, 2019.

Sokolić, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 6541–6550, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pp. 3835–3844, 2018.

Wang, X., Girshick, R., Gupta, A., and He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention Generative Adversarial Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 7354–7363, 2019.

A. Power Iteration

Although $\|W\|_\infty$ can be computed efficiently in $O(nm)$ time for $W \in \mathbb{R}^{m\times n}$, naïvely computing $\|W\|_2 = \sigma_{\max}(W) := \sqrt{\lambda_{\max}(W^\top W)}$ requires $O(n^3)$ operations. (By $\lambda_{\max}(A)$ we denote the greatest eigenvalue of a symmetric matrix A.) We can however obtain an underestimate $\tilde{\sigma}(W)$ via *power iteration*:

$$b_{k+1} = \frac{W^\top W b_k}{\|W^\top W b_k\|_2}, \quad \tilde{\sigma}_k(W) = \sqrt{\frac{b_k^\top W^\top W b_k}{b_k^\top b_k}}, \tag{4}$$

with each iteration taking $O(n^2)$ time. Then using $K \ll n$ iterations gives us an underestimate $\tilde{\sigma}_K$ in $O(Kn^2)$ time. Since this is an underestimate, the resulting approximation to the Lipschitz constant of the linear map will not be an upper bound. However the number of power iterations is usually chosen so that $\tilde{\sigma}$ is accurate enough — $K = 5$ is shown to be sufficient in the context of fully connected networks or convolutions considered by Behrmann et al. (2019).

The iteration will converge if $W^\top W$ has an eigenvalue that is strictly greater in magnitude than its other eigenvalues, and the starting vector b_0 has a nonzero component in the direction of an eigenvector associated with the dominant eigenvalue. This happens with probability 1 if b_0 is chosen at random, and the convergence is geometric with ratio $|\lambda_2/\lambda_{\max}|$ where λ_2 is the eigenvalue with second largest magnitude (Mises & Pollaczek-Geiringer, 1929).

B. Efficient Computation of L2 Self-attention

Dot-product self-attention only requires a few matrix multiplications to compute the logits (i.e. the inputs to the softmax) between all pairs of inputs, without having to loop over pairs, hence it can be computed efficiently. Similarly, we can show that L2 self-attention can also be computed in an efficient manner. Using the identity $\|a-b\|_2^2 = \|a\|_2^2 - 2a^\top b + \|b\|_2^2$ we can compute the logits of L2 attention between all pairs via matrix multiplications and computation of row-wise L2 norms, with negligible overhead compared to dot-product self-attention. Specifically, for L2 self-attention we can show that

$$P = \text{softmax}\left(-\frac{\|XW^Q\|_{\text{row}}^2 \mathbb{1}^\top - 2XW^Q(XW^K)^\top + \mathbb{1}\|XW^K\|_{\text{row}}^{2\top}}{\sqrt{D/H}}\right), \tag{5}$$

where $\|A\|_{\text{row}}^2$ applies the squared L2 norm to each row of A, so if $A \in \mathbb{R}^{m\times n}$ then $\|A\|_{\text{row}}^2 \in \mathbb{R}^m$.

C. Proof of Theorem 3.1

It is sufficient to show that a single head DP is not Lipschitz, since MHA is a linear combination of the outputs of each head. Let us write Equation (1) as $DP(X) = PXW^V$, where $P \in \mathbb{R}^{N\times N}$ is the output of the softmax (we suppress the dependence of P on X to reduce clutter below). P is a stochastic matrix, i.e. its entries are non-negative and its rows sum to 1. Since the rows of X are the x_i's, a linear transformation of each x_i by some matrix A is equivalent to right multiplication of X by $A^\top$. So right multiplication of X by W^V is a linear map and thus Lipschitz. Therefore, we are interested in the mapping $f(X) = PX$; this is *not* a linear mapping because P itself is a non-linear function of X. In fact, we show that f is *not* Lipschitz, thus proving the first main result of the paper. The mapping f can be written as

$$f(X) = PX = \text{softmax}\left(XA^\top X^\top\right)X = \begin{bmatrix} f_1(X)^\top \\ \vdots \\ f_N(X)^\top \end{bmatrix}, \tag{6}$$

where $A = W^K W^{Q^\top}/\sqrt{D/H} \in \mathbb{R}^{D\times D}$ and $f_i(X) = \sum_{j=1}^N P_{ij}x_j$ with $P_{i:}^\top = \text{softmax}(XAx_i)$. Hence f can be interpreted as a map of each x_i to a point in the convex hull of $x_1, ..., x_N$. Since f is a map from $\mathbb{R}^{N\times D}$ to $\mathbb{R}^{N\times D}$, its Jacobian is

$$J_f = \begin{bmatrix} J_{11} & \dots & J_{1N} \\ \vdots & \ddots & \vdots \\ J_{N1} & \dots & J_{NN} \end{bmatrix} \in \mathbb{R}^{ND\times ND}, \tag{7}$$

where $J_{ij} = \frac{\partial f_i(X)}{\partial x_j} \in \mathbb{R}^{D\times D}$. By taking partial derivatives we can show that $J_{ij} = X^\top P^{(i)}\left[e_{ji}XA^\top + XA\delta_{ij}\right] + P_{ij}I$ where $e_{ij} \in \mathbb{R}^{N\times N}$ is a binary matrix with zeros everywhere except the (i,j)th entry, δ_{ij} is the Kronecker delta, and

$P^{(i)} := \mathrm{diag}(P_{i:}) - P_{i:}^\top P_{i:}$. So for $i = j$:

$$\begin{aligned} J_{ii} &= X^\top P^{(i)} e_{ii} X A^\top + X^\top P^{(i)} X A + P_{ii} I \\ &= P_{ii} \left(x_i - \textstyle\sum_k P_{ik} x_k\right) x_i^\top A^\top + X^\top P^{(i)} X A + P_{ii} I. \end{aligned} \tag{8}$$

For the last equality, note $e_{ii} X$ has all rows equal to zero except for the ith row given by $x_i^\top$. We can then verify that $X^\top P^{(i)} e_{ii} X$ simplifies to $P_{ii}(x_i - \sum_k P_{ik} x_k) x_i^\top$.

For vector p-norms, $\|J_f\|_p$ is bounded if and only if its entries are bounded, by definition of the operator norm. The entries of $X^\top P^{(i)} X A$ are bounded for arbitrary A only if the entries of $X^\top P^{(i)} X$ are bounded. So let us investigate the entries of this $D \times D$ matrix. Writing out each term of the matrix, we observe that it is in fact a covariance matrix of a discrete distribution. Specifically:

$$[X^\top P^{(i)} X]_{lm} = \textstyle\sum_k P_{ik} x_{kl} x_{km} - \left(\sum_k P_{ik} x_{kl}\right) \left(\sum_k P_{ik} x_{km}\right) = \mathrm{Cov}(\mathbb{X}_l, \mathbb{X}_m), \tag{9}$$

where $\mathbb{X}$ is a discrete distribution with support at the inputs $\{x_1, \ldots, x_N\}$ and probability mass function given by their softmax probabilities $\mathbb{P}(\mathbb{X} = x_j) = P_{ij}$. A consequence of this interpretation is that $P^{(i)}$ is *positive semi-definite* (PSD) since for $D = 1$, Equation (9) becomes $X^\top P^{(i)} X = \mathrm{Var}(\mathbb{X}) \geq 0$, with equality if and only if the x_j are all equal.

We use this observation to show that the terms of J_{ii} are unbounded, and so DP-MHA is *not* Lipschitz. Consider the case $x_i = 0$. Then $P_{i:}^\top = \mathrm{softmax}\left(X A x_i\right) = \frac{1}{N}\mathbb{1}$, i.e. we have uniform attention regardless of $x_{\neq i}$. The first term of J_{ii} in Equation (8) disappears since $x_i = 0$, and the last term becomes $\frac{1}{N} I$. For the second term, the entries $[X^\top P^{(i)} X]_{ll} = \mathrm{Var}(\mathbb{X}_l)$ are unbounded since the latter is equal to the sample variance of $x_{1l}, \ldots, x_{Nl}$, which can be arbitrarily large.

The high-level intuition for this result is as follows. At $x_i = 0$, $f_i(X) = \frac{1}{N}\sum_k x_k$, the mean of the inputs. The rate of change of f_i is governed by how fast the softmax saturates when x_i is perturbed, which is determined by how spread out the $x_{\neq i}$ are. The more spread out they are (the higher the sample variance), the greater the rate of saturation of the softmax, and the faster the rate of change of f_i. Since the sample variance of $x_{\neq i}$ can be arbitrarily large, the rate of change of f_i can also be arbitrarily large, i.e. the entries of the Jacobian (and hence its p-norm) can become arbitrarily large.

A natural question to ask is whether we can add bias terms b^Q to $x_i^\top W^Q$ and b^K to $x_j^\top W^K$ to resolve this issue. The answer is *no* in general. It can again be shown that J_{ii} is unbounded when x_i is chosen such that $x_i^\top W^Q + b^Q = 0$ (such a choice is possible assuming W^Q is full rank, a dense set in $\mathbb{R}^{D \times D/H}$). Then we again have $P_{i:}^\top = \frac{1}{N}\mathbb{1}$, and the diagonal entries of $X^\top P^{(i)} X$ are again unbounded.

D. Proof of Theorem 3.2

Recall the formulation of L2-MHA:

$$\begin{aligned} F &: \mathbb{R}^{N \times D} \to \mathbb{R}^{N \times D} \\ F(X) &= \left[f^1(X) W^{V,1}, \ldots, f^H(X) W^{V,H}\right] W^O \\ f^h(X) &= P^h X A_h \\ P^h_{ij} \propto \exp(L_{ij}) &:= \exp\left(-\frac{\|x_i^\top W^{Q,h} - x_j^\top W^{K,h}\|_2^2}{\sqrt{D/H}}\right), \quad \sum_j P^h_{ij} = 1 \end{aligned}$$

where we have that $W^{Q,h}, W^{K,h}, W^{V,h} \in \mathbb{R}^{D \times D/H}$, $W^O \in \mathbb{R}^{D \times D}$, $P^h \in \mathbb{R}^{N \times N}$ and $A_h := W^{Q,h} W^{Q,h^\top} / \sqrt{D/H} \in \mathbb{R}^{D \times D}$, and the softmax is applied to each row of the input matrix. Recall Equation (5):

$$P^h = \mathrm{softmax}\left(-\frac{\|X W^{Q,h}\|_{\text{row}}^2 \mathbb{1}^\top - 2 X W^{Q,h} (X W^{K,h})^\top + \mathbb{1} \|X W^{K,h}\|_{\text{row}}^{2\top}}{\sqrt{D/H}}\right).$$

D.1. L2 self-attention is *not* Lipschitz for general W^Q, W^K

Let us first look at the case of $H = 1$ and suppress the index h to reduce clutter. Consider the map $\tilde{f}(X) := P X$, so $f(X) = \tilde{f}(X) A$. We need $\tilde{f}$ to be Lipschitz for f and hence F to be Lipschitz. Note that P is defined as:

$$P_{ij} \propto \exp(L_{ij}) := \exp\left(-\frac{\|x_i^\top W^Q - x_j^\top W^K\|_2^2}{\sqrt{D/H}}\right)$$

and the normalisation constant satisfies $\sum_j P_{ij} = 1$, for $P \in \mathbb{R}^{N\times N}$, $X \in \mathbb{R}^{N\times D}$.

For L2 self-attention, we may take partial derivatives and use the chain rule to show that the Jacobian of $\tilde{f}$ is:

$$J_{\tilde{f}} = \begin{bmatrix} \tilde{J}_{11} & \dots & \tilde{J}_{1N} \\ \vdots & \ddots & \vdots \\ \tilde{J}_{N1} & \dots & \tilde{J}_{NN} \end{bmatrix} \in \mathbb{R}^{ND\times ND} \tag{10}$$

with

$$\tilde{J}_{ij} = X^\top P^{(i)} \frac{\partial L_{i:}}{\partial x_j} + P_{ij} I \in \mathbb{R}^{D\times D} \tag{11}$$

where

$$\frac{\partial L_{i:}}{\partial x_j} = \frac{2}{\sqrt{D/H}} \left[\left(XW^K - \mathbb{1} x_i^\top W^Q \right) W^{Q^\top} \delta_{ij} + \left(e_{ji} X W^Q - e_{jj} X W^K \right) W^{K^\top} \right] \tag{12}$$

and

$$P^{(i)} := \operatorname{diag}(P_{i:}) - P_{i:}^\top P_{i:} = \begin{bmatrix} P_{i1}(1-P_{i1}) & -P_{i1}P_{i2} & \dots & -P_{i1}P_{iN} \\ -P_{i2}P_{i1} & P_{i2}(1-P_{i2}) & \dots & -P_{i2}P_{iN} \\ \vdots & \vdots & \ddots & \vdots \\ -P_{iN}P_{i1} & -P_{iN}P_{i2} & \dots & P_{iN}(1-P_{iN}) \end{bmatrix},$$

$$P_{ij} = \frac{\exp\left(-\|x_i^\top W^Q - x_j^\top W^K\|_2^2\right)}{\sum_k \exp\left(-\|x_i^\top W^Q - x_k^\top W^K\|_2^2\right)}.$$

Recall that $e_{ji} \in \mathbb{R}^{N\times N}$ is a binary matrix with zeros everywhere except the (j,i)th entry. Hence $e_{ji}X$ has all rows equal to zero except for the jth row given by $x_i^\top$. We can then verify:

$$X^\top P^{(i)} e_{ji} X = P_{ij}(x_j - \sum_k P_{ik} x_k) x_i^\top. \tag{13}$$

Also note $P^{(i)}$ is symmetric, and each row/colum sums to 0, i.e. $P^{(i)}\mathbb{1} = \mathbb{1}^\top P^{(i)} = 0$. Hence we may simplify the Jacobian terms as follows:

$$\begin{aligned} \tilde{J}_{ii} &= \frac{2}{\sqrt{D/H}} \left[X^\top P^{(i)} (XW^K - \mathbb{1} x_i^T W^Q) W^{Q^\top} + X^\top P^{(i)} e_{ii} X (W^Q - W^K) W^{K^\top} \right] + P_{ii} I \\ &= \frac{2}{\sqrt{D/H}} \left[X^\top P^{(i)} (XW^K - \mathbb{1} x_i^T W^Q) W^{Q^\top} + P_{ii}(x_i - \sum_k P_{ik} x_k) x_i^\top (W^Q - W^K) W^{K^\top} \right] + P_{ii} I \\ &= \frac{2}{\sqrt{D/H}} \left[X^\top P^{(i)} X W^K W^{Q^\top} + P_{ii}(x_i - \sum_k P_{ik} x_k) x_i^\top (W^Q - W^K) W^{K^\top} \right] + P_{ii} I, \end{aligned} \tag{14}$$

and for $i \neq j$:

$$\begin{aligned} \tilde{J}_{ij} &= \frac{2}{\sqrt{D/H}} X^\top P^{(i)} (e_{ij} X W^Q - e_{jj} X W^K) W^{K^\top} + P_{ij} I \\ &= \frac{2}{\sqrt{D/H}} P_{ij} (x_j - \sum_k P_{ik} x_k)(x_i^\top W^Q - x_j^\top W^K) W^{K^\top} + P_{ij} I. \end{aligned} \tag{15}$$

We are now ready to show that $\tilde{f}$ is *not* Lipschitz for general W^Q, W^K:

Lemma D.1. *If $W^K \in \mathbb{R}^{D\times D/H}$ is full rank (i.e. full column rank), and $W^K \neq W^Q$, then J_{ij} has terms that are unbounded for $i \neq j$, hence $\tilde{f}$ is* not *Lipschitz.*

Proof. Let us investigate the expression $\tilde{K}_{ij} := P_{ij} W^{K^\top} (x_j - \sum_k P_{ik} x_k)(x_i^\top W^Q - x_j^\top W^K)$ for $i \neq j$, which is related to $\tilde{J}_{ij}$ as follows by Equation (15):

$$W^{K^\top} \tilde{J}_{ij} = \frac{2}{\sqrt{D/H}} \tilde{K}_{ij} + P_{ij} I.$$

It suffices to show that $\tilde{K}_{ij}$ is unbounded to show that $\tilde{J}_{ij}$ is unbounded, since W^K is full rank and $P_{ij} \in [0,1]$.

Let $y_j^\top = x_i^\top W^Q - x_j^\top W^K$. Then we have:

$$\begin{aligned} y_j - \sum_k P_{ik} y_k &= W^{Q^\top} x_i - W^{K^\top} x_j - \sum_k P_{ik}(W^{Q^\top} x_i - W^{K^\top} x_k) \\ &= W^{Q^\top} x_i - W^{K^\top} x_j - (W^{Q^\top} x_i - \sum_k P_{ik} W^{K^\top} x_k) \\ &= -W^{K^\top}(x_j - \sum_k P_{ik} x_k). \end{aligned}$$

Hence $\tilde{K}_{ij} = -P_{ij}(y_j - \sum_k P_{ik} y_k) y_j^\top$. Note y_i can take an arbitrary value in $\mathbb{R}^{D/H}$, since $W^K \neq W^Q$ and W^K is full-rank.

For all $j \neq i$, let us choose x_j such that $y_j = -y_i$. This is possible for any value of y_i since W^K is full-rank. Note $y_j = -y_i$ and not y_i. We then have that $\|y_j\|_2^2$ is equal for all j, hence $P_{ij} := \frac{\exp(-\|y_j\|_2^2)}{\sum_k \exp(-\|y_k\|_2^2)} = \frac{1}{N}$ for all j. Then for $i \neq j$, $\tilde{K}_{ij}$ simplifies to

$$\tilde{K}_{ij} = -\frac{1}{N}\left(-y_i - \frac{1}{N}(N-2)(-y_i)\right)(-y_i)^\top = -\frac{2N-2}{N^2} y_i y_i^\top$$

whose entries are unbounded since y_i can be any vector in $\mathbb{R}^{D/H}$ (note we assume $N \geq 2$ for self-attention to be well-defined, hence $2N - 2 \neq 0$). □

D.2. L2 self-attention is Lipschitz for $W^Q = W^K$

Hence we impose the restriction that $W^K = W^Q$. With this assumption we have

$$P_{ij} \propto \exp\left(-\|(x_i - x_j)^\top \sqrt{A}\|_2^2\right) \tag{16}$$

where $A = W^Q W^{Q^\top} / \sqrt{D/H} \in \mathbb{R}^{D \times D}$ and $\sqrt{A}$ is chosen such that $A = \sqrt{A}\sqrt{A}^\top$, in particular $\sqrt{A} := W^Q/(D/H)^{\frac{1}{4}}$. The terms in the Jacobian of $\tilde{f}$ simplify to:

$$\tilde{J}_{ii} = 2X^\top P^{(i)} XA + P_{ii} I \quad \text{(note } P^{(i)}\mathbb{1} = 0\text{)}, \tag{17}$$

$$\tilde{J}_{ij} = 2P_{ij}(x_j - \sum_k P_{ik} x_k)(x_i - x_j)^\top A + P_{ij} I \quad \text{for } i \neq j. \tag{18}$$

Let the Jacobian of $f(X)$ be:

$$J_f = \begin{bmatrix} J_{11} & \dots & J_{1N} \\ \vdots & \ddots & \vdots \\ J_{N1} & \dots & J_{NN} \end{bmatrix} \in \mathbb{R}^{ND \times ND}. \tag{19}$$

Since $f(X) = \tilde{f}(X)A$, and by the chain rule $\frac{\partial}{\partial x_j}[\tilde{f}_i(X)A] = A^\top \frac{\partial \tilde{f}_i(X)}{\partial x_j} = A\frac{\partial \tilde{f}_i(X)}{\partial x_j}$ (by symmetry of A), we have that $J_{ij} = A\tilde{J}_{ij}$. Hence

$$J_{ii} = 2AX^\top P^{(i)} XA + P_{ii} A \quad \text{(note } P^{(i)}\mathbb{1} = 0\text{)}, \tag{20}$$

$$J_{ij} = 2P_{ij} A(x_j - \sum_k P_{ik} x_k)(x_i - x_j)^\top A + P_{ij} A \quad \text{for } i \neq j. \tag{21}$$

Noting $\mathrm{Lip}_p(f) = \sup_X \|J_f(X)\|_p$, we would like to upper bound $\|J_f\|_p$.

D.2.1. UPPER BOUND ON $\text{Lip}_\infty(F)$ FOR L2-MHA

Consider the choice $p = \infty$, where $\|J_f\|_\infty$ is the maximum absolute row sum of J_f. A key observation is that if we can bound the ∞-norm of the Jacobian of f_i, a single output of f (i.e. a single block row $\|[J_{i1}, ..., J_{iN}]\|_\infty$ of J_f), then this is also a bound on $\|J_f\|_\infty$ due to permutation equivariance of self-attention; all block rows have the same maximal $\|\cdot\|_\infty$ when each is optimised over the input X. Using this, we can prove that $\|J_f\|_\infty$ admits an upper bound that is $O(\log N - \log\log N)$. Below we state and prove lemmas that lead to the proof of this upper bound.

First we analyse the term $\sqrt{A}^\top X^\top P^{(i)} X \sqrt{A}$, that appears in the first term of J_{ii}. Note that for $Y := X\sqrt{A}$, so that the rows of Y are $y_i^\top := x_i^\top \sqrt{A}$, we have

$$\sqrt{A}^\top X^\top P^{(i)} X \sqrt{A} = Y^\top P^{(i)} Y = \text{Cov}(\mathbb{Y}) \tag{22}$$

where $\mathbb{P}(\mathbb{Y} = y_j) = P_{ij} = \exp(-\|y_j - y_i\|_2^2)/\sum_k \exp(-\|y_k - y_i\|_2^2)$. The last equality uses the observation in Equation (9).

The central inequality used throughout the proof of the main theorem is the following:

Lemma D.2. $\text{Tr}(\text{Cov}(\mathbb{Y})) = \sum_j P_{ij}\|y_j - \sum_k P_{ik} y_k\|_2^2 \leq \sum_j P_{ij}\|y_j - y_i\|_2^2 \leq \phi^{-1}(N-1)$ *where* $\phi(c) = c\exp(c+1)$ *is a one-dimensional invertible function on* $\mathbb{R}_{\geq 0}$.

Proof. The first equality holds since $\text{Tr}(\text{Cov}(\mathbb{Y})) = \sum_j \text{Cov}(\mathbb{Y})_{jj} = \sum_j \text{Var}(\mathbb{Y}_j) = \sum_j \mathbb{E}[(\mathbb{Y}_j - \mathbb{E}[\mathbb{Y}_j])^2]$. The next inequality holds since $\text{Var}(\mathbb{Y}_j) = \text{Var}(\overline{\mathbb{Y}}_j) = \mathbb{E}[\overline{\mathbb{Y}}_j^2] - \mathbb{E}[\overline{\mathbb{Y}}_j]^2 \leq \mathbb{E}[\overline{\mathbb{Y}}_j^2]$ where $\overline{\mathbb{Y}} = \mathbb{Y} - y_i$. The final inequality can be proved as follows.

We would like to bound

$$\sum_j P_{ij}\|y_j - y_i\|_2^2 = \frac{\sum_j \|y_j - y_i\|_2^2 \exp(-\|y_j - y_i\|_2^2)}{\sum_k \exp(-\|y_k - y_i\|_2^2)} = \frac{\sum_j z_j^2 \exp(-z_j^2)}{\sum_k \exp(-z_k^2)} \tag{23}$$

where $z_j := \|y_j - y_i\|_2$ (hence $z_i = 0$). Define:

$$g(z) := \frac{\sum_j z_j^2 \exp(-z_j^2)}{\sum_k \exp(-z_k^2)} = \frac{\sum_{j \neq i} z_j^2 \exp(-z_j^2)}{1 + \sum_{k \neq i} \exp(-z_k^2)}. \tag{24}$$

First note that as $z_j \to \infty$, $\exp(-z_j^2) \to 0$ exponentially fast, causing the product $z_j^2 \exp(-z_j^2) \to 0$. Hence we expect the above quantity to be bounded and attain its maximum.

Let $h(z_j) := \exp(-z_j^2)$ for notational conciseness, and note $h(z_j) > 0$. By taking partial derivatives with the chain rule, we have that for $j \neq i$

$$\frac{\partial g(z)}{\partial z_j} = \frac{2 y_j h(z_j)}{(\sum_k h(z_k))^2}\left[(1 - z_j^2)\sum_k h(z_k) + \sum_k h(z_k) z_k^2\right]. \tag{25}$$

Hence the derivative is 0 if and only if $z_j = 0$ or $(1 - z_j^2)\sum_k h(z_k) + \sum_k h(z_k) z_k^2 = 0$, the latter being equivalent to $z_j^2 = 1 + \frac{\sum_k h(z_k) z_k^2}{\sum_k h(z_k)} = 1 + g(z)$. Hence at the maximum, the non-zero values among $\{z_j\}_{j=1}^N$ must be equal to one another. It is clear now that the maximum value c is attained when $z_j^2 = 1 + c$ for $j \neq i$ (and recall $z_i = 0$). So $h(z_j) = \exp(-1 - c)$ for $j \neq i$. Substituting this into $g(z)$, and rearranging, we obtain $c\exp(c+1) = N - 1$. Note $\phi(x) := x\exp(x+1)$ is increasing for $x > 0$ hence $c = \phi^{-1}(N-1)$. □

Note $\phi(\log N) = (\log N)\exp(\log N + 1) \geq N \log N \geq N - 1$ for $N \geq 3$. Since ϕ is increasing, we have $\phi^{-1}(N-1) \leq \log(N)$ for $N \geq 3$. In fact, it is known that $\phi^{-1}(N-1) = O(\log N - \log\log N)$ (Corless et al., 1996).

Note the A term in $f(X) = \tilde{f}(X)A$ allows us to use the above inequality, since $Y^\top P^{(i)} Y = \text{Cov}(\mathbb{Y})$ now appears in the terms of J_f:

$$J_{ii} = 2\sqrt{A}[Y^\top P^{(i)} Y]\sqrt{A}^\top + P_{ii} A, \tag{26}$$

$$J_{ij}, = 2\sqrt{A} P_{ij}(y_j - \sum_k P_{ik} y_k)(y_i - y_j)^\top \sqrt{A}^\top + P_{ij} A \quad \text{for } i \neq j. \tag{27}$$

Using the inequalities $\|BC\| \leq \|B\|\|C\|$, $\|B+C\| \leq \|B\| + \|C\|$ and $\|[A_1, \ldots, A_N]\| \leq \sum_i \|A_i\|$, we have:

$$
\begin{aligned}
&\|[J_{i1}, \ldots, J_{iN}]\|_\infty \\
&\quad \leq \|J_{ii}\|_\infty + \sum_{j \neq i} \|J_{ij}\|_\infty \\
&\quad \leq 2\|\sqrt{A}\|_\infty \|Y^\top P^{(i)} Y\|_\infty \|\sqrt{A}^\top\|_\infty + P_{ii}\|A\|_\infty \\
&\qquad + 2\sum_{j \neq i} \|\sqrt{A}\|_\infty \|P_{ij}(y_j - \sum_k P_{ik} y_k)(y_i - y_j)^\top\|_\infty \|\sqrt{A}^\top\|_\infty + P_{ij}\|A\|_\infty \\
&\quad = 2\|\sqrt{A}\|_\infty \|\sqrt{A}^\top\|_\infty \left(\|Y^\top P^{(i)} Y\|_\infty + \sum_{j \neq i} \|P_{ij}(y_j - \sum_k P_{ik} y_k)(y_i - y_j)^\top\|_\infty \right) + \|A\|_\infty \\
&\quad = 2\frac{\|W^Q\|_\infty \|W^{Q^\top}\|_\infty}{\sqrt{D/H}} \left(\|Y^\top P^{(i)} Y\|_\infty + \sum_j \|P_{ij}(y_j - \sum_k P_{ik} y_k)(y_i - y_j)^\top\|_\infty \right) + \frac{\|W^Q W^{Q^\top}\|_\infty}{\sqrt{D/H}}.
\end{aligned}
$$

For the first equality, note that $\sum_j P_{ij} = 1$. For the second equality, note that the summand for $j = i$ is 0 because the term $y_i - y_j = 0$. Each of the terms in the brackets are bounded by the following lemmas:

Lemma D.3. $\|Y^\top P^{(i)} Y\|_\infty \leq \phi^{-1}(N-1)\sqrt{D/H}$ *(ϕ defined as in Lemma D.2).*

Proof. Recall that $Y^\top P^{(i)} Y = \mathrm{Cov}(\mathbb{Y})$. Let $\sigma(\mathbb{Y}_m)$ denote the standard deviation of $\mathbb{Y}_m$. Then $[\mathrm{Cov}(\mathbb{Y})]_{lm} \leq \sigma(\mathbb{Y}_l)\sigma(\mathbb{Y}_m)$. Hence

$$
\begin{aligned}
\|\mathrm{Cov}(\mathbb{Y})\|_\infty = \max_l \sum_m |[\mathrm{Cov}(\mathbb{Y})]_{lm}| &\leq \max_l \sigma(\mathbb{Y}_l) \sum_m \sigma(\mathbb{Y}_m) \\
&\leq \sqrt{\frac{D}{H}} \sum_m \sigma^2(\mathbb{Y}_m) = \sqrt{\frac{D}{H}} \mathrm{Tr}(\mathrm{Cov}(\mathbb{Y})) \\
&\leq \sqrt{\frac{D}{H}} \phi^{-1}(N-1),
\end{aligned}
$$

since $\sum_m \sigma(\mathbb{Y}_m) \leq \sqrt{\frac{D}{H}} \sqrt{\sum_m \sigma^2(\mathbb{Y}_m)}$ (by e.g. using the Cauchy–Schwartz inequality on $[\sigma(\mathbb{Y}_1), \ldots, \sigma(\mathbb{Y}_{D/H})]$ and $\mathbb{1}$) and $\max_l \sigma(\mathbb{Y}_l) \leq \sqrt{\sum_m \sigma^2(\mathbb{Y}_m)}$, and the last inequality is from Lemma D.2. □

Lemma D.4. $\sum_j \|P_{ij}(y_j - \sum_k P_{ik} y_k)(y_i - y_j)^\top\|_\infty \leq \phi^{-1}(N-1)\sqrt{D/H}$.

Proof. Note $\|ab^\top\|_\infty = \|a\|_\infty \|b\|_1$ for real vectors a, b. Hence

$$
\begin{aligned}
\sum_j \|P_{ij}(y_j - \sum_k P_{ik} y_k)(y_i - y_j)^\top\|_\infty &= \sum_j P_{ij} \|y_j - \sum_k P_{ik} y_k\|_\infty \|y_i - y_j\|_1 \\
&= a^\top b \leq \|a\|_2 \|b\|_2,
\end{aligned}
$$

where $a_j = \sqrt{P_{ij}} \|y_j - \sum_k P_{ik} y_k\|_\infty$, $b_j = \sqrt{P_{ij}} \|y_i - y_j\|_1$.

Note $a_j \leq c_j := \sqrt{P_{ij}} \|y_j - \sum_k P_{ik} y_k\|_2$ since $\|x\|_\infty \leq \|x\|_2$ for vector x. Hence $\|a\|_2 \leq \|c\|_2$. Also $b_j \leq \sqrt{\frac{D}{H}} d_j := \sqrt{\frac{D}{H}} \sqrt{P_{ij}} \|y_i - y_j\|_1$ since $\|x\|_1 \leq \sqrt{\frac{D}{H}} \|x\|_2$ (e.g. by the Cauchy–Schwartz inequality on $[|x_1|, \ldots, |x_{D/H}|]$ and $\mathbb{1}$) for $x \in \mathbb{R}^{D/H}$. Hence $\|b\|_2 \leq \sqrt{\frac{D}{H}} \|d\|_2$.

Note $\|c\|_2^2 = \sum_j P_{ij} \|y_j - \sum_k P_{ik} y_k\|_2^2 = \mathrm{Tr}(\mathrm{Cov}(\mathbb{Y})) \leq \phi^{-1}(N-1)$ from Lemma D.2, and $\|d\|_2^2 = \sum_j P_{ij} \|y_i - y_j\|_2^2 \leq \phi^{-1}(N-1)$ also from Lemma D.2. Hence $\|a\|_2 \|b\|_2 \leq \sqrt{\frac{D}{H}} \|c\|_2 \|d\|_2 \leq \sqrt{\frac{D}{H}} \phi^{-1}(N-1)$. □

Putting the above lemmas altogether, with the observation $\sup_X \|J_f(X)\|_\infty = \sup_X \|[J_{i1}(X), \dots, J_{iN}(X)]\|_\infty$ by permutation invariance of $\|J_f\|_\infty$ (since f is permutation equivariant and $\|\cdot\|_\infty$ is the maximum absolute row sum), we have

$$\begin{aligned} \|J_f\|_\infty &\leq 4\|W^Q\|_\infty \|W^{Q^\top}\|_\infty \phi^{-1}(N-1) + \frac{\|W^Q W^{Q^\top}\|_\infty}{\sqrt{D/H}} \\ &\leq \|W^Q\|_\infty \|W^{Q^\top}\|_\infty \left(4\phi^{-1}(N-1) + \frac{1}{\sqrt{D/H}}\right) \\ &\leq \|W^Q\|_\infty \|W^{Q^\top}\|_\infty \left(4\log N + \frac{1}{\sqrt{D/H}}\right), \end{aligned} \tag{28}$$

where the last inequality holds for $N \geq 3$.

The full multihead attention map that combines the heads $f^h(X)$ is:

$$F : X \mapsto \left[f^1(X)W^{V,1}, \dots f^H(X)W^{V,H}\right] W^O = g(X)W^V W^O$$

where $g : X \mapsto [f^1(X), \dots, f^H(X)]$, $W^O \in \mathbb{R}^{D \times D}$ and

$$W^V = \begin{bmatrix} W^{V,1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & W^{V,H} \end{bmatrix} \in \mathbb{R}^{DH \times D}.$$

Note the Jacobian J_g is a block matrix whose rows are J_{f^h}, hence $\|J_g\|_\infty = \max_h \|J_{f^h}\|_\infty$, and similarly $\|W^{V^\top}\|_\infty = \max_h \|W^{V,h^\top}\|_\infty$. Hence we have

$$\mathrm{Lip}_\infty(F) \leq \max_h \|J_{f^h}\|_\infty \max_h \|W^{V,h^\top}\|_\infty \|W^{O^\top}\|_\infty.$$

Combining this with Inequality (28), we have:

$$\mathrm{Lip}_\infty(F) \leq \left(4\phi^{-1}(N-1) + \frac{1}{\sqrt{D/H}}\right) \max_h \|W^{Q,h}\|_\infty \|W^{Q,h^\top}\|_\infty \max_h \|W^{V,h^\top}\|_\infty \, \|W^{O^\top}\|_\infty.$$

D.2.2. Upper bound on $\mathrm{Lip}_2(F)$ for L2-MHA

For $p = 2$, we use the following lemma:

Lemma D.5. *Let A be a block matrix with block rows $A_1, \dots, A_N$. Then $\|A\|_2 \leq \sqrt{\sum_i \|A_i\|_2^2}$, and equality holds if and only if the first right singular vectors of the A_i align.*

Proof.

$$\|A\|_2^2 = \left\| \begin{bmatrix} A_1 \\ \vdots \\ A_N \end{bmatrix} \right\|_2^2 = \sup_{\|x\|_2=1} \left\| \begin{bmatrix} A_1 \\ \vdots \\ A_N \end{bmatrix} x \right\|_2^2 = \sup_{\|x\|_2=1} \sum_i \|A_i x\|_2^2 \leq \sum_i \sup_{\|x\|_2=1} \|A_i x\|_2^2 = \sum_i \|A_i\|_2^2.$$

Note that equality holds if and only if the first right singular vectors of the A_i align. □

Hence a bound on the spectral norm of each block row of J_f can give us an $O(\sqrt{N})$ bound on $\|J_f\|_2$, which may be loose, and it remains an open question as to whether this bound can be tightened.

To bound the $\|\cdot\|_2$ norm of each row of J_f, we use the following lemmas:

Lemma D.6. $\|Y^\top P^{(i)} Y\|_2 \leq \phi^{-1}(N-1)$

Proof. $\|Y^\top P^{(i)} Y\|_2 = \|\mathrm{Cov}(\mathbb{Y})\|_2 = \lambda_{\max}(\mathrm{Cov}(\mathbb{Y})) \leq \mathrm{Tr}(\mathrm{Cov}(\mathbb{Y})) \leq \phi^{-1}(N-1)$, where the first equality holds by symmetry of $\mathrm{Cov}(\mathbb{Y})$ and the next holds by $\mathrm{Cov}(\mathbb{Y})$ being positive semi-definite, so all its eigenvalues are non-negative, and hence the maximal eigenvalue is bounded by the sum of the eigenvalues, equal to its trace. The final inequality is from Lemma D.2. □

Lemma D.7. $\sum_j \|P_{ij}(y_j - \sum_k P_{ik} y_k)(y_i - y_j)^\top\|_2 \leq \phi^{-1}(N-1)$

Proof. Directly use Cauchy–Schwartz on c and d in the proof of Lemma D.4. □

Again using the inequalities $\|BC\| \leq \|B\|\|C\|$, $\|B+C\| \leq \|B\| + \|C\|$ and $\|[A_1, \ldots, A_N]\| \leq \sum_i \|A_i\|$, with the additional equality $\|B^\top\|_2 = \|B\|_2$, we have the bound:

$$
\begin{aligned}
&\|[J_{i1}, \ldots, J_{iN}]\|_2 \\
&\leq 2\frac{\|W^Q\|_2 \|W^{Q^\top}\|_2}{\sqrt{D/H}} \left(\|Y^\top P^{(i)} Y\|_2 + \sum_j \|P_{ij}(y_j - \sum_k P_{ik} y_k)(y_i - y_j)^\top\|_2 \right) + \frac{\|W^Q W^{Q^\top}\|_2}{\sqrt{D/H}} \\
&\leq 4\phi^{-1}(N-1) \frac{\|W^Q\|_2^2}{\sqrt{D/H}} + \frac{\|W^Q W^{Q^\top}\|_2}{\sqrt{D/H}} \\
&\leq \frac{\|W^Q\|_2^2}{\sqrt{D/H}} \left(4\phi^{-1}(N-1) + 1 \right).
\end{aligned}
$$

Using Lemma D.5, we have that

$$
\begin{aligned}
\|J_f\|_2 &\leq \frac{\sqrt{N}\|W^Q\|_2^2}{\sqrt{D/H}} \left(4\phi^{-1}(N-1) + 1 \right) \\
&\leq \frac{\sqrt{N}\|W^Q\|_2^2}{\sqrt{D/H}} (4 \log N + 1).
\end{aligned} \tag{29}
$$

To obtain the final result for the full multihead self-attention F, we need a final lemma:

Lemma D.8. *Let A be a block matrix with block columns $A_1, \ldots, A_N$. Then $\|A\|_2 \leq \sqrt{\sum_i \|A_i\|_2^2}$.*

Proof.

$$
\begin{aligned}
\|A\|_2 = \|[A_1, \ldots, A_N]\|_2 &= \sup_{\sum_i \|x_i\|_2^2 = 1} \left\| [A_1, \ldots, A_N] \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \right\|_2^2 = \sup_{\sum_i \|x_i\|_2^2 = 1} \|\sum_i A_i x_i\|_2 \\
&\leq \sup_{\sum_i \|x_i\|_2^2 = 1} \sum_i \|A_i x_i\|_2 = \sup_{\|e_i\|_2 = 1, \sum_i \lambda_i^2 = 1} \sum_i \lambda_i \|A_i e_i\|_2 = \sup_{\sum_i \lambda_i^2 = 1} \sum_i \lambda_i \|A_i\|_2 \\
&\leq \sqrt{\sum_i \|A_i\|_2^2},
\end{aligned}
$$

where we are using the substitution $x_i = \lambda_i e_i$, and the last inequality holds by e.g. Cauchy–Schwartz inequality on $[\lambda_1, \ldots, \lambda_N]$ and $[\|A_1\|_2, \ldots, \|A_N\|_2]$. □

Recall that

$$
F : X \mapsto \left[f^1(X) W^{V,1}, \ldots, f^H(X) W^{V,H} \right] W^O.
$$

Since $\|f^h(X) W^{V,h}\|_2 \leq \|J_{f^h}\|_2 \|W^{V,h}\|_2$, by Lemma D.8 we have that

$$
\left\| [f^1(X) W^{V,1}, \ldots, f^H(X) W^{V,H}] \right\|_2 \leq \sqrt{\sum_h \|J_{f^h}\|_2^2 \|W^{V,h}\|_2^2}
$$

and hence

$$\mathrm{Lip}_2(F) \leq \left(\sqrt{\sum_h \|J_{f^h}\|_2^2 \|W^{V,h}\|_2^2} \right) \|W^O\|_2. \tag{30}$$

Combining this with Inequality (29), we have:

$$\mathrm{Lip}_2(F) \leq \frac{\sqrt{N}}{\sqrt{D/H}} \left(4\phi^{-1}(N-1)+1\right) \left(\sqrt{\textstyle\sum_h \|W^{Q,h}\|_2^2 \, \|W^{V,h}\|_2^2} \right) \|W^O\|_2.$$

E. The Case with Masking

Since self-attention is often used with *masking*, a natural question is how masking affects the derived bounds. In self-attention (for any choice of attention function), masking is implemented as follows: given a set of mask indices $\mathcal{M} \subset \{1,\ldots,N\} \times \{1,\ldots,N\}$, the logits (i.e. the inputs to the softmax) are set to $-\infty$ at the mask indices. That is,

$$L_{ij} = \begin{cases} \tilde{L}_{ij} & \text{if } (i,j) \notin \mathcal{M} \\ -\infty & \text{if } (i,j) \in \mathcal{M} \end{cases}$$

where $\tilde{L}_{ij}$ is the original logit (e.g. for L2 self-attention, $\tilde{L}_{ij} = -(x_i - x_j)^\top A(x_i - x_j)$).

Masking implies $f_i(X)$ is not a function of x_j for $(i,j) \in \mathcal{M}$, hence $J_{ij} = 0$ for $(i,j) \in \mathcal{M}$. Thus $f_i(X)$ is equal to the ith output for self-attention with inputs restricted to $\{x_j : (i,j) \notin \mathcal{M}\}$, the unmasked inputs with respect to the ith output. Hence J_{ij} will no longer contribute to the bound on $\|[J_{i1},\ldots,J_{iN}]\|$, and hence the bound for the unmasked case will continue to hold as long as $(i,i) \in \mathcal{M}$ i.e. x_i attends to itself (this is necessary for the proof of Lemma D.2 to hold). The bound can in fact be tightened by replacing N with $|\{x_j : (i,j) \notin \mathcal{M}\}|$, the number of unmasked inputs with respect to the ith output.

F. Asymptotic tightness of the upper bound on the Lipschitz constant

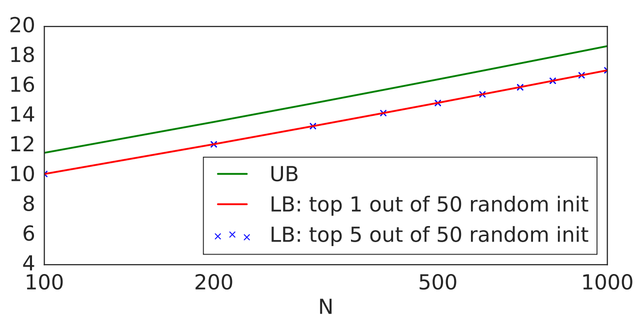

Figure 3. Lower and upper bound on $\mathrm{Lip}_\infty(f)$ for L2-MHA f, with $H = D = 1$ and varying N.

We investigate whether the bound on the Lipschitz constant of `L2-MHA` is tight. The Lipschitz constant is a supremum over the space of inputs $X \in \mathbb{R}^{N \times D}$ and approximating it requires solving an intractable optimisation problem. Hence it is infeasible to estimate accurately in general, especially when X is high-dimensional. However, we may compute a lower bound on the Lipschitz constant by maximising the norm of the Jacobian $\|J_f(X)\|$ with respect to X until convergence. This local optimum will form a lower bound by Theorem 2.1, and we can expect this lower bound to be fairly tight for the low-dimensional case, provided the optimisation is thorough.

We use this observation to provide empirical evidence for the asymptotic tightness of the upper bound on $\mathrm{Lip}_\infty(f)$ in Theorem 3.2. In Figure 3, we show the upper bound as well as the lower bound on $\mathrm{Lip}_\infty(f)$ obtained by optimising $\|J_f(X)\|_\infty$ with respect to X for `L2-MHA` f with 50 different random initialisations of X, with $H = D = 1$ and N varying between 100 and 1000. We fix all free parameters of f (namely W^Q, W^V) to be the identity, and only optimise the input X. We use 50 random initialisations of X for each N, where $X_{ij} \sim \mathcal{U}[-c, c]$ for $c \sim \mathcal{U}[0, 10]$ (we observed that having c

itself be random improves optimisation). We display the top 5 results for each value of N after optimising each random initialisation till convergence using Adam (Kingma & Ba, 2015) with a learning rate of 0.1. Note that we use a log-scale for the x-axis, and recall that the upper bound is $O(\log N - \log\log N)$, dominated by the $O(\log N)$ term for large N. Hence the plot for the upper bound shows a linear trend. We also observe that the slope of the lower bound is very similar, providing empirical evidence that the $O(\log N - \log\log N)$ upper bound is asymptotically tight.

There are at least two possible explanations for the gap between the upper and lower bounds. (1) The lower bound is only a local optimum — the true Lipschitz constant is a global optimum across inputs, which can be difficult to attain especially for high values of N. (2) The multiplicative constant of the upper bound may be loose. Assuming asymptotic tightness, it remains an open question whether the multiplicative constant can be tightened.

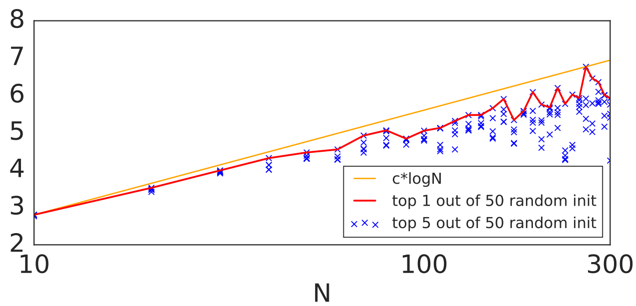

Figure 4. Lower bound on $\mathrm{Lip}_2(f)$ where f is L2-MHA, with $D = 1$ and varying N, obtained by optimising $\|J_f(X)\|_2$ with respect to X, with 50 random initialisations of X for each N.

Similarly, in Figure 4, we show the lower bound on $\mathrm{Lip}_2(f)$ obtained by optimising $\|J_f(X)\|_2$ using the same optimisation procedure as before. Here the optimisation is more difficult, evident in the variance of the top 5 values, and the trend is less clear, but it appears that $\mathrm{Lip}_2(f)$ grows at a rate of $O(\log N)$. The message is less clear here, and there are at least two possibilities. (1) The optimisation is difficult even for small values of N, hence Figure 4 shows a loose lower bound. (2) If the lower bound is tight, this suggests that the $O(\sqrt{N}\log N)$ bound in Theorem 3.2 is not asymptotically tight, and could be improved to $O(\log N)$ (or $O(\log N - \log\log N)$ as for $p = \infty$).

G. Experimental Details

For the experiments in Section 5, we compare the performance of the original Transformer and the Transformer with Lipschitz/invertible self-attention at character-level language modelling on the Penn Treebank dataset (Marcus et al., 1993).[1] Each training example is a sentence represented as a variable-length sequence of characters, and examples are batched according to length such that padding is minimised, with the maximum sequence length set to 288. All models are autoregressive, outputting the logits for the categorical likelihood predicting the next character, and are trained using maximum likelihood (cross-entropy loss) with a batch size of 64. The LSTM models have the dimensionality of the hidden state equal to the dimensionality D of the cell state (the usual default implementation). The Transformer models are trained with a varying number of blocks (number of layers) with $H = 8$ heads and $D = 512$, tuning hyperparameters for dropout rate in $\{0, 0.1, 0.2\}$ and base learning rate $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0\}$ with number of warmup iterations $w \in \{1000, 2000, 4000, 8000\}$ for the standard custom learning rate schedule in Vaswani et al. (2017):

$$\epsilon_t = \frac{\gamma}{\sqrt{D}} \min(t^{-1/2}, tw^{-3/2}),$$

where ϵ_t is the learning rate at training iteration t. Hence the learning rate linearly increases from 0 to $(Dw)^{-1/2}$ over w iterations, then decays proportionally to $t^{-1/2}$.

[1] We use the standard training-validation split, and the dataset can be found at e.g. `https://github.com/harvardnlp/TextFlow/tree/master/data/ptb`.

H. Numerical Invertibility of MHA Residual Map

As a sanity check, in Figure 5 we compare the numerical invertibility of the residual map $g(x) = x + cf(x)$ between the cases where f is `L2-MHA` and `DP-MHA`. For each, we take `MHA` with 8 heads and randomly initialised weights, and quantify the maximum reconstruction error across a batch of 128 inputs whose outputs are inverted via the fixed-point iteration described in Section 4.1. We use $N \in \{64, 128\}$, $D \in \{64, 128\}$, and $c \in \{0.5, 0.7, 0.9\}$ To highlight the difference between the two types of self-attention, recall in the proof of Theorem 3.1 (showing that `DP-MHA` is not Lipschitz) that when one of the inputs x_i is 0, some terms of the Jacobian grow with the sample variance of $x_{\neq i}$. Hence we check numerical invertibility at a set of N inputs where $x_i = 0$ and $x_{\neq i}$ are chosen uniformly at random. In this case, we expect `DP-MHA` to be non-invertible whereas `L2-MHA` will be invertible for sufficiently small c. This is precisely what we observe in Figure 5. We note that the figure shows local invertibility at the sampled inputs, and not global invertibility across the whole input space, yet this clearly highlights the difference between the two choices of self-attention.

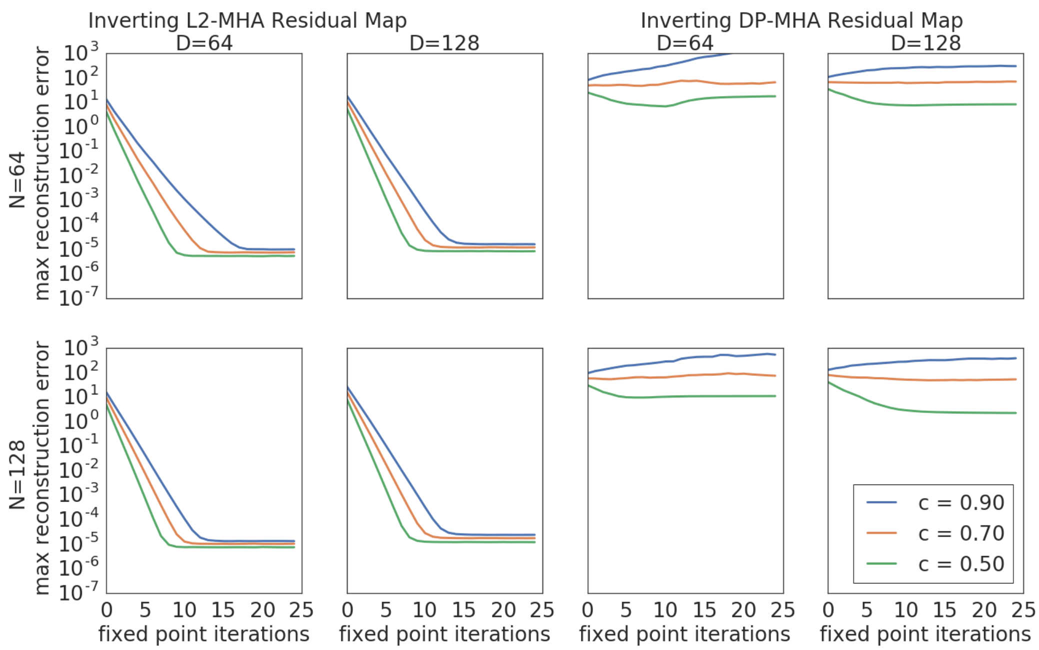

Figure 5. Numerical invertibility of $g(x) = x + cf(x)$ where f is L2-MHA (left) or DP-MHA (right), for different values of N and D.