Appendix

A. Useful Identities for deriving Jacobian expressions

In this section, we list some useful identities for deriving the Jacobians of the expressions in the paper.

Suppose λ is a scalar, $\mathbf{u}, \mathbf{v}, \mathbf{x} \in \mathbb{R}^{n \times 1}$ are column vectors, and $f(\mathbf{u})$ is a vector valued function. We use the standard convention that for $\mathbf{a} \in \mathbb{R}^m$, $\mathbf{b} \in \mathbb{R}^n$, we have $\frac{\partial \mathbf{a}}{\partial \mathbf{b}} \in \mathbb{R}^{m \times n}$. Then we have the following chain rule identities:

- $\frac{\partial}{\partial \mathbf{x}}[\lambda \mathbf{u}] = \lambda \frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{x} \frac{\partial \lambda}{\partial \mathbf{x}}$
- $\frac{\partial f(\mathbf{u})}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$
- $\frac{\partial}{\partial \mathbf{x}}[\mathbf{u}^\top \mathbf{v}] = \mathbf{u}^\top \frac{\partial \mathbf{v}}{\partial \mathbf{x}} + \mathbf{v}^\top \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$

Note $\frac{\partial \lambda}{\partial \mathbf{x}}$ is a row vector, so $\mathbf{u} \frac{\partial \lambda}{\partial \mathbf{x}}$ is a matrix.

The Jacobian of the softmax is also well-known. Suppose $\mathbf{v} = \operatorname{softmax}(\mathbf{u}) \in \mathbb{R}^{n \times 1}$. Then

$$\frac{\partial \mathbf{v}}{\partial \mathbf{u}} = \operatorname{diag}(\mathbf{v}) - \mathbf{v}\mathbf{v}^\top = \begin{bmatrix} v_1(1-v_1) & -v_1 v_2 & \dots & -v_1 v_n \\ -v_2 v_1 & v_2(1-v_2) & \dots & -v_2 v_n \\ \vdots & \vdots & \ddots & \vdots \\ -v_n v_1 & -v_n v_2 & \dots & v_n(1-v_n) \end{bmatrix}.$$

B. 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*:

$$b_{k+1} = \frac{W^\top W b_k}{\|W^\top W b_k\|_2}, \quad \tilde{\sigma}_k(W) = \sqrt{\frac{b_k^\top W^\top W b_k}{b_k^\top b_k}}, \tag{9}$$

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).

C. Proof of Theorem 3.1 for General D

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

Proof. The mapping f can be written as

$$f(X) = PX = \operatorname{softmax}\left(XA^\top X^\top\right) X = \begin{bmatrix} f_1(X)^\top \\ \vdots \\ f_N(X)^\top \end{bmatrix} \in \mathbb{R}^{N \times D}, \tag{10}$$

where $A = W^K W^{Q^\top} / \sqrt{D/H} \in \mathbb{R}^{D \times D}$ and $f_i(X) = \sum_{j=1}^N P_{ij} \mathbf{x}_j$ with $P_{i:}^\top = \operatorname{softmax}(XA\mathbf{x}_i)$. Hence f can be interpreted as a map of each $\mathbf{x}_i$ to a point in the convex hull of $\mathbf{x}_1, ..., \mathbf{x}_N$. Since f is a map from $\mathbb{R}^{N \times D}$ to $\mathbb{R}^{N \times D}$, its Jacobian is

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

where $J_{ij} = \frac{\partial f_i(X)}{\partial \mathbf{x}_j} \in \mathbb{R}^{D \times D}$. By taking partial derivatives we can show that $J_{ij} = X^\top P^{(i)} \left[E_{ji} X A^\top + X A \delta_{ij} \right] + P_{ij} I$ 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 $i = j$:

$$\begin{aligned} J_{ii} &= X^\top P^{(i)} E_{ii} X A^\top + X^\top P^{(i)} X A + P_{ii} I \\ &= P_{ii} \left(\mathbf{x}_i - \textstyle\sum_k P_{ik} \mathbf{x}_k \right) \mathbf{x}_i^\top A^\top + X^\top P^{(i)} X A + P_{ii} I. \end{aligned} \tag{12}$$

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

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:

$$[X^\top P^{(i)} X]_{lm} = \textstyle\sum_k P_{ik} x_{kl} x_{km} - \left(\sum_k P_{ik} x_{kl} \right) \left(\sum_k P_{ik} x_{km} \right) = \mathrm{Cov}(\mathbb{X}_l, \mathbb{X}_m), \tag{13}$$

where $\mathbb{X}$ is a discrete distribution with support at the inputs $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and probability mass function given by their softmax probabilities $\mathbb{P}(\mathbb{X} = \mathbf{x}_j) = P_{ij}$. A consequence of this interpretation is that $P^{(i)}$ is *positive semi-definite* (PSD) since for $D = 1$, Equation (13) becomes $X^\top P^{(i)} X = \mathrm{Var}(\mathbb{X}) \geq 0$, with equality if and only if the $\mathbf{x}_j$ 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 $\mathbf{x}_i = 0$. Then $P_{i:}^\top = \mathrm{softmax}\left(X A \mathbf{x}_i \right) = \frac{1}{N} \mathbb{1}$, i.e. we have uniform attention regardless of $\mathbf{x}_{\neq i}$. The first term of J_{ii} in Equation (12) disappears since $\mathbf{x}_i = \mathbf{0}$, 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.

Note that we have shown that single head dot-product self-atttention ($H = 1$) is not Lipschitz, but it is clear that this implies multihead self-attention `DP-MHA` is also not Lipschitz, since the output of multihead attention is a linear combination of the outputs of each head. □

D. Bias term in DP Self-Attention

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

E. 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

$$P = \mathrm{softmax}\left(- \frac{\|XW^Q\|_{\text{row}}^2 \mathbb{1}^\top - 2XW^Q (XW^K)^\top + \mathbb{1} \|XW^K\|_{\text{row}}^{2\top}}{\sqrt{D/H}} \right), \tag{14}$$

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 $\|A\|_{\text{row}}^2 \in \mathbb{R}^m$.

In Table 2 we show the wall-clock training times for the Transformer models with different attention functions and a varying number of layers. It is evident that the differences between the models are rather small.

	1 Layer	2 Layers	3 Layers	4 Layers	5 Layers
Transformer **(DP)**	37	56	77	92	110
Transformer **(L2)**	35	56	73	99	115
Transformer, $W^Q = W^K$ **(L2)**	39	58	79	91	108
Transformer, **(Contractive-L2)**	37	60	81	102	127

Table 2. Wall clock training times for one epoch of training (seconds)

F. Proof of Theorem 3.2

Recall the formulation of `L2-MHA`:

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

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 (14):

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

F.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 $\tilde{f}(X) := PX$, 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:

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

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:

$$J_{\tilde{f}} = \begin{bmatrix} \tilde{J}_{11} & \dots & \tilde{J}_{1N} \\ \vdots & \ddots & \vdots \\ \tilde{J}_{N1} & \dots & \tilde{J}_{NN} \end{bmatrix} \in \mathbb{R}^{ND\times ND} \tag{15}$$

with

$$\tilde{J}_{ij} = X^\top P^{(i)} \frac{\partial L_{i:}}{\partial x_j} + P_{ij} I \in \mathbb{R}^{D\times D} \tag{16}$$

where

$$\frac{\partial L_{i:}}{\partial \mathbf{x}_j} = \frac{2}{\sqrt{D/H}} \left[\left(XW^K - \mathbb{1}\mathbf{x}_i^\top W^Q\right) W^{Q^\top} \delta_{ij} + \left(E_{ji} X W^Q - E_{jj} X W^K\right) W^{K^\top}\right] \tag{17}$$

and

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

$$P_{ij} = \frac{\exp\left(-\|\mathbf{x}_i^\top W^Q - \mathbf{x}_j^\top W^K\|_2^2\right)}{\sum_k \exp\left(-\|\mathbf{x}_i^\top W^Q - \mathbf{x}_k^\top W^K\|_2^2\right)}.$$

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 $\mathbf{x}_i^\top$. We can then verify:

$$X^\top P^{(i)} E_{ji} X = P_{ij}(\mathbf{x}_j - \sum_k P_{ik}\mathbf{x}_k)\mathbf{x}_i^\top. \tag{18}$$

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:

$$\begin{aligned}
\tilde{J}_{ii} &= \frac{2}{\sqrt{D/H}}\left[X^\top P^{(i)}(XW^K - \mathbb{1}\mathbf{x}_i^T W^Q)W^{Q^\top} + X^\top P^{(i)} E_{ii} X (W^Q - W^K) W^{K^\top}\right] + P_{ii} I \\
&= \frac{2}{\sqrt{D/H}}\left[X^\top P^{(i)}(XW^K - \mathbb{1}\mathbf{x}_i^T W^Q)W^{Q^\top} + P_{ii}(\mathbf{x}_i - \sum_k P_{ik}\mathbf{x}_k)\mathbf{x}_i^\top (W^Q - W^K) W^{K^\top}\right] + P_{ii} I \\
&= \frac{2}{\sqrt{D/H}}\left[X^\top P^{(i)} X W^K W^{Q^\top} + P_{ii}(\mathbf{x}_i - \sum_k P_{ik}\mathbf{x}_k)\mathbf{x}_i^\top (W^Q - W^K) W^{K^\top}\right] + P_{ii} I,
\end{aligned} \tag{19}$$

and for $i \neq j$:

$$\begin{aligned}
\tilde{J}_{ij} &= \frac{2}{\sqrt{D/H}} X^\top P^{(i)}(E_{ij} X W^Q - E_{jj} X W^K) W^{K^\top} + P_{ij} I \\
&= \frac{2}{\sqrt{D/H}} P_{ij}(\mathbf{x}_j - \sum_k P_{ik}\mathbf{x}_k)(\mathbf{x}_i^\top W^Q - \mathbf{x}_j^\top W^K) W^{K^\top} + P_{ij} I.
\end{aligned} \tag{20}$$

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

Lemma F.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 $\tilde{K}_{ij} := P_{ij} W^{K^\top}(\mathbf{x}_j - \sum_k P_{ik}\mathbf{x}_k)(\mathbf{x}_i^\top W^Q - \mathbf{x}_j^\top W^K) \in \mathbb{R}^{\frac{D}{H} \times \frac{D}{H}}$ for $i \neq j$, which is related to $\tilde{J}_{ij}$ as follows by Equation (20):

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

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 $\mathbf{y}_j^\top = \mathbf{x}_i^\top W^Q - \mathbf{x}_j^\top W^K$. Then we have:

$$\begin{aligned}
\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k &= W^{Q^\top}\mathbf{x}_i - W^{K^\top}\mathbf{x}_j - \sum_k P_{ik}(W^{Q^\top}\mathbf{x}_i - W^{K^\top}\mathbf{x}_k) \\
&= W^{Q^\top}\mathbf{x}_i - W^{K^\top}\mathbf{x}_j - (W^{Q^\top}\mathbf{x}_i - \sum_k P_{ik} W^{K^\top}\mathbf{x}_k) \\
&= -W^{K^\top}(\mathbf{x}_j - \sum_k P_{ik}\mathbf{x}_k).
\end{aligned}$$

Hence $\tilde{K}_{ij} = -P_{ij}(\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k)\mathbf{y}_j^\top$. Note $\mathbf{y}_i$ 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 $\mathbf{x}_j$ such that $\mathbf{y}_j = -\mathbf{y}_i$. This is possible for any value of $\mathbf{y}_i$ since W^K is full-rank. Note $\mathbf{y}_j = -\mathbf{y}_i$ and not $\mathbf{y}_i$. We then have that $\|\mathbf{y}_j\|_2^2$ is equal for all j, hence $P_{ij} := \frac{\exp(-\|\mathbf{y}_j\|_2^2)}{\sum_k \exp(-\|\mathbf{y}_k\|_2^2)} = \frac{1}{N}$ for all j. Then for $i \neq j$, $\tilde{K}_{ij}$ simplifies to

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

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

The intuition for this result is as follows: a reason for DP-MHA not being Lipschitz is that for $\mathbf{x}_i = 0$,, the attention weights P_{ij} become uniform regardless of the values of $\mathbf{x}_j$ for $j \neq i$. A similar issue arises for L2-MHA with $W^Q \neq W^K$ and full-rank W^K, as shown above: given any $\mathbf{x}_i$, we can choose $\mathbf{x}_j$ such that the P_{ij} become uniform.

F.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

$$P_{ij} \propto \exp\left(-\|(\mathbf{x}_i - \mathbf{x}_j)^\top \sqrt{A}\|_2^2\right) \tag{21}$$

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:

$$\tilde{J}_{ii} = 2X^\top P^{(i)} XA + P_{ii}I \;\; (\text{note } P^{(i)}\mathbb{1} = 0), \tag{22}$$

$$\tilde{J}_{ij} = 2P_{ij}(\mathbf{x}_j - \sum_k P_{ik}\mathbf{x}_k)(\mathbf{x}_i - \mathbf{x}_j)^\top A + P_{ij}I \;\; \text{for } i \neq j. \tag{23}$$

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

$$J_f = \begin{bmatrix} J_{11} & \dots & J_{1N} \\ \vdots & \ddots & \vdots \\ J_{N1} & \dots & J_{NN} \end{bmatrix} \in \mathbb{R}^{ND\times ND}. \tag{24}$$

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

$$J_{ii} = 2AX^\top P^{(i)} XA + P_{ii}A \;\; (\text{note } P^{(i)}\mathbb{1} = \mathbf{0}), \tag{25}$$

$$J_{ij} = 2P_{ij}A(\mathbf{x}_j - \sum_k P_{ik}\mathbf{x}_k)(\mathbf{x}_i - \mathbf{x}_j)^\top A + P_{ij}A \;\; \text{for } i \neq j. \tag{26}$$

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

F.2.1. Upper bound on $\mathbf{Lip}_\infty(F)$ 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 $\mathbf{y}_i^\top := \mathbf{x}_i^\top\sqrt{A}$, we have

$$\sqrt{A}^\top X^\top P^{(i)} X\sqrt{A} = Y^\top P^{(i)} Y = \mathrm{Cov}(\mathbb{Y}) \tag{27}$$

where $\mathbb{P}(\mathbb{Y} = \mathbf{y}_j) = P_{ij} = \exp(-\|\mathbf{y}_j - \mathbf{y}_i\|_2^2) / \sum_k \exp(-\|\mathbf{y}_k - \mathbf{y}_i\|_2^2)$. The last equality uses the observation in Equation (7).

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

Lemma F.2. $\mathrm{Tr}(\mathrm{Cov}(\mathbb{Y})) = \sum_j P_{ij} \|\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k\|_2^2 \leq \sum_j P_{ij} \|\mathbf{y}_j - \mathbf{y}_i\|_2^2 \leq \phi^{-1}(N-1)$ *where* $\phi(c) = c\exp(c+1)$ *is a one-dimensional invertible function on* $\mathbb{R}_{\geq 0}$.

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

We would like to bound

$$\sum_j P_{ij}\|\mathbf{y}_j - \mathbf{y}_i\|_2^2 = \frac{\sum_j \|\mathbf{y}_j - \mathbf{y}_i\|_2^2 \exp(-\|\mathbf{y}_j - \mathbf{y}_i\|_2^2)}{\sum_k \exp(-\|\mathbf{y}_k - \mathbf{y}_i\|_2^2)} = \frac{\sum_j z_j^2 \exp(-z_j^2)}{\sum_k \exp(-z_k^2)} \tag{28}$$

where $z_j := \|\mathbf{y}_j - \mathbf{y}_i\|_2$ (hence $z_i = 0$). Define:

$$g(\mathbf{z}) := \frac{\sum_j z_j^2 \exp(-z_j^2)}{\sum_k \exp(-z_k^2)} = \frac{\sum_{j \neq i} z_j^2 \exp(-z_j^2)}{1 + \sum_{k \neq i} \exp(-z_k^2)}. \tag{29}$$

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 $j \neq i$

$$\frac{\partial g(\mathbf{z})}{\partial z_j} = \frac{2z_j h(z_j)}{(\sum_k h(z_k))^2}\left[(1 - z_j^2)\sum_k h(z_k) + \sum_k h(z_k) z_k^2\right]. \tag{30}$$

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 $z_j^2 = 1 + \frac{\sum_k h(z_k) z_k^2}{\sum_k h(z_k)} = 1 + g(\mathbf{z})$. 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 $Y^\top P^{(i)} Y = \mathrm{Cov}(\mathbb{Y})$ now appears in the terms of J_f:

$$J_{ii} = 2\sqrt{A}[Y^\top P^{(i)} Y]\sqrt{A}^\top + P_{ii}A, \tag{31}$$

$$J_{ij}, = 2\sqrt{A}P_{ij}(\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k)(\mathbf{y}_i - \mathbf{y}_j)^\top \sqrt{A}^\top + P_{ij}A \quad \text{for } i \neq j. \tag{32}$$

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

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

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 $\mathbf{y}_i - \mathbf{y}_j = \mathbf{0}$. Each of the terms in the brackets are bounded by the following lemmas:

Lemma F.3. $\|Y^\top P^{(i)} Y\|_\infty \leq \phi^{-1}(N-1)\sqrt{D/H}$ *(ϕ defined as in Lemma F.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 $[\sigma(\mathbb{Y}_1), \dots, \sigma(\mathbb{Y}_{D/H})]$ 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 F.2. □

Lemma F.4. $\sum_j \|P_{ij}(\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k)(\mathbf{y}_i - \mathbf{y}_j)^\top\|_\infty \leq \phi^{-1}(N-1)\sqrt{D/H}$.

Proof. Note $\|\mathbf{u}\mathbf{v}^\top\|_\infty = \|\mathbf{u}\|_\infty \|\mathbf{v}\|_1$ for real vectors $\mathbf{u}, \mathbf{v}$. Hence

$$
\begin{aligned}
\sum_j \|P_{ij}(\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k)(\mathbf{y}_i - \mathbf{y}_j)^\top\|_\infty &= \sum_j P_{ij}\|\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k\|_\infty \|\mathbf{y}_i - \mathbf{y}_j\|_1 \\
&= \mathbf{a}^\top \mathbf{b} \leq \|\mathbf{a}\|_2 \|\mathbf{b}\|_2,
\end{aligned}
$$

where $a_j = \sqrt{P_{ij}}\|\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k\|_\infty$, $b_j = \sqrt{P_{ij}}\|\mathbf{y}_i - \mathbf{y}_j\|_1$.

Note $a_j \leq c_j := \sqrt{P_{ij}}\|\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k\|_2$ since $\|\mathbf{u}\|_\infty \leq \|\mathbf{u}\|_2$ for vector $\mathbf{u}$. Hence $\|\mathbf{a}\|_2 \leq \|\mathbf{c}\|_2$.

Also $b_j \leq \sqrt{\frac{D}{H}} d_j := \sqrt{\frac{D}{H}}\sqrt{P_{ij}}\|\mathbf{y}_i - \mathbf{y}_j\|_2$ since $\|\mathbf{u}\|_1 \leq \sqrt{\frac{D}{H}}\|\mathbf{u}\|_2$ for $\mathbf{u} \in \mathbb{R}^{D/H}$ (e.g. by the Cauchy–Schwartz inequality on $[|\mathbf{u}_1|, \dots, |\mathbf{u}_{D/H}|]$ and $\mathbb{1}$). 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 F.2, and $\|d\|_2^2 = \sum_j P_{ij}\|y_i - y_j\|_2^2 \leq \phi^{-1}(N-1)$ also from Lemma F.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 $\sup_X \|J_f(X)\|_\infty = \sup_X \|[J_{i1}(X), \ldots, J_{iN}(X)]\|_\infty$ by permutation invariance of $\|J_f\|_\infty$ (since f is permutation equivariant and $\|\cdot\|_\infty$ is the maximum absolute row sum), we have

$$\begin{aligned}
\|J_f\|_\infty &\leq 4\|W^Q\|_\infty \|W^{Q^\top}\|_\infty \phi^{-1}(N-1) + \frac{\|W^Q W^{Q^\top}\|_\infty}{\sqrt{D/H}} \\
&\leq \|W^Q\|_\infty \|W^{Q^\top}\|_\infty \left(4\phi^{-1}(N-1) + \frac{1}{\sqrt{D/H}}\right) \\
&\leq \|W^Q\|_\infty \|W^{Q^\top}\|_\infty \left(4\log N + \frac{1}{\sqrt{D/H}}\right),
\end{aligned} \tag{33}$$

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

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

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

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

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

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 (33), we have:

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

F.2.2. Upper bound on $\mathbf{Lip_2}(F)$ for L2-MHA

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

Lemma F.5. *Let A be a block matrix with block rows $A_1, \ldots, A_N$. 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.

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

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

Proof. $\|Y^\top P^{(i)} Y\|_2 = \|\text{Cov}(\mathbb{Y})\|_2 = \lambda_{\max}(\text{Cov}(\mathbb{Y})) \leq \text{Tr}(\text{Cov}(\mathbb{Y})) \leq \phi^{-1}(N-1)$, where the first equality holds by symmetry of $\text{Cov}(\mathbb{Y})$ and the next holds by $\text{Cov}(\mathbb{Y})$ 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 F.2. □

Lemma F.7. $\sum_j \|P_{ij}(\mathbf{y}_j - \sum_k P_{ik}\mathbf{y}_k)(\mathbf{y}_i - \mathbf{y}_j)^\top\|_2 \leq \phi^{-1}(N-1)$

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

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

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

Using Lemma F.5, we have that

$$\begin{aligned}
\|J_f\|_2 &\leq \frac{\sqrt{N}\|W^Q\|_2^2}{\sqrt{D/H}}\left(4\phi^{-1}(N-1) + 1\right) \\
&\leq \frac{\sqrt{N}\|W^Q\|_2^2}{\sqrt{D/H}}(4\log N + 1).
\end{aligned} \tag{34}$$

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

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

Proof.

$$\begin{aligned}
\|A\|_2 = \|[A_1, \dots, A_N]\|_2 &= \sup_{\sum_i \|\mathbf{x}_i\|_2^2 = 1} \left\| [A_1, \dots, A_N] \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_N \end{bmatrix} \right\|_2^2 = \sup_{\sum_i \|\mathbf{x}_i\|_2^2 = 1} \|\sum_i A_i \mathbf{x}_i\|_2 \\
&\leq \sup_{\sum_i \|\mathbf{x}_i\|_2^2 = 1} \sum_i \|A_i \mathbf{x}_i\|_2 = \sup_{\|\mathbf{e}_i\|_2 = 1, \sum_i \lambda_i^2 = 1} \sum_i \lambda_i \|A_i \mathbf{e}_i\|_2 = \sup_{\sum_i \lambda_i^2 = 1} \sum_i \lambda_i \|A_i\|_2 \\
&\leq \sqrt{\sum_i \|A_i\|_2^2},
\end{aligned}$$

where we are using the substitution $\mathbf{x}_i = \lambda_i \mathbf{e}_i$, and the last inequality holds by e.g. Cauchy–Schwartz inequality on $[\lambda_1, \dots, \lambda_N]$ and $[\|A_1\|_2, \dots, \|A_N\|_2]$. □

Recall that

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

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

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

and hence

$$\mathrm{Lip}_2(F) \leq \left(\sqrt{\sum_h \|J_{f^h}\|_2^2 \|W^{V,h}\|_2^2} \right) \|W^O\|_2. \tag{35}$$

Combining this with Inequality (34), we have:

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

G. 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 $\mathcal{M} \subset \{1,\dots,N\} \times \{1,\dots,N\}$, 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, $\tilde{L}_{ij} = -(\mathbf{x}_i - \mathbf{x}_j)^\top A(\mathbf{x}_i - \mathbf{x}_j)$).

Masking implies $f_i(X)$ is not a function of $\mathbf{x}_j$ 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 $\{\mathbf{x}_j : (i,j) \notin \mathcal{M}\}$, the unmasked inputs with respect to the ith output. Hence J_{ij} will no longer contribute to the bound on $\|[J_{i1},\dots,J_{iN}]\|$, and hence the bound for the unmasked case will continue to hold as long as $(i,i) \in \mathcal{M}$ i.e. $\mathbf{x}_i$ attends to itself (this is necessary for the proof of Lemma F.2 to hold). The bound can in fact be tightened by replacing N with $|\{\mathbf{x}_j : (i,j) \notin \mathcal{M}\}|$, the number of unmasked inputs with respect to the ith output.

H. Experimental Details

For the experiment in Section 5.1, showing the asymptotic tightness of the upper bound on $\mathrm{Lip}_\infty(F)$ where F is L2-MHA, 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.

For the experiments in Section 5.3, we comparing 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 $t^{-1/2}$. We use Glorot Uniform initialisation (Glorot & Bengio, 2010) for all weights ($U\left[-\sqrt{\frac{1}{d_{in}+d_{out}}}, \sqrt{\frac{1}{d_{in}+d_{out}}}\right]$), except for weights in L2-MHA that are initialised from $U\left[-\frac{s}{\sqrt{D}}, \frac{s}{\sqrt{D}}\right]$, and s is a hyperparameter. For $D = 512$, we used $s = \frac{1}{2^4}$. All experiments were done in Tensorflow 1.14 (Abadi et al., 2016) on single Nvidia Tesla V100 GPUs.

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

I. Numerical Invertibility of MHA Residual Map

Following Section 5.2, Figure 6 confirms that numerical invertibility does not hold for trained weights for dot-product multihead self-attention (DP-MHA) (obtained from one-layer Transformer (DP) model used for Figure 4), similar to the randomly initialised weight case. Figure 7 shows additional results for different values of N and D.

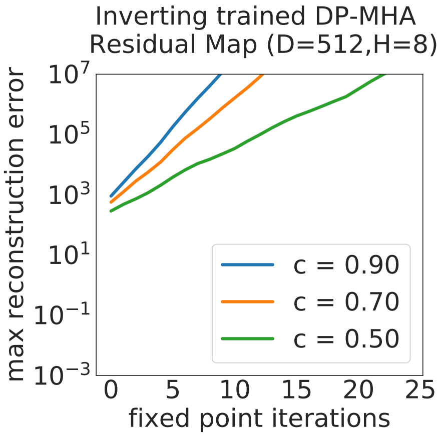

Figure 6. Invertibility of $g(\mathbf{x}) = \mathbf{x} + cf(\mathbf{x})$ for trained DP-MHA f.

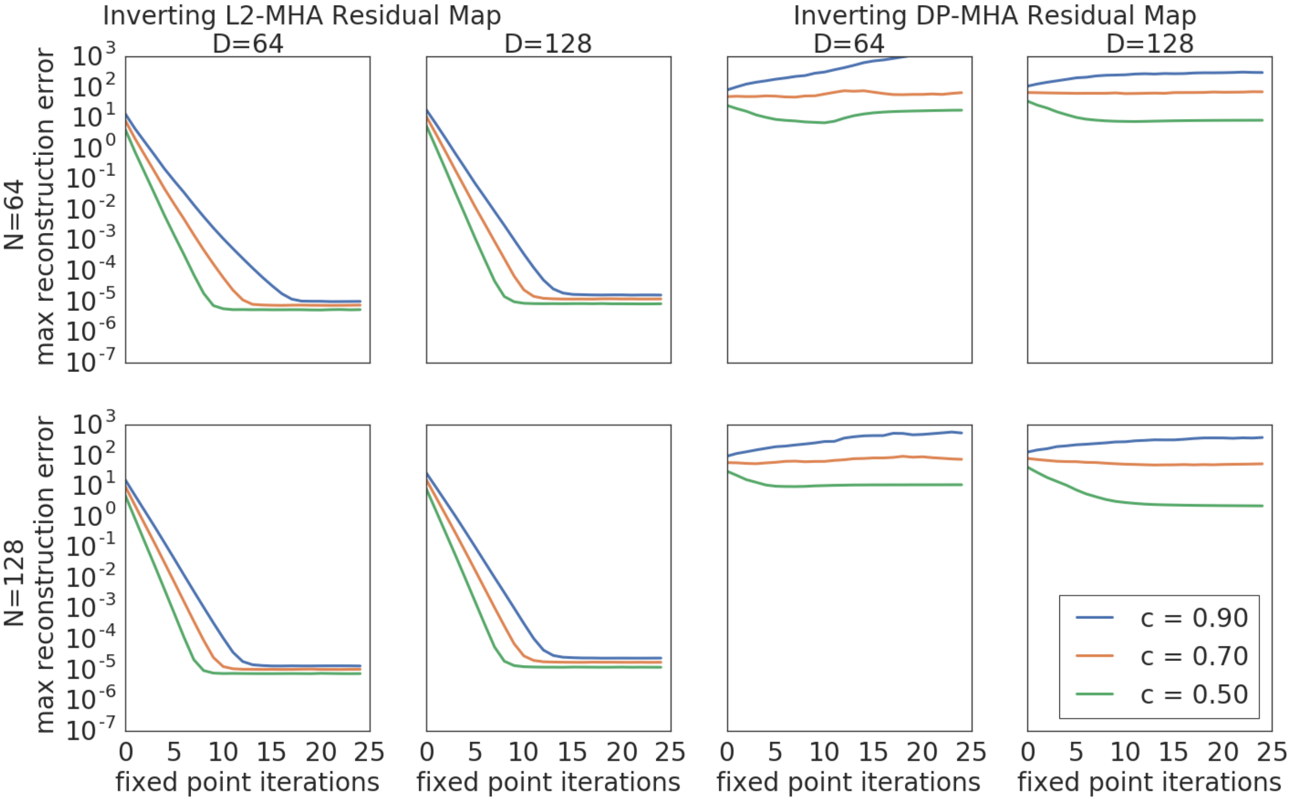

Figure 7. Numerical invertibility of $g(\mathbf{x}) = \mathbf{x} + cf(\mathbf{x})$ where f is L2-MHA(left) or DP-MHA (right), for different values of N and D.

J. Behaviour of Lower Bound on $\mathrm{Lip}_2(F)$

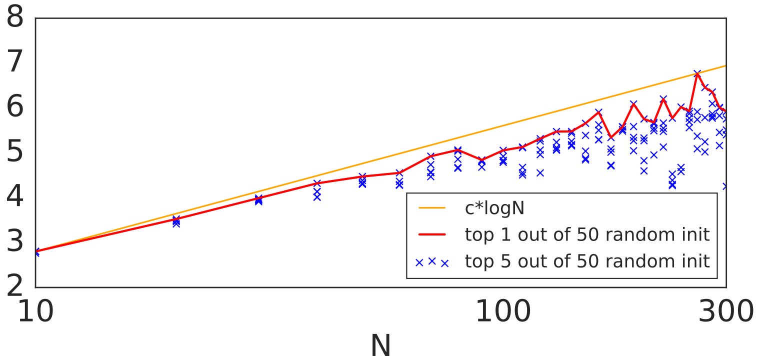

Figure 8. Lower bound on $\mathrm{Lip}_2(F)$ where F is L2-MHA, with $D = 1$ and varying N, obtained by optimising $\|J_F(X)\|_2$ with respect to X, with 50 random initialisations of X for each N.

In Figure 8, we show the lower bound on $\mathrm{Lip}_2(F)$ obtained by optimising $\|J_F(X)\|_2$ using the same optimisation procedure as for Figure 2 of Section 5.1. 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 8 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 $p = \infty$).

K. Optimising the norm of the Jacobian of DP-MHA

In Figure 9, we show how the norm of the Jacobian $\|J_f(X)\|_\infty$ for DP−MHA f keeps increasing when being optimised with respect to X. This is a useful sanity check validating our theoretical result of Theorem 3.1, that DP−MHA is *not* Lipshchitz. The oscillations are likely due to momentum term of Adam optimizer that was used to optimise the norm.

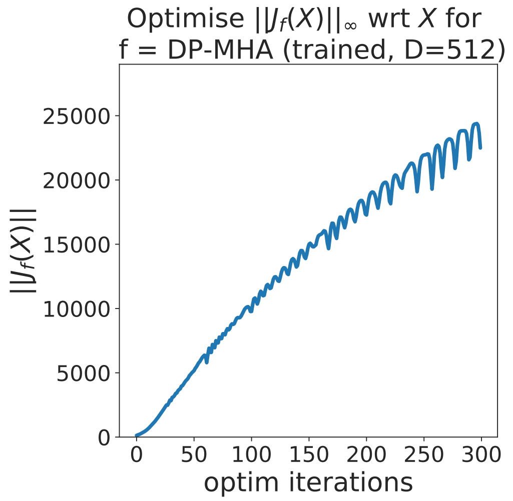

Figure 9. Optimise $\|J_f(X)\|_\infty$ w.r.t. X for trained DP-MHA f.

L. Experiment tying keys and queries of L2-MHA but preserving parameter count

In Figure 4 of Section 5.3, we have shown that there is a clear reduction in performance when tying the keys and queries. To test whether this can be attributed to the reduction in parameter count, we tried doubling the number of columns of W^Q when the keys and queries are shared (i.e. from D/H to $2D/H$) so that the shared model has the same number of parameters as the unshared model. In Figure 10, the third column shows results for shared `L2-MHA`, but with the same number of parameters as the unshared `L2-MHA` i.e. without tying the keys and queries. The performance is similar to the second column (tying with a reduced number of parameters), suggesting that there is an inherent limitation in expressiveness to tying the keys and queries, and that the reduction in number of parameters is an insufficient explanation this phenomenon.

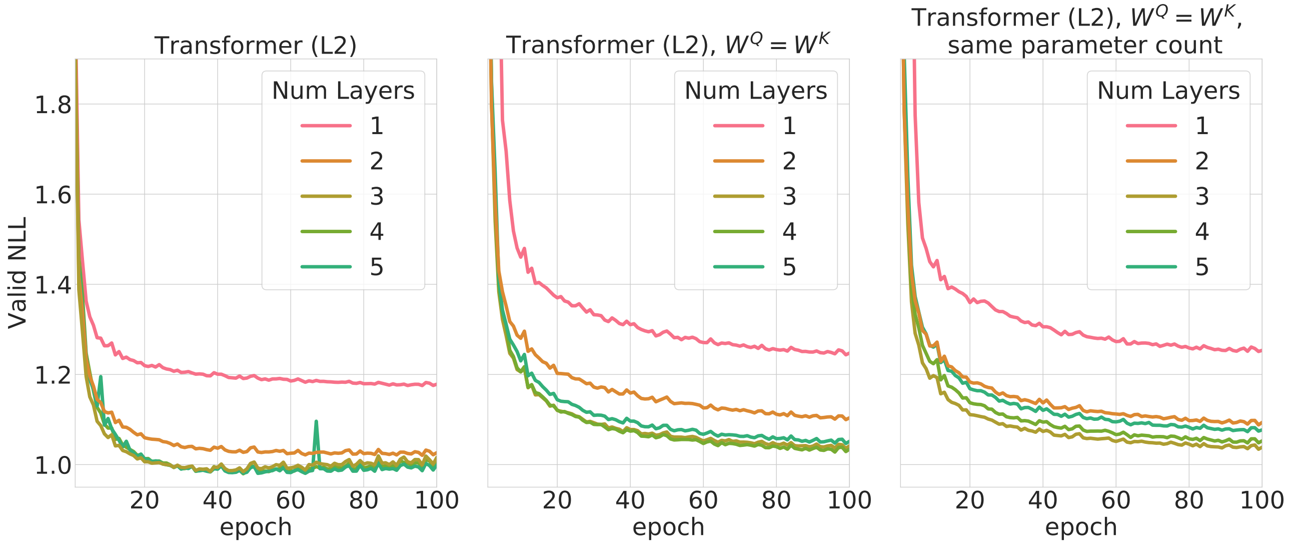

Figure 10. Experiment tying keys/queries but preserving parameter count.

M. Training curves for fixed learning rate DP-MHA vs L2-MHA

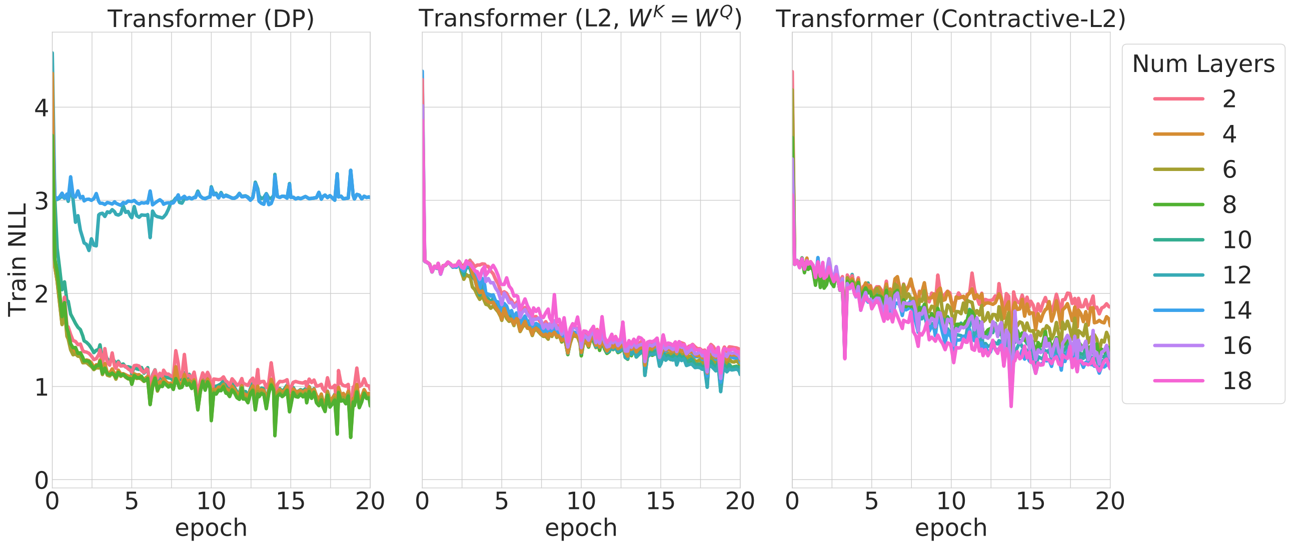

Figure 11. Train NLL for Transformer (DP), Transformer (L2) and Transformer (Contractive-L2)

N. The Lipschitz constant of LayerNorm

In this section, we show that `LayerNorm` is Lipschitz, with a loose bound on its Lipschitz constant w.r.t. to the ∞-norm. `LayerNorm` is defined as follows:

$$\begin{aligned}
\mathrm{LN}(\mathbf{x}) &= \frac{\mathbf{x}-\mu(\mathbf{x})}{\sqrt{\sigma^2(\mathbf{x})+\epsilon}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta} \\
\mu(\mathbf{x}) &= \frac{1}{D}\sum_{d=1}^{D} x_d \\
\sigma^2(\mathbf{x}) &= \frac{1}{D}\sum_{d=1}^{D}(x_d-\mu(\mathbf{x}))^2
\end{aligned}$$

where $\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^D$. We will omit dependence on x to write μ, σ^2 in cases when there is no ambiguity to reduce clutter.

In the trivial case where x_d are all equal or when $D = 1$, $\mathbf{x} = \mu$ hence $LN(\mathbf{x}) = \boldsymbol{\beta}$, so its Lipschitz constant is 0. Thus let us assume $D > 2$ and not all x_d are equal.

First let us compute the derivative of μ and σ^2 w.r.t x:

$$\begin{aligned}
\frac{\partial \mu}{\partial \mathbf{x}} &= \frac{1}{D}\mathbb{1}^\top \\
\frac{\partial \sigma^2}{\partial \mathbf{x}} &= \frac{1}{D}\sum_d 2(x_d-\mu)\frac{\partial}{\partial \mathbf{x}}(x_d-\mu) \\
&= \frac{2}{D}\sum_d (x_d-\mu)(\mathbf{e}_d - \frac{1}{D}\mathbb{1})^\top \\
&= \frac{2}{D}\left[\sum_d (x_d-\mu)\mathbf{e}_d - \frac{1}{D}\mathbb{1}\sum_d (x_d-\mu)\right]^\top \\
&= \frac{2}{D}\sum_d (x_d-\mu)\mathbf{e}_d^\top \\
&= \frac{2}{D}(\mathbf{x}-\mu)^\top
\end{aligned}$$

where $\mathbf{e}_d \in \mathbb{R}^D$ is a one-hot vector with 1 at the dth element. Note the penultimate equality holds because $\sum_d (x_d - \mu) = 0$.

Now the derivative of $\mathrm{LN}(\mathbf{x})_d$, the dth element of $\mathrm{LN}(\mathbf{x})$, w.r.t.$\mathbf{x}$ is

$$\begin{aligned}
\frac{\partial \mathrm{LN}(\mathbf{x})_d}{\partial \mathbf{x}} &= \gamma_d\left[\frac{\partial}{\partial \mathbf{x}}(x_d-\mu)(\sigma^2+\epsilon)^{-\frac{1}{2}} + (x_d-\mu)\Big(-\frac{1}{2}(\sigma^2+\epsilon)^{-\frac{3}{2}}\Big)\frac{\partial \sigma^2}{\partial \mathbf{x}}\right] \\
&= \gamma_d(\sigma^2+\epsilon)^{-\frac{1}{2}}\left[(\mathbf{e}_d - \frac{1}{D}\mathbb{1})^\top - \frac{1}{2}(x_d-\mu)(\sigma^2+\epsilon)^{-1}\frac{2}{D}(\mathbf{x}-\mu)^\top\right] \\
&= \gamma_d(\sigma^2+\epsilon)^{-\frac{1}{2}}\left[(\mathbf{e}_d - \frac{1}{D}\mathbb{1})^\top - \frac{1}{D}(\sigma^2+\epsilon)^{-1}(x_d-\mu)(\mathbf{x}-\mu)^\top\right].
\end{aligned}$$

Hence

$$\frac{\partial \mathrm{LN}(\mathbf{x})}{\partial \mathbf{x}} = (\sigma^2+\epsilon)^{-\frac{1}{2}}\left[\mathrm{diag}(\boldsymbol{\gamma}) - \frac{1}{D}\boldsymbol{\gamma}\mathbb{1}^\top - \frac{1}{D}(\sigma^2+\epsilon)^{-1}\mathrm{diag}(\boldsymbol{\gamma})(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top\right].$$

Note

$$\mathrm{diag}(\boldsymbol{\gamma}) - \frac{1}{D}\boldsymbol{\gamma}\mathbb{1}^\top = \begin{bmatrix} \gamma_1(D-1)/D & -\gamma_1/D & \dots & -\gamma_1/D \\ -\gamma_2/D & \gamma_2(D-1)/D & \dots & -\gamma_2/D \\ \vdots & \vdots & \ddots & \vdots \\ -\gamma_D/D & -\gamma_D/D & \dots & \gamma_D(D-1)/D \end{bmatrix},$$

hence

$$\left\|\operatorname{diag}(\boldsymbol{\gamma}) - \frac{1}{D}\boldsymbol{\gamma}\mathbb{1}^\top\right\|_\infty = \frac{2(D-1)}{D}\max_d |\gamma_d|, \tag{36}$$

recalling that $\|\cdot\|_\infty$ is the maximum absolute row sum.

Let $z_d \coloneqq x_d - \mu$. Hence $\sum_d z_d = 0$, $\sigma^2 = \frac{1}{D}\sum_d z_d^2$ and

$$\operatorname{Cov}(\mathbf{x}) = (\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top = \begin{bmatrix} z_1^2 & \dots & z_1 z_D \\ \vdots & \ddots & \vdots \\ z_D z_1 & \dots & z_D^2 \end{bmatrix}.$$

Hence

$$\frac{\|\operatorname{Cov}(\mathbf{x})\|_\infty}{\sigma^2} = \frac{\max_d |z_d| \sum_{d'} |z_{d'}|}{\frac{1}{D}\sum_d z_d^2}.$$

Noting that this expression is scale-invariant in $\mathbf{z}$, we may assume WLOG $\max_d |z_d| = z_D = 1$, since we are assuming not all x_d are equal and hence at least one z_d is non-zero.

The expression now becomes

$$\frac{\|\operatorname{Cov}(\mathbf{x})\|_\infty}{\sigma^2} = D\left(\frac{1+\sum_{d<D}|z_d|}{1+\sum_{d<D} z_d^2}\right). \tag{37}$$

Since all terms $|z_d| \leq 1$ are bounded, this continuous expression reaches a global maximum for some value of $\mathbf{z}$ with $z_D = 1$.

It is easy to see that at the global maximum, $z_d \neq 0 \; \forall d$: suppose this were to be true, WLOG $z_1 = 0$. Then let us see how the quantity (37) changes when $z_1 = 0$ is increased by $0 < \delta < 1$ and $z_D = 1$ is decreased by δ, keeping the sum constant. It is easy to see that the numerator $\sum_d |z_d|$ stays constant, but the denominator $\sum_d z_d^2$ changes by $2\delta^2 - 2\delta < 0$. Since for small δ, the numerator of (37) stays constant but the denominator decreases, the quantity (37) increases, contradicting that the global max is obtained for $z_1 = 0$. Hence we may assume that $z_d \neq 0 \; \forall d$.

Hence the quantity (37) (in particular, $\sum_d |z_d|$) is differentiable at the global maximum, at which the partial derivatives of the following Lagrangian are zero:

$$\mathcal{L}(z_1, \dots, z_{D-1}, \lambda) = \frac{1+\sum_{d<D}|z_d|}{1+\sum_{d<D} z_d^2} - \lambda(\sum_{d<D} z_d + 1).$$

From now on let us write $\sum$ for $\sum_{d<D}$ below to reduce clutter. Setting $\frac{\partial \mathcal{L}}{\partial z_k} = 0$ and noting $\frac{d|z_k|}{dz_k} = \operatorname{sgn}(z_k)$, we obtain

$$\begin{aligned} &\frac{\operatorname{sgn}(z_k)(1+\sum z_d^2) - 2z_k(1+\sum|z_d|)}{(1+\sum z_d^2)^2} - \lambda = 0 \\ \iff &\operatorname{sgn}(z_k)(1+\sum z_d^2) - 2z_k(1+\sum|z_d|) = \lambda(1+\sum z_d^2)^2 \\ \iff &z_k = \frac{\operatorname{sgn}(z_k)(1+\sum z_d^2) - \lambda(1+\sum z_d^2)^2}{2(1+\sum|z_d|)} \\ \iff &z_k = \frac{(\operatorname{sgn}(z_k) - \lambda(1+\sum z_d^2))(1+\sum z_d^2)}{2(1+\sum|z_d|)} \end{aligned}$$

Hence at the global maximum, z_k takes one of two values $a > 0$ and $b < 0$. Further we have that

$$\frac{1+\sum|z_d|}{1+\sum z_d^2} = \frac{\operatorname{sgn}(z_k) - \lambda(1+\sum z_d^2)}{2z_k} \tag{38}$$

If both a and b are among the z_k, we have that $\frac{1-\lambda(1+\sum z_d^2)}{2a} = \frac{-1-\lambda(1+\sum z_d^2)}{2b}$. Solving for $\lambda(1+\sum z_d^2)$ and plugging it in back to Equation (38), we get:

$$\frac{1+\sum|z_d|}{1+\sum z_d^2} = \frac{1}{a-b}$$

Since $a > 0$, $b < 0$ and $\sum z_d = -1$, $a - b$ is minimised when only one of the z_d is a and the rest are b. Hence a crude lower bound on $a - b$ is $\frac{1}{D-2}$, giving a bound:

$$\frac{\|\mathrm{Cov}(\mathbf{x})\|_\infty}{\sigma^2} \leq D(D-2) \tag{39}$$

However we conjecture that the true global maximum is attained when $z_d = -\frac{1}{D-1}$ $\forall d < D$ (i.e. all the z_d for $d < D$ are equal to $b < 0$), for which it is easy to show that $\frac{1+\sum_{d<D}|z_d|}{1+\sum_{d<D}z_d^2} = 2(D-1)/D$.

Putting together the above, we have:

$$\begin{aligned}
\left\|\frac{\partial \mathrm{LN}(\mathbf{x})}{\partial \mathbf{x}}\right\|_\infty &= (\sigma^2+\epsilon)^{-\frac{1}{2}} \left\|\mathrm{diag}(\boldsymbol{\gamma}) - \frac{1}{D}\boldsymbol{\gamma}\mathbb{1}^\top - \frac{1}{D}(\sigma^2+\epsilon)^{-1}\mathrm{diag}(\boldsymbol{\gamma})(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top\right\|_\infty \\
&\leq \epsilon^{-\frac{1}{2}}\left(\left\|\mathrm{diag}(\boldsymbol{\gamma}) - \frac{1}{D}\boldsymbol{\gamma}\mathbb{1}^\top\right\|_\infty + \frac{1}{D}\|\mathrm{diag}(\boldsymbol{\gamma})\|_\infty \left\|(\sigma^2+\epsilon)^{-1}(\mathbf{x}-\mu)(\mathbf{x}-\mu)^\top\right\|_\infty\right) \\
&\leq \epsilon^{-\frac{1}{2}}\left(\left\|\mathrm{diag}(\boldsymbol{\gamma}) - \frac{1}{D}\boldsymbol{\gamma}\mathbb{1}^\top\right\|_\infty + \frac{1}{D}\|\mathrm{diag}(\boldsymbol{\gamma})\|_\infty \left\|\mathrm{Cov}(\mathbf{x})/\sigma^2\right\|_\infty\right) \\
&\leq \epsilon^{-\frac{1}{2}}\left(\frac{2(D-1)}{D}\max_d|\gamma_d| + \frac{1}{D}\max_d|\gamma_d|D(D-2)\right) \\
&= \epsilon^{-\frac{1}{2}}\max_d|\gamma_d|\left(\frac{2(D-1)}{D} + D - 2\right) \\
&= \epsilon^{-\frac{1}{2}}\max_d|\gamma_d|\left(\frac{D^2-2}{D}\right).
\end{aligned}$$