Spectral Bound of Attention

Shengqi Qiu

03/03/2026

0. Notation (aligned with *Attention Is All You Need*). We follow the standard scaled dot-product attention notation. Let n denote the sequence length, and let d denote the model dimension. For a single head with key/query dimension d_k (so $h := d/d_k$ heads if multi-head is used), define

$$Q := XW_Q, \qquad K := XW_K, \qquad V := XW_V,$$

and the row-wise softmax attention

$$P(X) := \mathrm{SM}\Big(\frac{1}{\sqrt{d_k}} QK^\top\Big).$$

Equivalently, letting

$$A := W_Q W_K^\top \in \mathbb{R}^{d\times d}, \qquad S(X) := \frac{1}{\sqrt{d_k}} XAX^\top \in \mathbb{R}^{n\times n},$$

we have $P(X) = \mathrm{SM}(S(X))$.

Throughout, $\|\cdot\|_2$ denotes the spectral norm for matrices and the Euclidean norm for vectors. For a linear operator $\mathcal{T}$ between matrix normed spaces, $\|\mathcal{T}\|_2$ denotes the induced operator norm.

1. Definition. We define our function $f(X)$ as the product of two matrix-valued functions $P(X)$ and $V(X)$:

$$f(X) = P(X)\, V(X).$$

Here $P(X)$ is the attention matrix of size $n \times n$, and

$$V(X) = XW_V$$

is the value matrix of size $n \times d_v$ (in the Transformer one often has $d_v = d_k$, but we do not need this here).

2. Perturbation. To find the Jacobian (Fréchet derivative), perturb X by a small matrix ΔX:

$$f(X + \Delta X) = P(X + \Delta X)\, V(X + \Delta X).$$

Since P is C^2 in a neighborhood of the fixed base point X, Taylor's theorem gives

$$\begin{aligned} P(X + \Delta X) &= P(X) + (D_X P)[\Delta X] + R_P(\Delta X), \qquad \|R_P(\Delta X)\|_2 = \mathcal{O}(\|\Delta X\|_2^2), \\ V(X + \Delta X) &= V(X) + (D_X V)[\Delta X] = V(X) + \Delta X W_V. \end{aligned}$$

For brevity, write $P := P(X)$, $V := V(X)$, and

$$\Delta P := (D_X P)[\Delta X], \qquad \Delta V := (D_X V)[\Delta X] = \Delta X W_V.$$

3. Expand the product. Substituting the Taylor expansion into the product gives

$$\begin{aligned} f(X+\Delta X) &= (P+\Delta P+R_P(\Delta X))(V+\Delta V) \\ &= PV+(\Delta P)V+P(\Delta V)+(\Delta P)(\Delta V)+R_P(\Delta X)V+R_P(\Delta X)\Delta V. \end{aligned}$$

Since $D_X P$ and $D_X V$ are bounded linear maps, $\|\Delta P\|_2 = \mathcal{O}(\|\Delta X\|_2)$ and $\|\Delta V\|_2 = \mathcal{O}(\|\Delta X\|_2)$. Hence

$$(\Delta P)(\Delta V)+R_P(\Delta X)V+R_P(\Delta X)\Delta V = \mathcal{O}(\|\Delta X\|_2^2)$$

in spectral norm.

4. Isolate the linear part. Therefore

$$f(X+\Delta X)-f(X) = (\Delta P)V+P(\Delta V)+\mathcal{O}(\|\Delta X\|_2^2).$$

Keeping only the linear part in ΔX gives

$$\Delta P = (D_X P)[\Delta X], \qquad \Delta V = (D_X V)[\Delta X],$$

and hence the Jacobian operator is

$$\mathcal{J}(\Delta X) = (D_X P)[\Delta X]\cdot V \;+\; P\cdot(D_X V)[\Delta X].$$

Bounding the second term $P\cdot(D_X V)[\Delta X]$. Since $(D_X V)[\Delta X] = \Delta X W_V$, sub-multiplicativity of the spectral norm gives

$$\|P(\Delta X)W_V\|_2 \le \|P\|_2\|\Delta X\|_2\|W_V\|_2. \tag{1}$$

Bounding the first term $(D_X P)[\Delta X]\cdot V$. We define $\Delta S := D_X S(X)[\Delta X]$ as the variation of the score matrix $S = \frac{1}{\sqrt{d_k}} XAX^\top$.

The crucial identity $(\Delta P)\mathbf{1} = (D_S \mathrm{SM}(S)[\Delta S])\mathbf{1} = \mathbf{0}$.

We claim that the rows of the softmax derivative ΔP sum to zero: $(\Delta P)\mathbf{1} = \mathbf{0}$.

Let $S : \mathbb{R}^{n\times d} \to \mathbb{R}^{n\times n}$ be a differentiable map and let

$$P(X) \;:=\; \mathrm{SM}(S(X)),$$

where $\mathrm{SM} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ denotes the row-wise softmax.

For any direction $\Delta X \in \mathbb{R}^{n\times d}$, define

$$\Delta P \;:=\; D_X P(X)[\Delta X].$$

By the Fréchet chain rule,

$$\Delta P \;=\; D_X P(X)[\Delta X] \;=\; D_S \mathrm{SM}(S(X))[\,D_X S(X)[\Delta X]\,] \;=\; D_S \mathrm{SM}(S)\,[\Delta S], \tag{2}$$

where we set

$$\Delta S \;:=\; D_X S(X)[\Delta X].$$

Therefore to prove $(\Delta P)\mathbf{1} = \mathbf{0}$, it suffices to show that for any ΔS,

$$(D_S \mathrm{SM}(S)[\Delta S])\mathbf{1} \;=\; \mathbf{0}. \tag{3}$$

This follows from the fact that, for every S, the row-wise softmax produces a row-stochastic matrix:

$$\mathrm{SM}(S)\mathbf{1} = \mathbf{1}.$$

Differentiating the identity $\mathrm{SM}(S)\mathbf{1} = \mathbf{1}$ with respect to S in the direction ΔS yields

$$D_S\big(\mathrm{SM}(S)\mathbf{1}\big)[\Delta S] = \mathbf{0}.$$

By the proof below, we have that

$$D_S\big(\mathrm{SM}(S)\mathbf{1}\big)[\Delta S] = \big(D_S \mathrm{SM}(S)[\Delta S]\big)\mathbf{1}.$$

Combining the last two displays proves (3). Consequently, (3) holds for any ΔS, regardless of how ΔS is produced (in particular, even when $\Delta S = D_X S(X)[\Delta X]$ via (2)).

A rigorous proof of $D_S(\mathrm{SM}(S)\mathbf{1})[\Delta S] = \Big(D_S \mathrm{SM}(S)[\Delta S]\Big)\mathbf{1}$

Let $\mathrm{SM} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be the row-wise softmax map and define

$$F(S) \;:=\; \mathrm{SM}(S)\mathbf{1} \in \mathbb{R}^n.$$

We prove that for every direction $\Delta S \in \mathbb{R}^{n\times n}$,

$$DF(S)[\Delta S] = \big(D\,\mathrm{SM}(S)[\Delta S]\big)\mathbf{1}. \tag{2}$$

Step 1: Introduce the right-multiplication operator. Define the linear map

$$R_{\mathbf{1}} : \mathbb{R}^{n\times n} \to \mathbb{R}^n, \qquad R_{\mathbf{1}}(A) := A \cdot \mathbf{1}.$$

Then $F = R_{\mathbf{1}} \circ \mathrm{SM}$, i.e.,

$$F(S) = R_{\mathbf{1}}(\mathrm{SM}(S)).$$

Lemma (boundedness of $R_{\mathbf{1}}$). With the spectral norm $\|\cdot\|_2$ on matrices and the Euclidean norm $\|\cdot\|_2$ on vectors, $R_{\mathbf{1}}$ is bounded:

$$\|R_{\mathbf{1}}(A)\|_2 = \|A\mathbf{1}\|_2 \le \|A\|_2\,\|\mathbf{1}\|_2 \qquad \forall\, A \in \mathbb{R}^{n\times n}.$$

Hence $R_{\mathbf{1}}$ is continuous.

Step 2: Use the Fréchet definition (small-o) for SM. Assume SM is Fréchet differentiable at S. Then there exists a linear map $D\,\mathrm{SM}(S) : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ such that as $\Delta S \to 0$,

$$\mathrm{SM}(S + \Delta S) = \mathrm{SM}(S) + D\,\mathrm{SM}(S)[\Delta S] + r(\Delta S), \qquad \frac{\|r(\Delta S)\|_2}{\|\Delta S\|_2} \to 0. \tag{3}$$

Step 3: Apply R_1 to the expansion and control the remainder. Apply $R_{\mathbf{1}}$ to (3):

$$F(S+\Delta S) = F(S) + R_{\mathbf{1}}(D\,\mathrm{SM}(S)[\Delta S]) + R_{\mathbf{1}}(r(\Delta S)).$$

Since $R_{\mathbf{1}}(A) = A\mathbf{1}$, this is

$$F(S+\Delta S) = F(S) + (D\,\mathrm{SM}(S)[\Delta S])\mathbf{1} + r(\Delta S)\mathbf{1}. \tag{4}$$

Using boundedness of $R_{\mathbf{1}}$,

$$\frac{\|r(\Delta S)\mathbf{1}\|_2}{\|\Delta S\|_2} \le \|\mathbf{1}\|_2 \frac{\|r(\Delta S)\|_2}{\|\Delta S\|_2} \longrightarrow 0 \qquad (\Delta S \to 0),$$

so $r(\Delta S)\mathbf{1} = o(\|\Delta S\|_2)$.

Step 4: Identify the Fréchet derivative of F. From (4) and the previous estimate,

$$F(S+\Delta S) - F(S) - (D\,\mathrm{SM}(S)[\Delta S])\mathbf{1} = o(\|\Delta S\|_2).$$

By the definition of the Fréchet derivative, this proves (2), i.e.,

$$D_S(\mathrm{SM}(S)\mathbf{1})[\Delta S] = (D_S\mathrm{SM}(S)[\Delta S])\mathbf{1}.$$

From $(\Delta P)\mathbf{1} = 0$ to $(\Delta P)\Pi Z = (\Delta P)Z$

Assume we have already established the *zero row-sum* property

$$(\Delta P)\mathbf{1} = \mathbf{0}.$$

Let

$$\Pi \;:=\; I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$$

be the orthogonal projector onto $\mathbf{1}^\perp$. Then for any matrix (or vector) Z of compatible size,

$$(\Delta P)\Pi Z = (\Delta P)Z.$$

Starting from the left-hand side and substituting the definition of Π,

$$(\Delta P)\Pi Z = (\Delta P)\Big(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top\Big)Z.$$

Using distributivity of matrix multiplication,

$$(\Delta P)\Pi Z = (\Delta P)IZ - \frac{1}{n}(\Delta P)\mathbf{1}\mathbf{1}^\top Z.$$

Since $(\Delta P)IZ = (\Delta P)Z$ and matrix multiplication is associative, we may regroup the second term as

$$(\Delta P)\Pi Z = (\Delta P)Z - \frac{1}{n}((\Delta P)\mathbf{1})\,\mathbf{1}^\top Z.$$

Finally, by $(\Delta P)\mathbf{1} = \mathbf{0}$, the second term vanishes:

$$(\Delta P)\Pi Z = (\Delta P)Z - \frac{1}{n}\mathbf{0}\,\mathbf{1}^\top Z = (\Delta P)Z.$$

Therefore, combining the results above, the first term $(D_X P)[\Delta X]\cdot V$ can be equivalently written as:

$$\Delta P\, V = (\Delta P)\Pi V = (\Delta P)\Pi X W_V,$$

whence

$$\|\Delta P\, V\|_2 \le \|\Delta P\|_2 \|\Pi X\|_2 \|W_V\|_2.$$

Chain rule along a line via the Fréchet derivative

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be Fréchet differentiable at every point of the line $u(t) = s + t\Delta s$, and define

$$g(t) := f(u(t)) = f(s + t\Delta s).$$

Fix $t \in \mathbb{R}$. By the definition of the (one-dimensional) derivative,

$$g'(t) = \lim_{h\to 0} \frac{g(t+h) - g(t)}{h} = \lim_{h\to 0} \frac{f(u(t+h)) - f(u(t))}{h}.$$

Since $u(t+h) = s + (t+h)\Delta s = u(t) + h\Delta s$, this becomes

$$g'(t) = \lim_{h\to 0} \frac{f(u(t) + h\Delta s) - f(u(t))}{h}.$$

Because f is Fréchet differentiable at $u(t)$, there exists a bounded linear map $Df(u(t)) : \mathbb{R}^n \to \mathbb{R}^n$ and a remainder $r(v)$ such that, for $v \to 0$,

$$f(u(t) + v) = f(u(t)) + Df(u(t))[v] + r(v), \qquad \frac{\|r(v)\|}{\|v\|} \to 0.$$

Apply this with $v = h\Delta s$:

$$f(u(t) + h\Delta s) - f(u(t)) = Df(u(t))[h\Delta s] + r(h\Delta s).$$

Divide by h and use linearity of $Df(u(t))$:

$$\frac{f(u(t) + h\Delta s) - f(u(t))}{h} = Df(u(t))[\Delta s] + \frac{r(h\Delta s)}{h}.$$

Moreover,

$$\left\| \frac{r(h\Delta s)}{h} \right\| = \frac{\|r(h\Delta s)\|}{\|h\Delta s\|} \|\Delta s\| \xrightarrow[h\to 0]{} 0,$$

hence the limit exists and equals $Df(u(t))[\Delta s]$. Therefore,

$$g'(t) = Df(u(t))[\Delta s] = Df(s + t\Delta s)[\Delta s].$$

Row-wise exponential normalization: exact increment and spectral bound. Let $\mathrm{SM} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ be the softmax function

$$(\mathrm{SM}(S))_{i:} := \mathrm{SM}(S_{i:}), \qquad \mathrm{SM}(s)_k := \frac{e^{s_k}}{\sum_{j=1}^n e^{s_j}}.$$

Fix $S \in \mathbb{R}^{n\times n}$ and a perturbation $\Delta S \in \mathbb{R}^{n\times n}$. Define

$$S(t) := S + t\Delta S, \qquad P(t) := \mathrm{SM}(S(t)), \qquad p_i(t) := P(t)_{i:}.$$

Then for each row i we have the *exact* identity

$$\Delta P^{\mathrm{inc}}_{i:} := P(1)_{i:} - P(0)_{i:} = \int_0^1 \Delta S_{i:}\, J(p_i(t))\, dt, \qquad J(p) := \mathrm{Diag}(p) - p^\top p. \tag{5}$$

Derivation. Set $g_i(t) := p_i(t) = \mathrm{SM}(S_{i:} + t\Delta S_{i:})$. Since SM is C^1, g_i is differentiable and, by the chain rule in Fréchet form,

$$g_i'(t) = D\,\mathrm{SM}(S_{i:} + t\Delta S_{i:})[\Delta S_{i:}].$$

Moreover, the Fréchet derivative of SM satisfies

$$D\,\mathrm{SM}(u)[h] = h\,J(\mathrm{SM}(u)).$$

Hence $g_i'(t) = \Delta S_{i:}\,J(p_i(t))$, and integrating from 0 to 1 yields (5).

Norm bounds for the finite increment. Let

$$\alpha_{\mathrm{inc}}(\Delta S) := \max_{1\le i\le n} \sup_{t\in[0,1]} \|J(p_i(t))\|_2.$$

Here $\|\cdot\|_2$ denotes the Euclidean norm on row vectors and its induced matrix norm on $J(p_i(t))$. Using $\|xA\|_2 \le \|x\|_2\|A\|_2$ and (5),

$$\|\Delta P_{i:}^{\mathrm{inc}}\|_2 \le \int_0^1 \|\Delta S_{i:}\|_2\,\|J(p_i(t))\|_2\,dt \le \alpha_{\mathrm{inc}}(\Delta S)\|\Delta S_{i:}\|_2.$$

Since for any matrix Y, we have that $\|Y\|_F^2 = \sum_i \|Y_{i:}\|_2^2$, squaring and summing over i gives

$$\|\Delta P^{\mathrm{inc}}\|_F \le \alpha_{\mathrm{inc}}(\Delta S)\|\Delta S\|_F.$$

Finally, using $\|M\|_2 \le \|M\|_F$ and $\|M\|_F \le \sqrt{\mathrm{rk}(M)}\|M\|_2$,

$$\|\Delta P^{\mathrm{inc}}\|_2 \le \|\Delta P^{\mathrm{inc}}\|_F \le \alpha_{\mathrm{inc}}(\Delta S)\|\Delta S\|_F \le \alpha_{\mathrm{inc}}(\Delta S)\sqrt{\mathrm{rk}(\Delta S)}\,\|\Delta S\|_2.$$

Lemma (rank bound used above). With $\Delta S = \frac{1}{\sqrt{d_k}}((\Delta X)AX^\top + XA(\Delta X)^\top)$,

$$\mathrm{rk}(\Delta S) \le \mathrm{rk}((\Delta X)AX^\top) + \mathrm{rk}(XA(\Delta X)^\top) \le \mathrm{rk}(\Delta X) + \mathrm{rk}(\Delta X) = 2\,\mathrm{rk}(\Delta X) \le 2\min\{n, d\} \le 2d.$$

Therefore,

$$\|\Delta P^{\mathrm{inc}}\|_2 \le \alpha_{\mathrm{inc}}(\Delta S)\sqrt{2d}\,\|\Delta S\|_2.$$

In particular, since $\sup_{p\in\Delta^{n-1}} \|J(p)\|_2 = \frac{1}{2}$, we have $\alpha_{\mathrm{inc}}(\Delta S) \le \frac{1}{2}$ and thus

$$\|\Delta P^{\mathrm{inc}}\|_2 \le \sqrt{\frac{d}{2}}\,\|\Delta S\|_2.$$

Passing from finite increments to the Fréchet derivative. The finite increment ΔP^{inc} is not the same object as the Fréchet derivative $\Delta P = D_S\mathrm{SM}(S)[\Delta S]$. To obtain a derivative bound, fix a direction $H \in \mathbb{R}^{n\times n}$ and apply the finite-increment estimate to the scaled perturbation εH. Define

$$\alpha_\varepsilon(H) := \max_{1\le i\le n} \sup_{t\in[0,1]} \|J(\mathrm{SM}(S_{i:} + t\varepsilon H_{i:}))\|_2\,.$$

Then

$$\frac{\|\mathrm{SM}(S+\varepsilon H) - \mathrm{SM}(S)\|_2}{\varepsilon} \le \alpha_\varepsilon(H)\sqrt{\mathrm{rk}(H)}\,\|H\|_2.$$

Since SM is Fréchet differentiable,

$$D_S\mathrm{SM}(S)[H] = \lim_{\varepsilon\to 0} \frac{\mathrm{SM}(S+\varepsilon H) - \mathrm{SM}(S)}{\varepsilon}$$

in spectral norm. Hence

$$\|D_S\mathrm{SM}(S)[H]\|_2 \le \alpha_0(H)\sqrt{\mathrm{rk}(H)}\,\|H\|_2, \qquad \alpha_0(H) := \limsup_{\varepsilon\to 0} \alpha_\varepsilon(H) \le \frac{1}{2}.$$

Now take $H = \Delta S = D_X S(X)[\Delta X]$. By the chain rule,

$$\Delta P = D_X P(X)[\Delta X] = D_S\mathrm{SM}(S)[\Delta S],$$

and the rank bound above gives the derivative estimate

$$\|\Delta P\|_2 \le \sqrt{2d}\,\alpha(\Delta X)\,\|\Delta S\|_2, \qquad \alpha(\Delta X) := \alpha_0(D_X S(X)[\Delta X]) \le \frac{1}{2}.$$

Bound for ΔS under spectral norm

We differentiate $S = \frac{1}{\sqrt{d_k}} XAX^\top$:

$$\Delta S = \frac{1}{\sqrt{d_k}} \left((\Delta X)AX^\top + XA(\Delta X)^\top \right). \tag{6}$$

Bounding the spectral norm using triangle inequality and sub-multiplicativity:

$$\|\Delta S\|_2 \le \frac{1}{\sqrt{d_k}} \left(\|(\Delta X)AX^\top\|_2 + \|XA(\Delta X)^\top\|_2 \right), \tag{7}$$

$$\le \frac{1}{\sqrt{d_k}} \left(\|\Delta X\|_2 \|AX^\top\|_2 + \|XA\|_2 \|(\Delta X)^\top\|_2 \right). \tag{8}$$

Note that for the spectral norm, $\|Z\|_2 = \|Z^\top\|_2$. Thus $\|(\Delta X)^\top\|_2 = \|\Delta X\|_2$.

$$\|\Delta S\|_2 \le \frac{1}{\sqrt{d_k}} \left(\|XA^\top\|_2 + \|XA\|_2 \right) \|\Delta X\|_2. \tag{9}$$

Combining all these terms together, we can finalize our bound for $\|\mathcal{J}\|_2$.

Indeed, for any ΔX,

$$\|\mathcal{J}(\Delta X)\|_2 \le \|W_V\|_2 \left(\|P\|_2 + \sqrt{2}\,\alpha(\Delta X)\sqrt{\frac{d}{d_k}}\,\|\Pi X\|_2 \left(\|XA^\top\|_2 + \|XA\|_2 \right) \right) \|\Delta X\|_2,$$

where $\alpha(\Delta X)$ is defined above. Hence, by the definition of the induced operator norm, we must take the supremum over all nonzero directions ΔX. Let

$$\bar{\alpha} := \sup_{\Delta X \neq 0} \alpha(\Delta X) \le \frac{1}{2}.$$

Then

$$\|\mathcal{J}\|_2 \le \|W_V\|_2 \left(\|P\|_2 + \sqrt{2}\,\bar{\alpha}\sqrt{\frac{d}{d_k}}\,\|\Pi X\|_2 \left(\|XA^\top\|_2 + \|XA\|_2 \right) \right),$$

where $h := d/d_k$ is the number of heads in the attention, assumed to be an integer.

Acknowledgements

We thank Prof. Grigorios Chrysos and Muhammad Ashiq for their helpful feedback.