Softmax is $\frac{1}{2}$-Lipschitz
(in a norm that may not matter)

Laker Newhouse
lakern@mit.edu

February 2025

Abstract

The best result in the literature for the ℓ_2 Lipschitz constant of softmax(γx) for $\gamma > 0$ is γ [Gao and Pavel, 2018]. We improve this bound to $\gamma/2$ and prove that the new bound is tight. The sensitivity of softmax has implications for deep learning, particularly for attention in Transformers.

1 Introduction

The softmax function for $x \in \mathbb{R}^n$ is defined in each component as

$$\text{softmax}(x)_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}. \tag{1}$$

We are interested in the maximum sensitivity of softmax(γx), where $\gamma > 0$ is the inverse temperature. The sensitivity is bounded by the spectral radius of the Jacobian. We calculate the Jacobian and show that its maximum singular value is bounded from above at $\gamma/2$. We construct an example to show the bound is tight.

2 Softmax Jacobian

Let $p = \text{softmax}(\gamma x)$. Let $P = \text{diag}(p_1, \ldots, p_n)$. The Jacobian of softmax(γx) is well known to be

$$J = \gamma(P - pp^\top), \tag{2}$$

or in component form $J_{ij} = \frac{\partial p_i}{\partial x_j} = \gamma p_i(\delta_{ij} - p_j)$, where δ_{ij} is the Kronecker delta.

This Jacobian matrix is symmetric and has real number entries. Therefore its eigenvalues are real. We seek its maximum eigenvalue λ, which is its spectral radius and thus the Lipschitz constant of softmax(γx).

The bound cited often in the literature is $\lambda \leq \gamma$. This loose bound follows from noting that J is positive semidefinite; thus its eigenvalues are nonnegative, and all of them together sum up to $\text{Tr}(J) = \gamma(1 - p^T p) \leq \gamma$.

3 The Gershgorin circle theorem

In 1931, Gershgorin proved that the eigenvalues of a matrix J lie within at least one of the disks

$$D(J_{ii}, R_i) \subseteq \mathbb{C}, \tag{3}$$

where J_{ii} is the center of the disk and $R_i = \sum_{j \neq i} |J_{ij}|$ is its radius [Gershgorin, 1931]. In our case, the eigenvalues of J are all real numbers. Thus all the eigenvalues of J lie within at least one interval

$$[J_{ii} - R_i, J_{ii} + R_i]. \tag{4}$$

Recall that $J_{ii} = \gamma p_i(1-p_i)$. Miraculously, the radius reduces to the same value:

$$R_i = \sum_{j \neq i} |J_{ij}| = \gamma \sum_{j \neq i} |-p_i p_j| = \gamma p_i \sum_{j \neq i} |p_j| = \gamma p_i(1-p_i). \tag{5}$$

The maximum of the function $f(p_i) = p_i(1-p_i)$ for $p_i \in [0,1]$ is $\frac{1}{4}$. Thus the maximum value that the center J_{ii} and the radius R_i can attain is $\frac{1}{4}\gamma$. The farthest any disk can reach is then the interval $\left[0, \frac{1}{4}\gamma + \frac{1}{4}\gamma\right] = \left[0, \frac{1}{2}\gamma\right]$. Even the maximum eigenvalue of the softmax Jacobian cannot exceed $\frac{1}{2}\gamma$.

4 The bound is tight

Consider $x = (0, 0, -\alpha, \ldots, -\alpha)$ as $\alpha \to \infty$. Then $\text{softmax}(\gamma x)$ approaches $p = (\frac{1}{2}, \frac{1}{2}, 0, \ldots, 0)$ with Jacobian

$$J = \gamma \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} & 0 & \cdots & 0 \\ -\frac{1}{4} & \frac{1}{4} & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}. \tag{6}$$

For $v = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 \end{bmatrix}^\top$ we can compute

$$v^\top J v = \frac{\gamma}{2} \begin{bmatrix} -1 & 1 \end{bmatrix} \begin{bmatrix} -\frac{1}{2} \\ \frac{1}{2} \end{bmatrix} = \frac{\gamma}{2}, \tag{7}$$

which attains the Lipschitz bound $\|Jv\|_2 \leq \frac{\gamma}{2}\|v\|_2$. We conclude the bound is tight. The example also shows that the highest sensitivity regime for softmax is when the softmax reduces to a choice between two indices.

5 Discussion

Several works aim to put Lipschitz bounds on neural networks, including using orthogonal weight constraints to improve gradient flow [Qi et al., 2023; Béthune, 2024]. Softmax is important for this program because it appears in almost every modern architecture, including Transformers [Vaswani et al., 2017].

While the original $1/\sqrt{d}$ scaling in dot product attention is not Lipschitz, subsequent work has proposed ways to modify attention to be Lipschitz [Kim et al., 2021]. In particular, Large et al. [2024] use the max-over-tokens RMS norm. Another possibility is the (computationally intractable) $L_{\infty \to 1}$ induced operator norm: given all modules are well-normed in the sense of Large et al., the input is unit L_∞ norm after scaling by $1/d$, meaning it is entrywise at most 1, and the output is a probability vector equipped with the L_1 norm.

If the useful input-output norms for softmax are not Euclidean, then the bound in this paper is moot.

6 Conclusion

We have proved that $\frac{\gamma}{2}$ is a tight bound on the ℓ_2 Lipschitz constant of $\text{softmax}(\gamma x)$. We hope this simple result might be useful in its own right and for attempts to control the dynamics of attention in Transformers.

7 Acknowledgments

Thank you to Claude AI for many rounds of brainstorming, including the idea to use Gershgorin's circle theorem. Thank you to Jeremy Bernstein for everything this year. Subsequently, I learned that Jeremy and his collaborator Tim developed a similar proof while writing the modular norm paper [Large et al., 2024].

References

Louis Béthune. *Deep learning with Lipschitz constraints*. Ph.d. thesis, Université de Toulouse, 2024. URL https://tel.archives-ouvertes.fr/tel-04674274. In English. NNT: 2024TLSES014. 2

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning, 2018. URL https://arxiv.org/abs/1704.00805. 1

S. A. Gershgorin. Uber die abgrenzung der eigenwerte einer matrix. *Bulletin of the Russian Academy of Sciences*, (6):749–754, 1931. 1

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention, 2021. URL https://arxiv.org/abs/2006.04710. 2

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm, 2024. URL https://arxiv.org/abs/2405.14813. 2

Xianbiao Qi, Jianan Wang, and Lei Zhang. Understanding optimization of deep learning via jacobian matrix and lipschitz constant, 2023. URL https://arxiv.org/abs/2306.09338. 2

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2