How Wide Does One Attention Head Need to Be?

Due: Wednesday, November 25 at 11:59pm

Instructions: Please submit your *handwritten* solutions as a single PDF on Gradescope.

Suppose we want one attention head to give each token a mix of local and global information. At each position $i = 1, \ldots, n$, we want the output to be halfway between that token's value and the average of all n values:

$$y_i = \frac{1}{2} v_i + \frac{1}{2} \bar{v}, \qquad \bar{v} = \frac{1}{n} \sum_{j=1}^{n} v_j.$$

We can choose any queries and keys in $\mathbb{R}^d$. Once chosen, they must give this output for every choice of scalar values $v_1, \ldots, v_n$. Assume $n \geq 2$ and allow each token to attend to all n positions.

1. The output of the head is $y_i = \sum_{j=1}^{n} a_{ij} v_j$. What must a_{ij} be for this weighted sum to give the desired output for every choice of values?
2. Let $\mathbf{s}_j$ be column j of the score matrix $\mathbf{S}$. For $j = 1, \ldots, n-1$, what is the difference $\mathbf{s}_j - \mathbf{s}_n$?
3. These $n-1$ differences are all in the column space of $\mathbf{S}$. Why do they force the query and key dimension to satisfy $d \geq n-1$?

Hint: For part 2, two weights in the same row have the same softmax denominator, so $a_{ij}/a_{in} = \exp(s_{ij} - s_{in})$.

Takeaway: *Every token follows the same averaging rule, but each one needs to favor a different position. The number of query and key coordinates needed for this rule grows with the sequence length.*