My solutions to

Deep Learning: Foundations and Concepts

Dario Miro Konopatzki

12 Transformers

12.2

For any $x_k, x_l \in \mathbb{R}^D$, $x_k^\top x_l \in \mathbb{R}$ and thus $e^{x_k^\top x_l} > 0$. Hence $a_{nm} = \frac{\overbrace{e^{x_n^\top x_m}}^{>0}}{\sum_{m'=1}^{N} \underbrace{e^{x_n^\top x_{m'}}}_{>0 \text{ f.a. } m'}} > 0$.

$$\begin{aligned}
\sum_{m=1}^{N} a_{nm} &= \sum_{m=1}^{N} \frac{e^{x_n^\top x_m}}{\sum_{m'=1}^{N} e^{x_n^\top x_{m'}}} \\
&= \frac{\cancel{\sum_{m=1}^{N} e^{x_n^\top x_m}}}{\cancel{\sum_{m'=1}^{N} e^{x_n^\top x_{m'}}}} \\
&= 1
\end{aligned}$$