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 $a_{nm} = \frac{\overbrace{e^{x_n^\top x_m}}^{>0}}{\sum_{m'=1}^{N} \underbrace{e^{x_n^\top x_{m'}}}_{>0 \text{ f.a. } m'}} > 0$.

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

12.4

$$\begin{aligned}
\mathbb{E}\left[\left(a^\top b\right)^2\right] &= \mathbb{E}\left[\left(\sum_{d=1}^{D} a_d b_d\right)^2\right] \\
&= \mathbb{E}\left[\sum_{d=1}^{D} \left(a_d b_d\right)^2 + \sum_{d=1}^{D} \sum_{\substack{d'=1 \\ d' \neq d}}^{D} a_d b_d a_{d'} b_{d'}\right] \\
&= \sum_{d=1}^{D} \left(\mathbb{E}\left[a_d^2 b_d^2\right] + \sum_{\substack{d'=1 \\ d' \neq d}}^{D} \mathbb{E}\left[a_d b_d a_{d'} b_{d'}\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{\star}{=} \sum_{d=1}^{D} \left(\mathbb{E}\left[a_d^2\right] \mathbb{E}\left[b_d^2\right] + \sum_{\substack{d'=1 \\ d' \neq d}}^{D} \mathbb{E}\left[a_d\right] \mathbb{E}\left[b_d\right] \mathbb{E}\left[a_{d'}\right] \mathbb{E}\left[b_{d'}\right] \right) \\
&\overset{\dagger}{=} \sum_{d=1}^{D} \left(1 \cdot 1 + \sum_{\substack{d'=1 \\ d' \neq d}}^{D} 0 \cdot 0 \cdot 0 \cdot 0 \right) \\
&= D
\end{aligned}
$$

$\star$ By assumption $a, b \sim \mathcal{N}(0, \mathbb{I})$. Diagonal covariance implies components of a (resp. b) independent. Taken together with a, b independent this gives $a_1, \ldots, a_D, b_1, \ldots, b_D$ independent.

$\dagger$ By assumption $a \sim \mathcal{N}(0, \mathbb{I})$, so $\mathbb{E}[a_d] = 0$ and $\mathbb{E}[a_d^2] = \mathbb{E}\left[aa^\top\right]_{dd} = \mathbb{I}_{dd} = 1$ for all d. The analogous result holds for b.

12.5

$$
\begin{aligned}
Y &= \left[\text{Attention}\left(Q_h, K_h, V_h\right)\right]_h W^{(o)} \\
&= \sum_{h=1}^{H} \text{Attention}\left(Q_h, K_h, V_h\right) W_h^{(o)} \\
&= \sum_{h=1}^{H} \text{Softmax}\left(\frac{Q_h K_h^\top}{\sqrt{D_k}}\right) V_h W_h^{(o)} \\
&= \sum_{h=1}^{H} \text{Softmax}\left(\frac{Q_h K_h^\top}{\sqrt{D_k}}\right) X W_h^{(v)} W_h^{(o)} \\
&= \sum_{h=1}^{H} \text{Softmax}\left(\frac{Q_h K_h^\top}{\sqrt{D_k}}\right) X W^{(h)} \\
&= \sum_{h=1}^{H} \text{Attention}\left(Q_h, K_h, X W^{(h)}\right)
\end{aligned}
$$