

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

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

12.4

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

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

★ 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 $a_1, \dots, a_D, b_1, \dots, b_D$ independent.

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

12.5

Let σ denote softmax and Att attention. Then

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

12.7

Let $\pi : N \rightarrow N$ be any permutation, $P := (\delta_{\pi(n)n'})_{nn'}$. Note that $P^\top P = (\delta_{\pi(n)n'})_{n'n} (\delta_{\pi(n)n'})_{nn'} = (\sum_n \delta_{\pi(n)n'} \delta_{\pi(n)n''})_{n'n''} = (\delta_{n'n''})_{n'n''} = \mathbb{I}_N$.

Consider any $A \in \mathbb{R}^{N \times N}$. Then

$$\begin{aligned}
\sigma(PA) &= \sigma \left(\left(\sum_{n'} \delta_{\pi(n)n'} A_{n'n''} \right)_{nn''} \right) \\
&= \sigma \left((A_{\pi(n)n''})_{nn''} \right) \\
&= \left(\frac{e^{A_{\pi(n)n''}}}{\sum_{n'} e^{A_{\pi(n)n'}}} \right)_{nn''} \\
&= (\sigma(A)_{\pi(n)n''})_{nn''} \\
&= \left(\sum_{n'} \delta_{\pi(n)n'} \sigma(A)_{n'n''} \right)_{nn''} \\
&= P\sigma(A)
\end{aligned}$$

and

$$\begin{aligned}
\sigma(AP^\top) &= \sigma \left(\left(\sum_{n'} A_{nn'} \delta_{\pi(n'')n'} \right)_{nn''} \right) \\
&= \sigma \left((A_{n\pi(n'')})_{nn''} \right) \\
&= \left(\frac{e^{A_{n\pi(n'')}}}{\sum_{n'} e^{A_{n\pi(n')}}} \right)_{nn''} \\
&= (\sigma(A)_{n\pi(n'')})_{nn''} \\
&= \left(\sum_{n'} \sigma(A)_{nn'} \delta_{\pi(n'')n'} \right)_{nn''} \\
&= \sigma(A)P^\top
\end{aligned}$$

Since A arbitrary, for every h where $(W^{(q)}, W^{(k)}, W^{(v)} := W_h^{(q)}, W_h^{(k)}, W_h^{(v)})$

$$\begin{aligned}
\text{Att}(PXW^{(q)}, PXW^{(k)}, PXW^{(v)}) &= \sigma \left(\frac{PXW^{(q)} (PXW^{(k)})^\top}{\sqrt{D_k}} \right) PXW^{(v)} \\
&= \sigma \left(\frac{PXW^{(q)} W^{(k)\top} X^\top P^\top}{\sqrt{D_k}} \right) PXW^{(v)} \\
&= P\sigma \left(\frac{XW^{(q)} W^{(k)\top} X^\top}{\sqrt{D_k}} \right) P^\top PXW^{(v)} \\
&= P\sigma \left(\frac{XW^{(q)} (XW^{(k)})^\top}{\sqrt{D_k}} \right) \mathbb{I}_N XW^{(v)} \\
&= \text{Att}(XW^{(q)}, XW^{(k)}, XW^{(v)})
\end{aligned}$$

and since the above holds for all h

$$\left[\text{Att} \left(PXW_h^{(q)}, PXW_h^{(k)}, PXW_h^{(v)} \right) \right]_h W^{(o)} = P [\text{Att} (Q_h, K_h, V_h)]_h W^{(o)}$$