$$\text{softmax}([x_1, x_2, x_3, x_4, x_5])$$

$$= \left[\frac{e^{x_1 - max}}{S}, \frac{e^{x_2 - max}}{S}, \frac{e^{x_3 - max}}{S}, \frac{e^{x_4 - max}}{S}, \frac{e^{x_5 - max}}{S} \right]$$

where: $S = \left(e^{x_1} + e^{x_2} + e^{x_3} + e^{x_4} + e^{x_5}\right) e^{-max}$

for numerical stability, add $-\max(x_1, x_2, x_3, x_4, x_5)$

Online softmax: compute softmax incrementally

new!

$$\text{softmax}([x_1, x_2, x_3]) \longrightarrow \text{softmax}([x_1, x_2, x_3, \boxed{x_4, x_5}])$$

Goal: We want to compute softmax block by block

$$\text{softmax}([x_1, x_2, x_3])$$

$$= \left[\frac{e^{x_1 - m_1}}{S_1}, \frac{e^{x_2 - m_1}}{S_1}, \frac{e^{x_3 - m_1}}{S_1} \right]$$

Where: $\boxed{\begin{array}{l} m_1 = \max(x_1, x_2, x_3) \\ S_1 = e^{x_1 - m_1} + e^{x_2 - m_1} + e^{x_3 - m_1} \end{array}}$

Then: How to compute softmax$([x_1, x_2, x_3, x_4, x_5])$

by reusing softmax$([x_1\ x_2\ x_3])$?

old new

$m_2 = \max(\boxed{x_1\, x_2\, x_3}\ \boxed{x_4\, x_5})$

$S_2 = \boxed{e^{x_1 - m_2} + e^{x_2 - m_2} + e^{x_3 - m_2}}$
$\boxed{+ e^{x_4 - m_2} + e^{x_5 - m_2}}$

Observe:

The end goal is:

$$\left[\frac{e^{x_1 - m_2}}{S_2} \quad \frac{e^{x_2 - m_2}}{S_2} \quad \frac{e^{x_3 - m_2}}{S_2} \quad \frac{e^{x_4 - m_2}}{S_2} \quad \frac{e^{x_5 - m_2}}{S_2} \right]$$

And we already have

$$\left[\frac{e^{x_1 - m_1}}{S_1}, \frac{e^{x_2 - m_1}}{S_1}, \frac{e^{x_3 - m_1}}{S_1} \right]$$

For the first 3 elements

We can rescale them by $\cdot S_1 \cdot S_2^{-1} \cdot e^{m_1} \cdot e^{-m_2}$

$$= \frac{S_1}{S_2} \cdot \frac{m_1}{m_2}$$

this scaling factor can make them

$$\left[\frac{e^{x_1-m_2}}{S_2} \quad \frac{e^{x_2-m_2}}{S_2} \quad \frac{e^{x_3-m_2}}{S_2} \right]$$

In the original paper,

it stores softmax result into 3 components

$$\mathrm{softmax}([x_1\ x_2]) = \begin{cases} ① & [e^{x_1-m_1} \quad e^{x_2-m_1}] \\ ② & m_1 = \max(x_1, x_2) \\ ③ & S_1 = e^{x_1-m_1} + e^{x_2-m_1} \end{cases}$$

the benefit of this

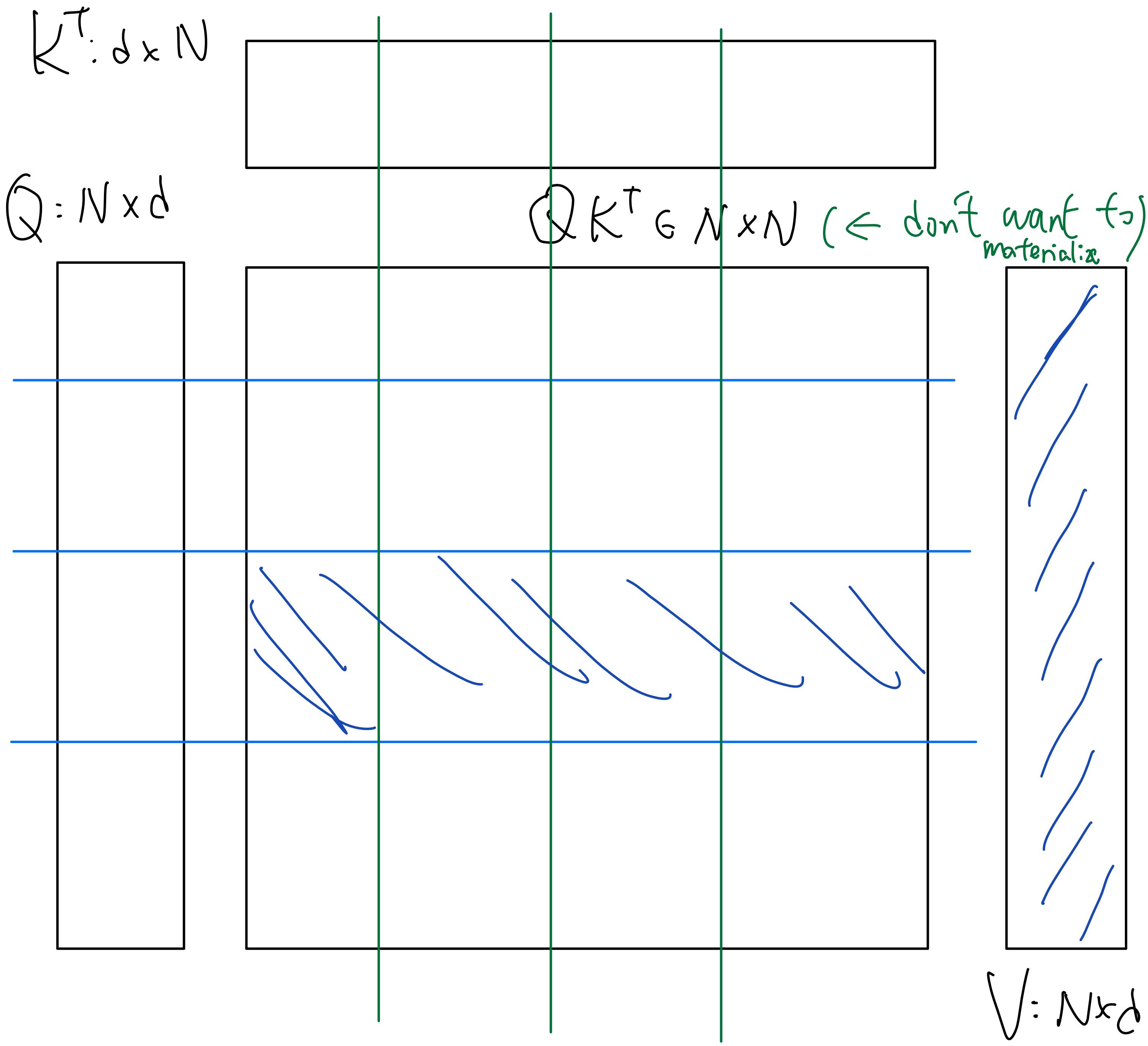
$\mathrm{softmax}([x_1\ x_2])$, $\mathrm{softmax}([x_3\ x_4\ x_5])$

can express $\mathrm{softmax}([x_1\ x_2\ x_3\ x_4\ x_5])$

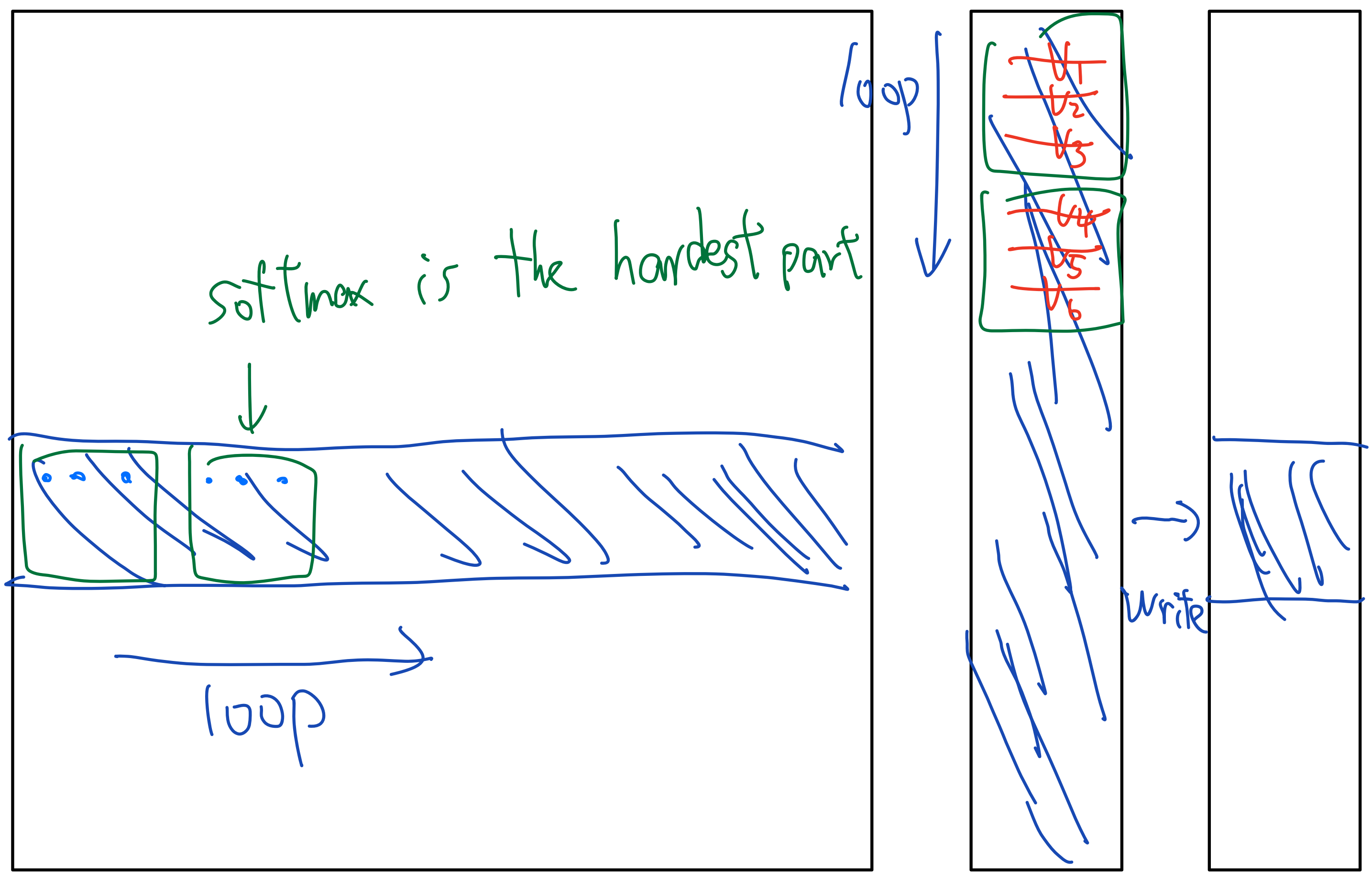
$$\text{by} \begin{cases} ① & \frac{e^{m_1}}{e^{m'}}[e^{x_1-m_1} \quad e^{x_2-m_1}] : \frac{e^{m_2}}{e^{m'}}[e^{x_3-m_2} \quad e^{x_4-m_2} \quad e^{x_5-m_2}] \\ ② & m' = \max(m_1, m_2) \\ ③ & S' = \frac{e^{m_1}}{e^{m'}} S_1 + \frac{e^{m_2}}{e^{m'}} S_2 \end{cases}$$

where $m_1 = \max(x_1, x_2)$ $\quad m_2 = \max(x_3\ x_4\ x_5)$

$S_1 = e^{x_1 - m_1} + e^{x_2 - m_1}$ $\quad S_2 = e^{x_3 - m_2} + e^{x_4 - m_2} + e^{x_5 - m_2}$



$$\mathrm{softmax}(QK^T) \qquad V$$



Can we express $\mathrm{softmax}([x_1\ x_2\ x_3\ x_4\ x_5\ x_6])\ [v_1\ v_2\ v_3\ v_4\ v_5\ v_6]$

$$\mathrm{softmax}([x_1\ x_2\ x_3]) \cdot [v_1\ v_2\ v_3]$$

by

$$\mathrm{softmax}([x_4\ x_5\ x_6]) \cdot [v_4\ v_5\ v_6] \quad ?$$

Recap:

$$\mathrm{softmax}([\,x_1\ x_2\ x_3\ x_4\ x_5\ x_6\,])$$

$$\begin{cases} ① \ \dfrac{e^{m_1}}{e^{m'}}\left[e^{x_1-m_1}\ \ e^{x_2-m_1}\ \ e^{x_3-m_1}\right] : \dfrac{e^{m_2}}{e^{m'}}\left[e^{x_4-m_2}\ \ e^{x_5-m_2}\ \ e^{x_6-m_2}\right] \\ ② \ m' = \max(m_1, m_2) \\ ③ \ s' = \dfrac{e^{m_1}}{e^{m'}} s_1 + \dfrac{e^{m_2}}{e^{m'}} s_2 \end{cases}$$

Computed independently

So $\mathrm{softmax}([x_1\, x_2\, x_3\, x_4\, x_5\, x_6]) \cdot [v_1\ v_2\ v_3\ v_4\ v_5\ v_6]$

will be represented as:

$$\begin{cases} ① \ \dfrac{e^{m_1}}{e^{m'}}\left(\boxed{v_1 e^{x_1-m_1} + v_2 e^{x_2-m_1} + v_3 e^{x_3-m_1}}\right) + \dfrac{e^{m_2}}{e^{m'}}\left(\boxed{v_4 e^{x_4-m_2} + v_5 e^{x_5-m_2} + v_6 e^{x_6-m_2}}\right) \\ ② \ m' = \max(m_1, m_2) \\ ③ \ s' = \dfrac{e^{m_1}}{e^{m'}} s_1 + \dfrac{e^{m_2}}{e^{m'}} s_2 \end{cases}$$

Sanity check: $\dfrac{①}{③}$ will be the actual softmax result