



Linear Attention using Polynomials

• Given  $Q, K, V \in \mathbb{R}^{n \times h}$

• Compute  $Q^{\otimes p}$  and  $K^{\otimes p}$

$$\bullet \text{LT}(Q^{\otimes p} \cdot (K^{\otimes p})^{\top}) \cdot V \text{ in } \mathcal{O}(nh^{p+1}) \text{ time}$$

• Typically,  $h = 64, 128, 256$

- Too expensive even for  $p \equiv 4$

• Use sketching to approximate!





# Linear Attention using Polynomials

- Given  $Q, K, V \in \mathbb{R}^{n \times h}$ 
  - Compute  $Q^{\otimes p}$  and  $K^{\otimes p}$
  - $\text{LT}(Q^{\otimes p} \cdot (K^{\otimes p})^T) \cdot V$  in  $O(nh^{p+1})$  time
- Typically,  $h = 64, 128, 256$ 
  - Too expensive even for  $p = 4$
- Use sketching to approximate!

# Sketching for Approximate Matrix Multiplication