Linear Attention using Polynomials

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

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

- $\mathrm{LT}(Q^{\otimes p} \cdot (K^{\otimes p})^{\mathsf{T}}) \cdot V$ in $O(nh^{p+1})$ time

- Typically, $h = 64, 128, 256$

- Too expensive even for $p = 4$

- Use sketching to approximate!

22

Linear Attention using Polynomials

- Given $Q, K, V \in \mathbb{R}^{n \times h}$
  - Compute $Q^{\otimes p}$ and $K^{\otimes p}$
  - $\mathrm{LT}(Q^{\otimes p} \cdot (K^{\otimes p})^\top) \cdot V$ 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