Previous Works

- Performer (Choromanski et al.,) uses a finite-dimensional map φ to approximate exponential
    - Vectors with larger norms require φ with larger dimension

- Other works consider arbitrary φ instead of first defining $\text{sim}(\cdot, \cdot)$
    - $\varphi(x) \doteq \text{elu}(x) + 1$ (Katharopoulos et al. '20), $\varphi(x) \doteq \text{relu}(x)$
    - Model quality is worse compared to softmax

- Is softmax necessary? Do any other functions with similar properties work?

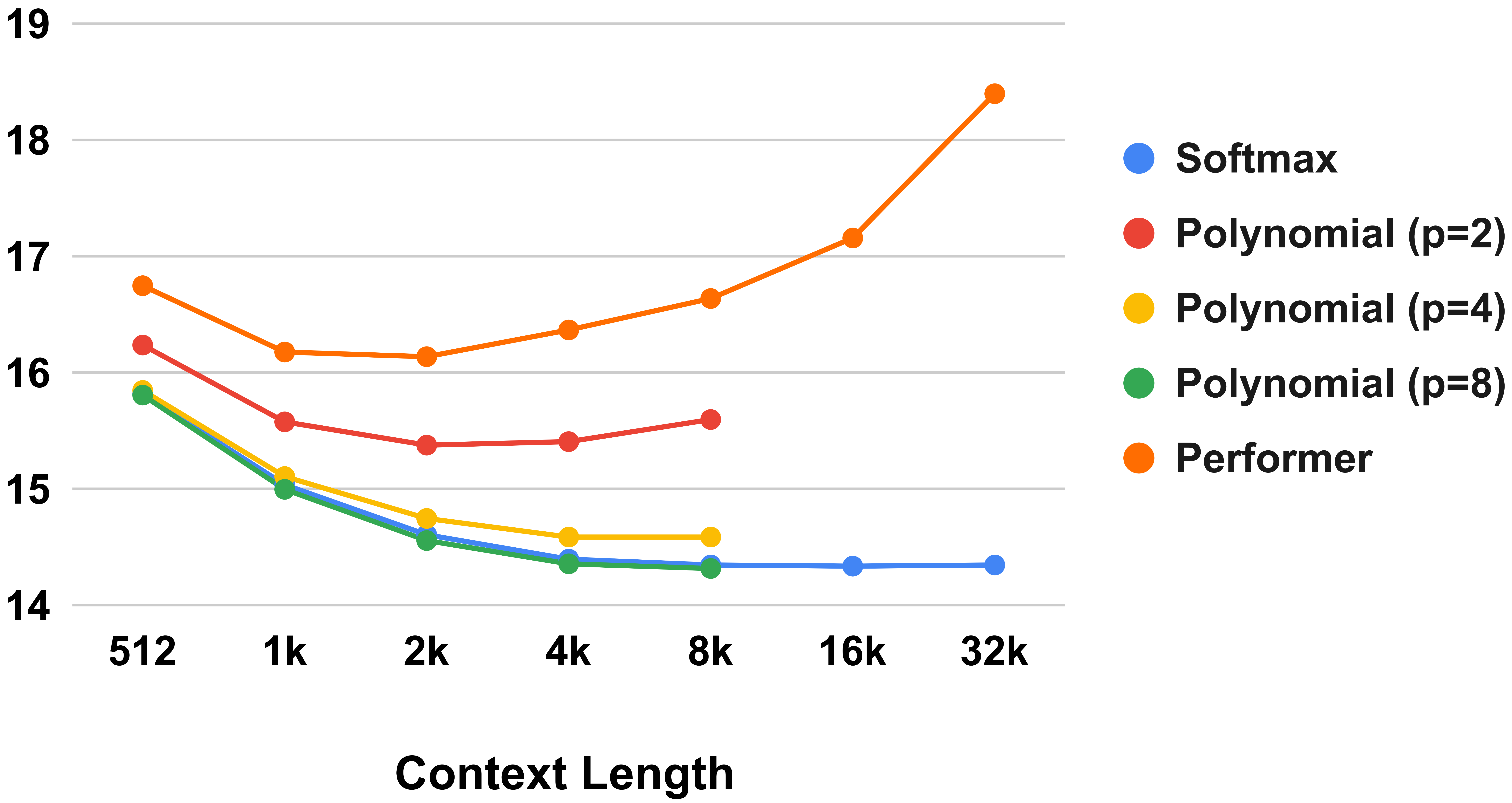
- Consider $\text{sim}(q, k) = \langle q, k \rangle^p$ where $p \geq 2$ is an even integer
    - Always ≥ 0
    - Increases as $\langle q, k \rangle$ goes up

17

Previous Works

- Performer (Choromanski et al.,) uses a finite-dimensional map φ to approximate exponential
  - Vectors with larger norms require φ with larger dimension
- Other works consider arbitrary φ instead of first defining sim$(\,\cdot\,,\cdot\,)$
  - $\varphi(x) \doteq$ elu$(x) + 1$ (Katharopoulos et al. '20), $\varphi(x) \doteq$ relu(x)
  - Model quality is worse compared to softmax
- Is softmax necessary? Do any other functions with similar properties work?
- Consider sim$(q, k) = \langle q, k\rangle^p$ where $p \geq 2$ is an even integer
  - Always ≥ 0
  - Increases as $\langle q, k\rangle$ goes up

Perplexities on Wiki-40B


1k
2k
4k
8k
16k
32k
Context Length
Softmax
Polynomial (p=2)
Polynomial (p=4)
Polynomial (p=8)
Performer