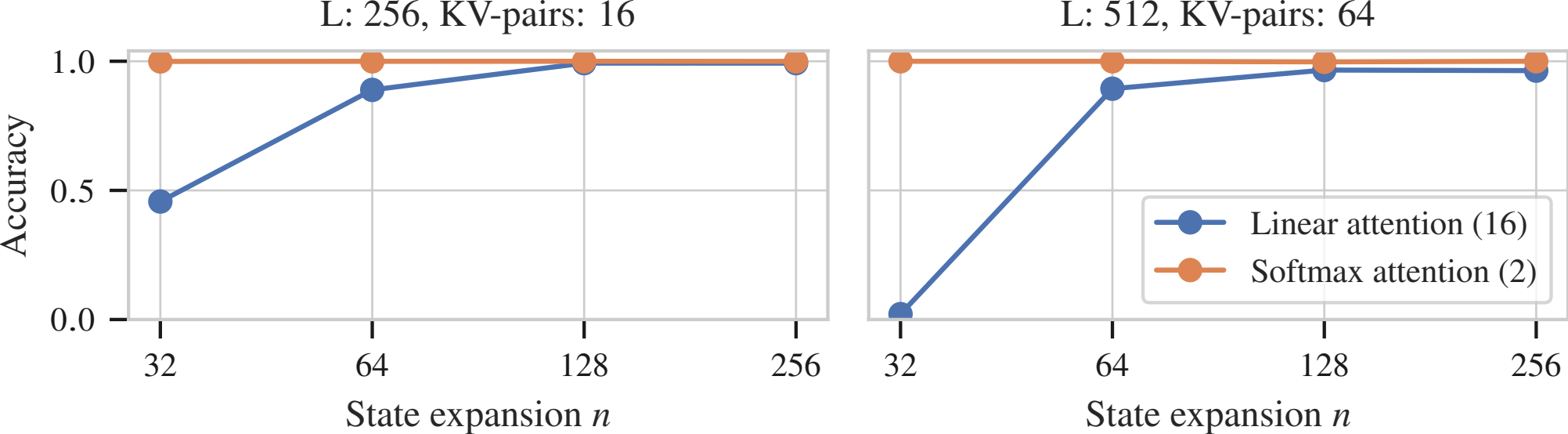
L: 256, KV-pairs: 16
L: 512, KV-pairs: 64
Accuracy
1.0
0.5
0.0
32
64
128
256
State expansion n
Linear attention (16)
Softmax attention (2)