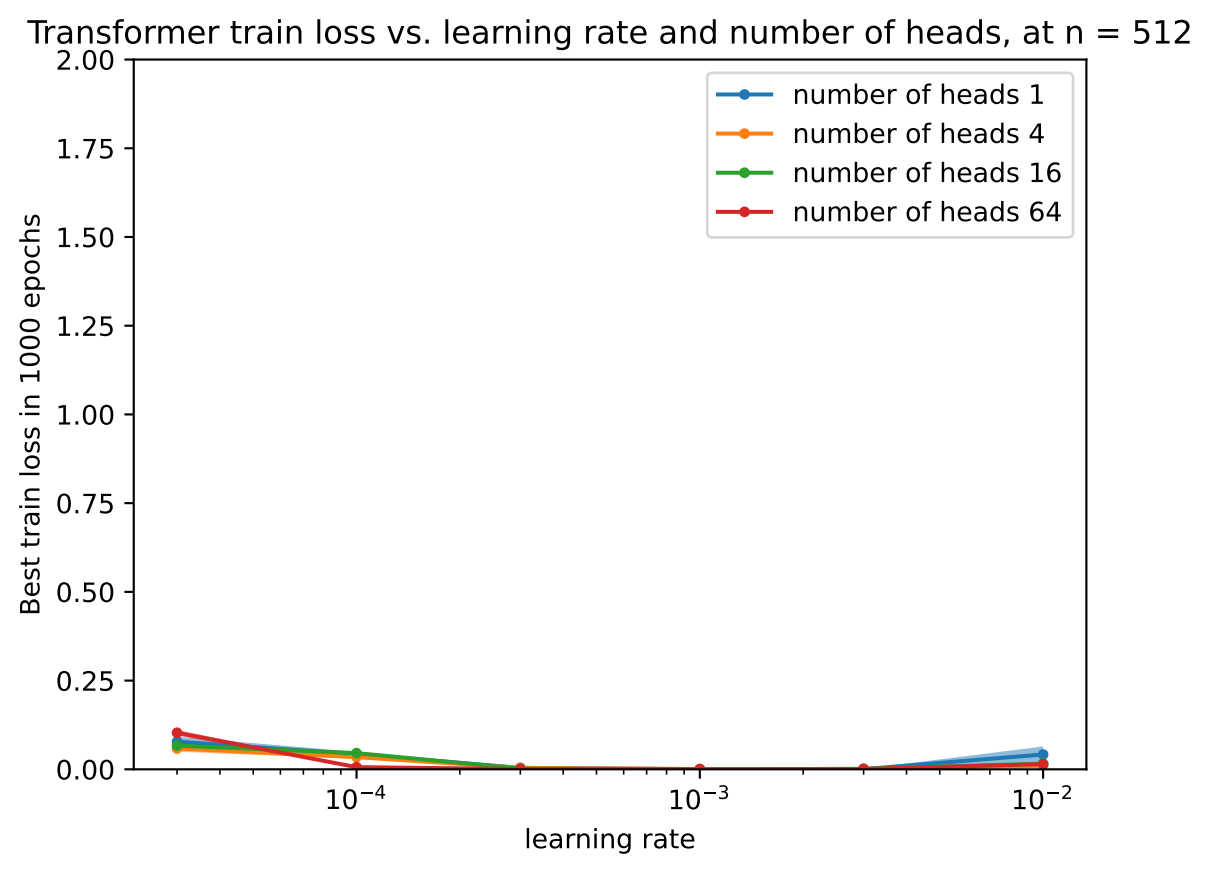
Transformer train loss vs. learning rate and number of heads, at n = 512
number of heads 1
number of heads 4
number of heads 16
number of heads 64
Best train loss in 1000 epochs
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^{-4}
10^{-3}
10^{-2}
learning rate