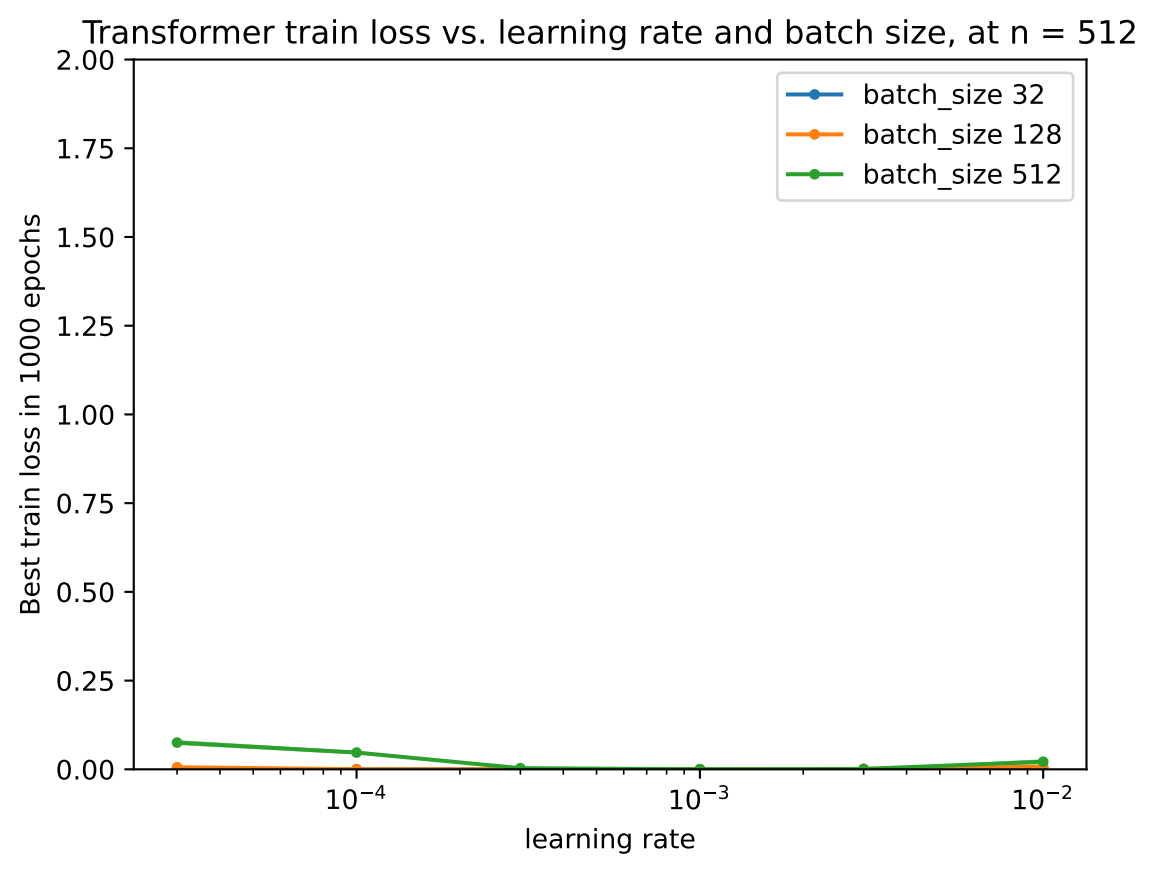
Transformer train loss vs. learning rate and batch size, at n = 512
batch_size 32
batch_size 128
batch_size 512
Best train loss in 1000 epochs
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^{-4}
10^{-3}
10^{-2}
learning rate