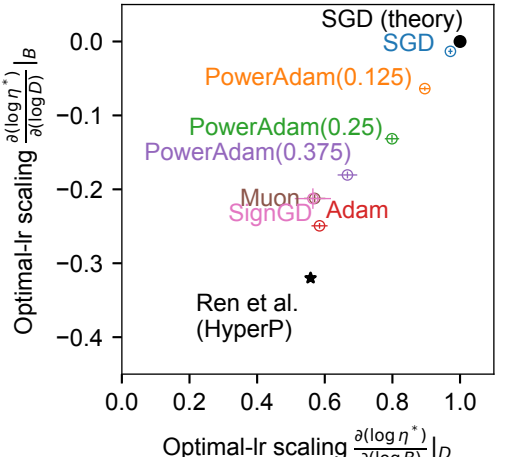

SGD (theory)
SGD
PowerAdam(0.125)
PowerAdam(0.25)
PowerAdam(0.375)
Muon
SignGD
Adam
Ren et al.
(HyperP)
0.0
−0.1
−0.2
−0.3
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
Optimal-lr scaling $\frac{\partial(\log \eta^*)}{\partial(\log D)}|_B$
Optimal-lr scaling $\frac{\partial(\log \eta^*)}{\partial(\log B)}|_D$