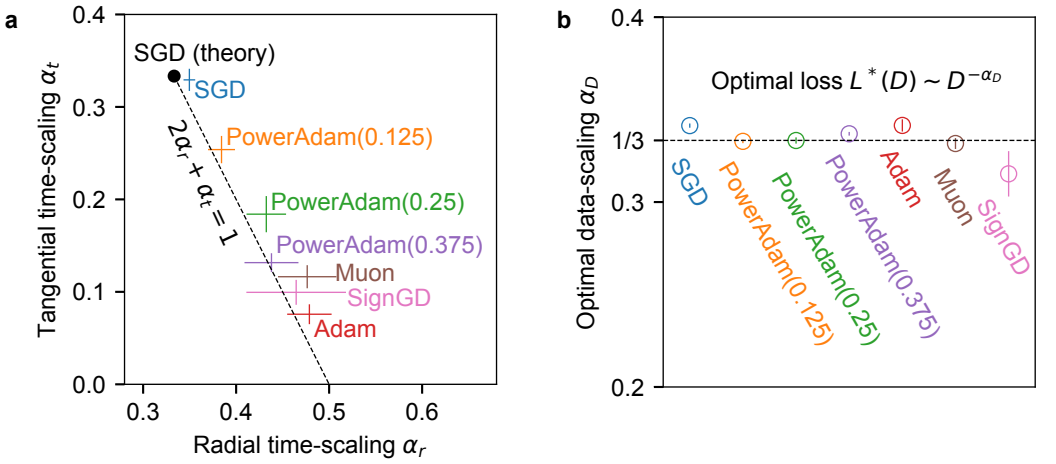

a
SGD (theory)
SGD
PowerAdam(0.125)
PowerAdam(0.25)
PowerAdam(0.375)
Muon
SignGD
Adam
$2\alpha_r + \alpha_t = 1$
Tangential time-scaling α_t
Radial time-scaling α_r
0.0
0.1
0.2
0.3
0.4
0.3
0.4
0.5
0.6
b
Optimal loss $L^*(D) \sim D^{-\alpha_D}$
Optimal data-scaling α_D
1/3
0.3
0.2
0.4
SGD
PowerAdam(0.125)
PowerAdam(0.25)
PowerAdam(0.375)
Adam
Muon
SignGD