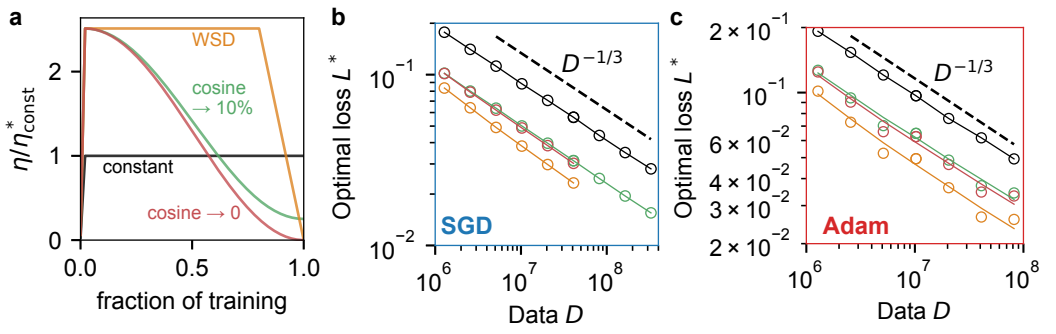

a
WSD
cosine
→ 10%
constant
cosine → 0
$\eta/\eta^*_{\text{const}}$
2
1
0
0.0
0.5
1.0
fraction of training
b
Optimal loss L^*
10^{-1}
10^{-2}
$D^{-1/3}$
SGD
10^6
10^7
10^8
Data D
c
Optimal loss L^*
2×10^{-1}
10^{-1}
6×10^{-2}
4×10^{-2}
3×10^{-2}
2×10^{-2}
$D^{-1/3}$
Adam
10^6
10^7
10^8
Data D