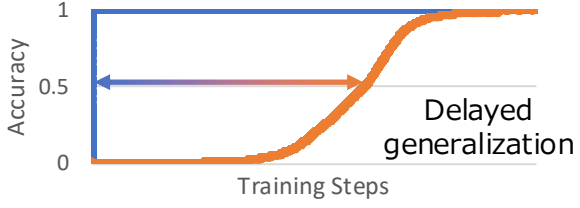
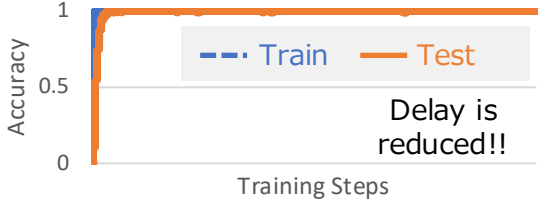
■ Observation (sec 3): Lottery ticket significantly reduces delayed generalization



(a) Dense network



(b) Lottery Ticket (Good Subnetworks)

■ What happen during delayed generalization? Why LT reduce delayed generalization?

H1 [Liu+]: Reducing weight norm	H2 [Merrill+]: Reducing Sparsity	H3 (Ours): Finding good structure
C_{mem} ⇩ C_{gen}	C_{mem} ⇩ C_{gen}	C_{mem} ⇩ C_{gen}
smaller $\|\boldsymbol{\theta}\|_2$	smaller $\|\boldsymbol{\theta}\|_0$	bettr structure

Experiments: sec 4: controlled experiments, sec 5.1 Progress measure using H3, sec 5.2 pruning as force to grokking, and sec 5.3 link to good representation