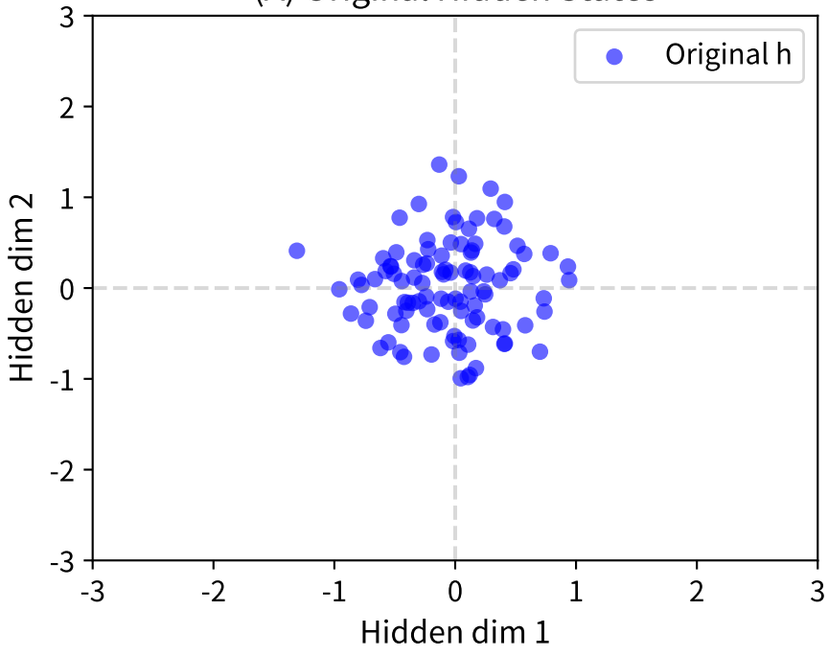
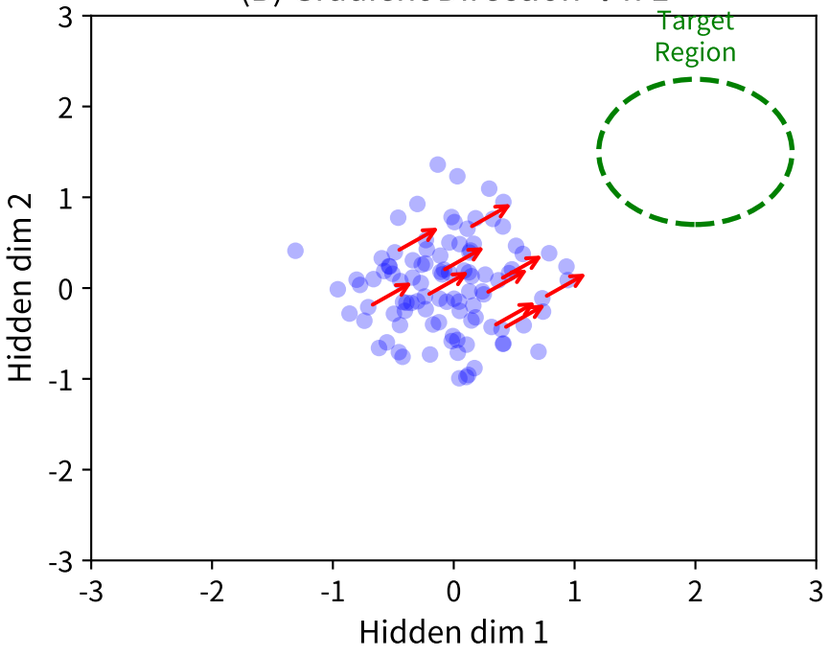
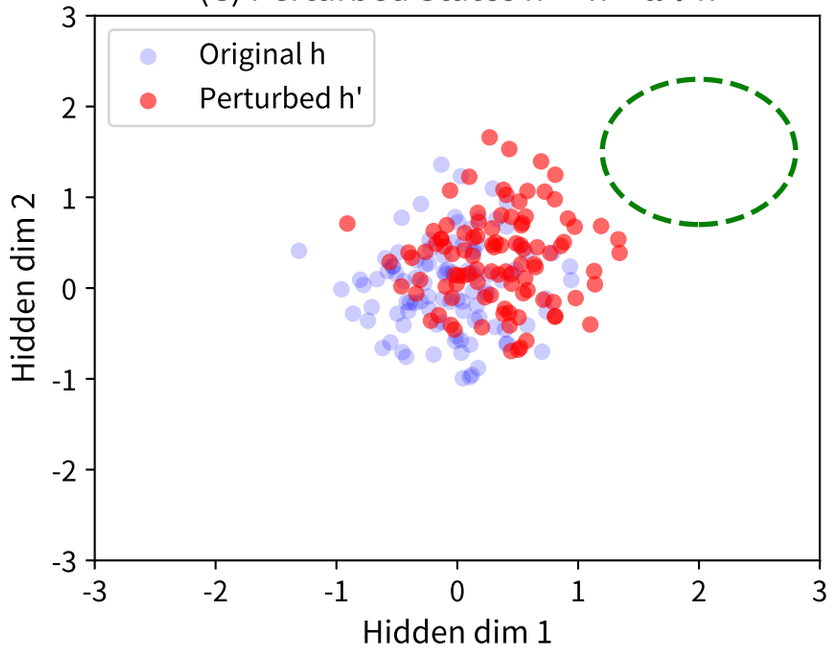

(A) Original Hidden States
Original h
Hidden dim 1
Hidden dim 2


(B) Gradient Direction ∇h L
Target Region
Hidden dim 1
Hidden dim 2


(C) Perturbed States h' = h + α∇h
Original h
Perturbed h'
Hidden dim 1
Hidden dim 2