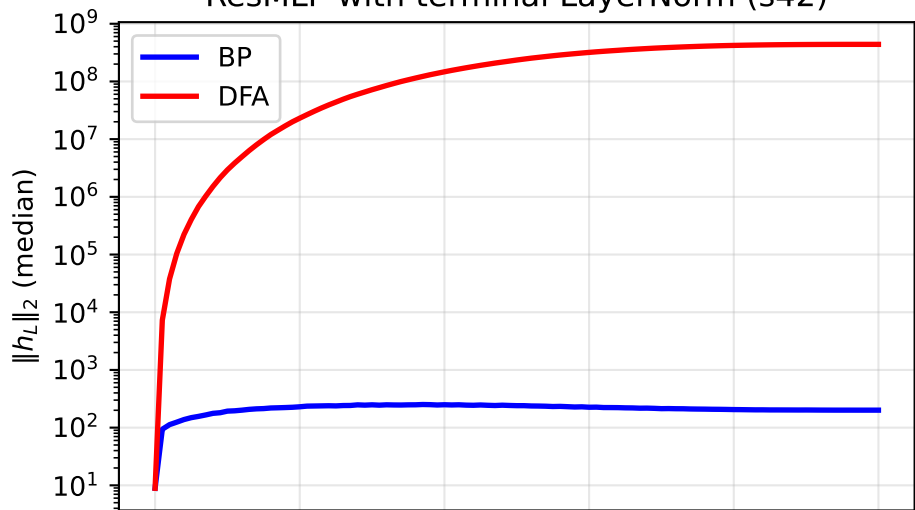


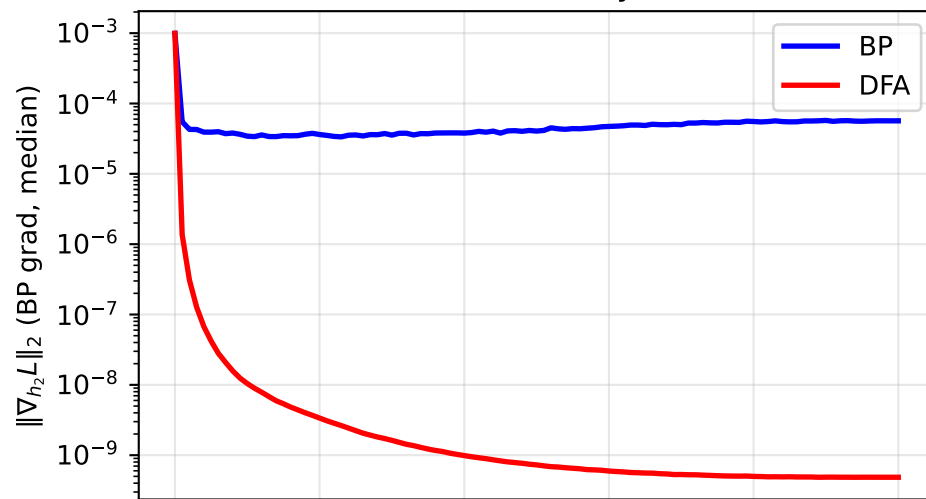
Snapshot evolution: residual stream + BP grad over training

(top: with terminal LN — DFA explodes; bottom: no terminal LN — DFA still grows but BP grad does NOT collapse)

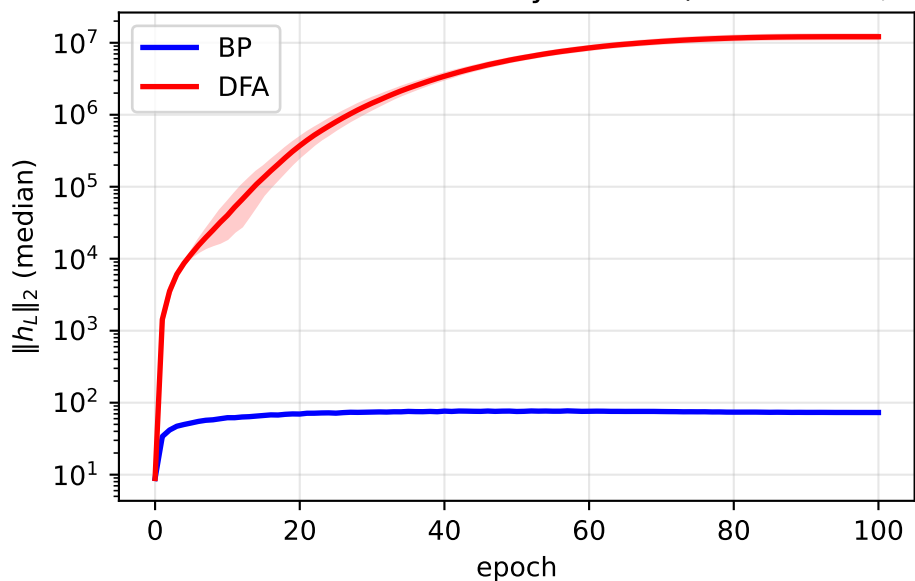
ResMLP with terminal LayerNorm (s42)



ResMLP with terminal LayerNorm (s42)



ResMLP WITHOUT terminal LayerNorm (mean \pm std, n=3)



ResMLP WITHOUT terminal LayerNorm (mean \pm std, n=3)

