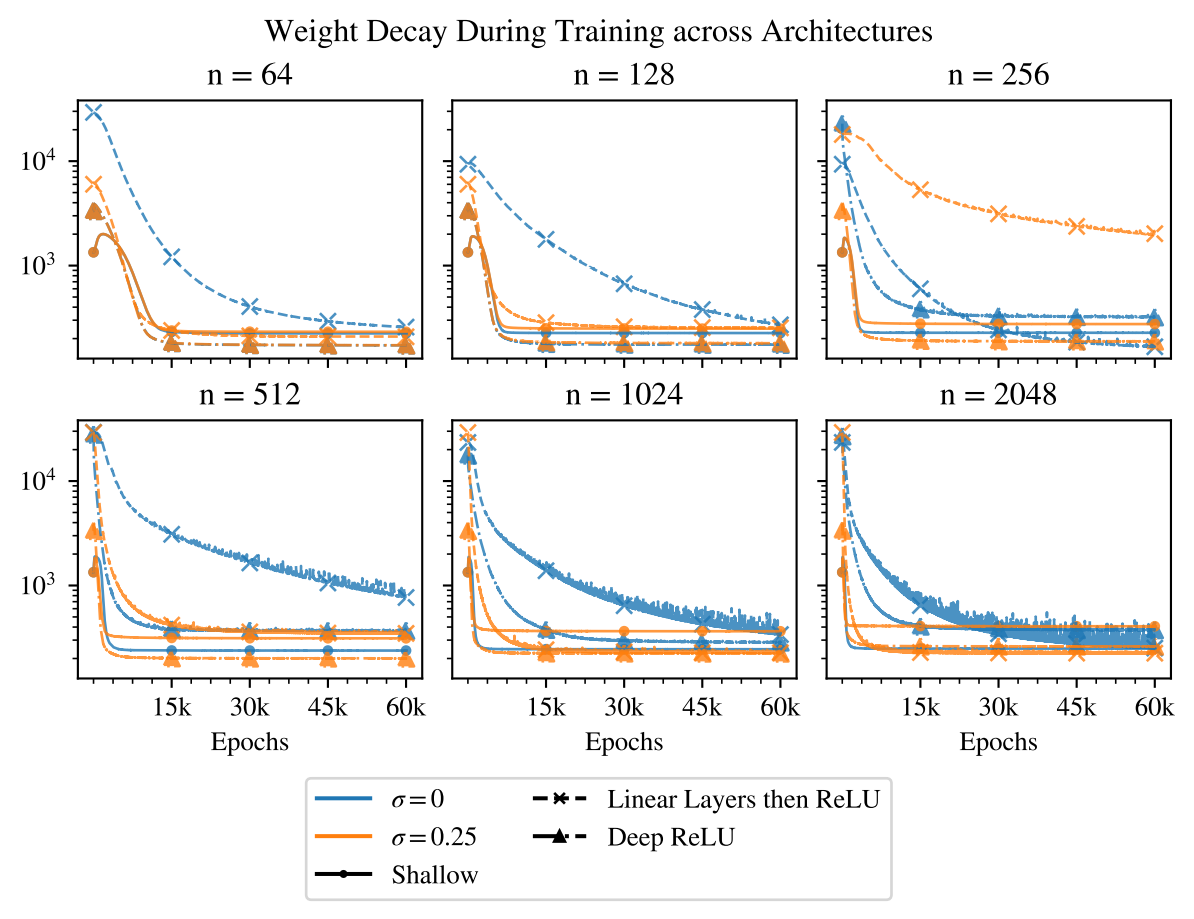
Weight Decay During Training across Architectures
n = 64
n = 128
n = 256
n = 512
n = 1024
n = 2048
10^4
10^3
15k
30k
45k
60k
Epochs
$\sigma = 0$
$\sigma = 0.25$
Shallow
Linear Layers then ReLU
Deep ReLU