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