Generalization across Architectures In-Distribution Out-of-Distribution 10^{0} 10^{-2} 10^{-4} 10^{0} 10^{-1} 2^{8} 29 26 29 27 28 2^6 2^{7} 2^{10} 2^{11} 2^{10} 2^{11} Number of training samples (n) Number of training samples (n) $\sigma = 0$ Linear Layers then ReLU $\sigma = 0.25$ Deep ReLU Shallow σ^2 , irreducible error