Generalization with SGD Training In-Distribution 1.0 10^{0} 0.8 10^{-1} 0.6 MSE 10^{-2} 0.4 10^{-3} 0.2 0.0 1.0 0.8 10^{0} 0.6 0.4 0.2 10^{-1} 0.0 2^7 $\frac{1}{2^{6}}$ 2^{T} 2^{7} 28 2⁹ 2^{10} 2^{8} 29 2^{10} 2^{6} 2^{11} Number of training samples (n) $\sigma = 0$ with extra linear layers $\sigma = 0.25$ σ^2 , irreducible error without extra linear layers