Generalization with SGD Training In-Distribution Out-of-Distribution 10^{0} MSE 10^{-2} 10^{1} 10^{0} 10^{-1} $\frac{1}{2^{6}}$ 2^{7} 2^{10} 2^{7} 28 29 2^8 2^{9} 2^{11} 2^{10} 2^{i_1} 2^6 Number of training samples (n) Number of training samples (n) $\sigma = 0$ with extra linear layers $\sigma = 0.25$ σ^2 , irreducible error without extra linear layers