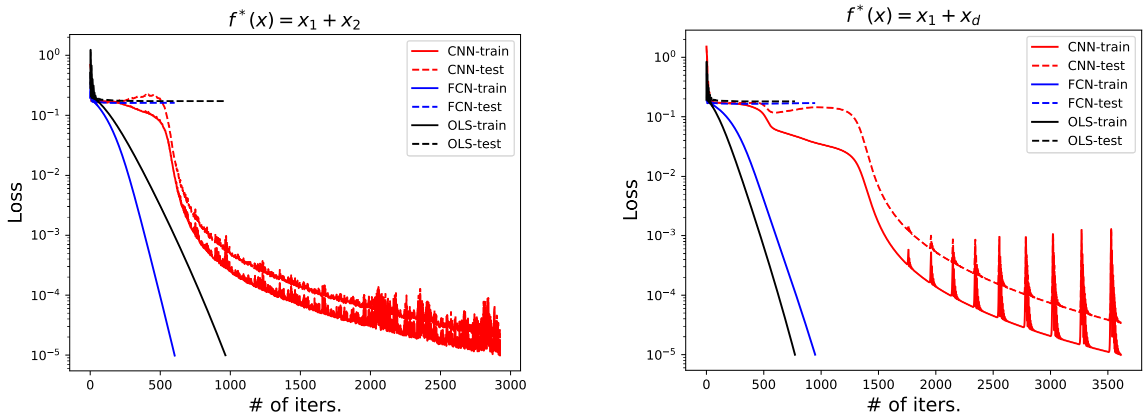

Figure 1: **CNN can learn sparse functions efficiently.** In this experiment, both short (**left**) and long (**right**) range interactions are considered and $d = 4096, n = 400$ and the noise is zero. Adam optimizer is used and *no regularization* is applied. For CNN, the filter size is $s = 4$ and as a result, the depth is $L = \log_4(d) = 6$; the number of channels is set to $C = 4$ across all layers. As a comparison, we also consider fully-connected networks (FCNs) whose architecture is given by $d \to 10 \to 1$ and ordinary the least linear regression (OLS). **Observation:** We can see that even without any explicit sparsity regularization, CNN can still learn sparse interactions efficiently for both short and long range cases. In contrast, FCN and OLS fail.

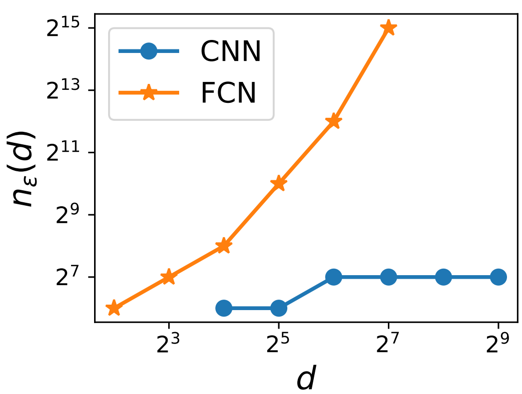

Figure 2: **The sample complexity separation for CNNs and FCNs.** A numerical comparison between fully-connected networks (FCNs) and CNNs for learning $f^*(x) = \sum_{i=1}^{d/2} x_i^2 - \sum_{j=d/2+1}^{d} x_j^2$. The x-axis and y-axis denote the input dimension d and sample complexity $n_\varepsilon(d)$ (defined as the number of samples required to achieve the target error ε), respectively. In experiments, we set ε=1e-3. We observe that when increasing d, $n_\varepsilon(d)$ keeps nearly unchanged for CNNs but increases significantly for FCNs. These results align well with our theoretical predictions.