Task complexity shapes internal representations and robustness in neural networks

Neural networks excel across a wide range of tasks, yet remain "black boxes". In particular, how their internal representations are shaped by the complexity of the input data and the problems they solve remains obscure. In this work, we introduce a suite of five data-agnostic probes—pruning, binarization, noise injection, sign flipping, and bipartite network randomization—to quantify how task difficulty influences the topology and robustness of representations in multilayer perceptrons (MLPs). MLPs are represented as signed, weighted bipartite graphs from a network science perspective. We contrast easy and hard classification tasks on the MNIST and Fashion-MNIST datasets. We show that binarizing weights in hard-task models collapses accuracy to chance, whereas easy-task models remain robust (Fig 1a,b). We also find that pruning low-magnitude edges in binarized hard-task models reveals a sharp phase-transition in performance (Fig 1a,b). Moreover, moderate noise injection can enhance accuracy (Fig 1c,d), similar to a stochastic-resonance effect linked to optimal sign flips of small-magnitude weights (Fig 1e,f). Finally, preserving only the sign structure—instead of precise weight magnitudes—through bipartite network randomization suffices to maintain high accuracy. These phenomena define a model- and modality-agnostic measure of task complexity: the performance gap between full-precision and binarized or shuffled neural network performance. Our findings highlight the crucial role of signed bipartite topology in learned representations and suggest practical strategies for model compression and interpretability that align with task complexity.

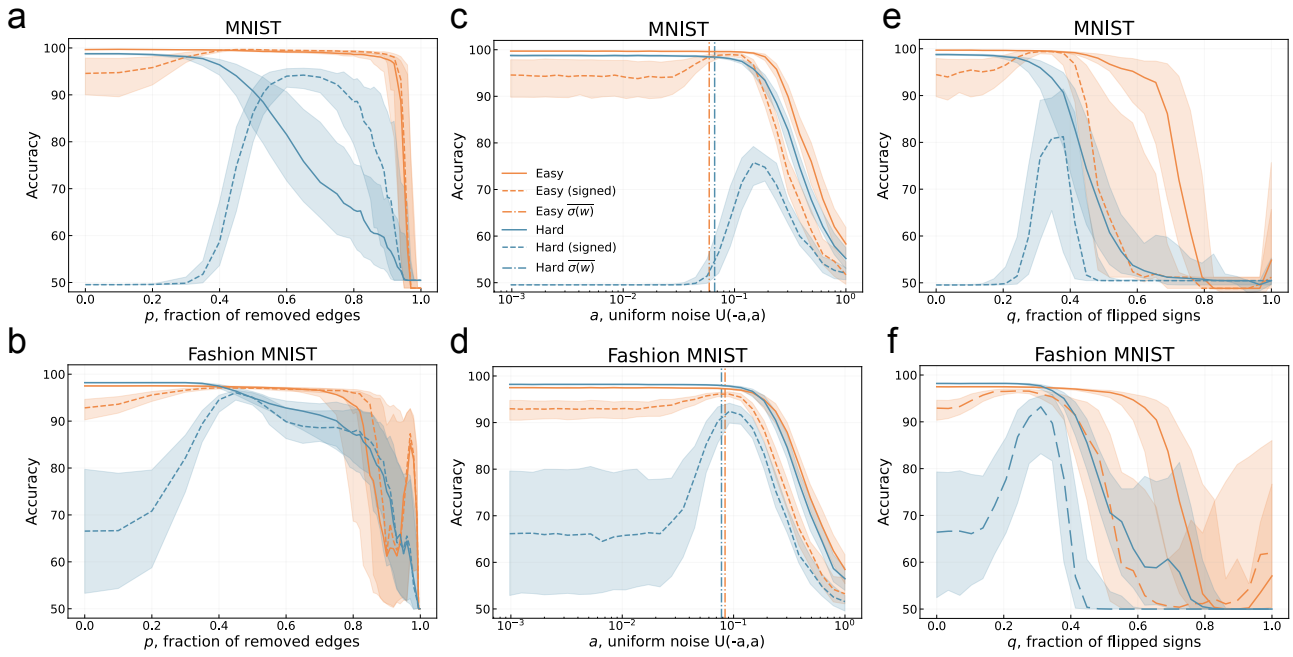

Figure 1: (**a, b**) Pruning experiment. Test accuracy as a function of the fraction of the smallest-magnitude weights removed. (**c, d**) Noise injection experiment. The test accuracy as a function of the additive uniform noise level injected into the weights. The vertical lines show the average standard deviation of the weights. (**e, f**) Sign flipping experiment. The test accuracy as a function of the fraction of the smallest-magnitude sign flipped. All curves are averaged over 100 random initializations. Shaded regions denote the interquartile range (IQR), and the solid lines represent the median.