The Abdus Salam International Centre for Theoretical Physics

United Nations Educational, Scientific and Cultural Organization

IAEA International Atomic Energy Agency

Sebastian Goldt

Title: The Gaussian world is not enough - how data shapes neural network representations

What do neural networks learn from their data ? We discuss this question in two learning paradigms: supervised classification with feed-forward networks, and masked language modelling with transformers. First, we give analytical and experimental evidence for a “distributional simplicity bias”, whereby neural networks learn increasingly complex distributions of their inputs. We then show that neural networks learn from the higher-order cumulants (HOCs) more efficiently than lazy methods, and show how HOCs shape the learnt features. We finally characterise the distributions that are learnt by single- and multi-layer transformers, and discuss implications for learning dynamics and transformer design.