ÁLVARO RIBOT, Harvard University
Do neural models prefer their submodels?

A neural model is the set of functions that can be expressed by a given neural network architecture. Neural models contain natural submodels, such as those obtained by setting some neurons to be inactive. We study the geometry of neural models and ask whether they prefer their submodels. By this we mean whether local minimizers of the loss function lie in a given submodel. We introduce a geometric framework to determine if this occurs. For shallow multilayer perceptrons, we show that subnetworks with inactive hidden neurons are never preferred. For deep multilayer perceptrons, we give conditions under which subnetworks with inactive hidden neurons are not preferred. For lightning self-attention, lower-rank submodels are usually not preferred, but we identify examples where they are because they lie on a boundary of the model. Our results show that, aside from a few exceptions, preference for submodels cannot be explained by the geometry of the neural model, and must instead come from other sources such as training dynamics. This is joint work with Yulia Alexandr and Anna Seigal.