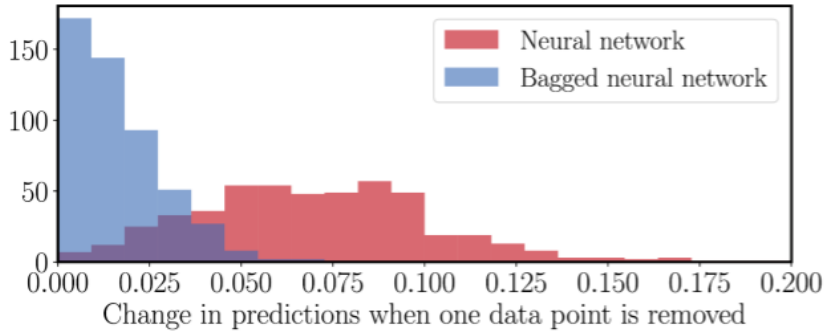
Making Machine Learning Stable

Algorithmic stability is an important concept in data science that refers to how making changes to the training data affects the performance of a learned model. It helps to ensure that a model performs well on new data, can be explained, and has reproducible results with uncertainty estimates. Bootstrap aggregating, or "bagging", is an approach that was developed in the 1990s to address this issue. It involves repeatedly learning a model for different subsets or variations of the training data and averaging the resulting predictions. While bagging has been shown to be effective in practice, it was difficult to guarantee its stability without making strong assumptions about the data or learning algorithms being used — making it inapplicable in many modern data science settings.

Our recent work shows that bagging is guaranteed to make *any* machine learning algorithm stable, regardless of the data set or base algorithm being used. This means that if we remove or replace a small fraction of the training data at random, the resulting prediction will typically change very little. This result has important implications for machine learning, as it can facilitate accuracy guarantees for new data, explainable algorithms, reproducible research, and uncertainty quantification. Bagging is widely applied in a broad range of machine learning settings to help ensure stability, and our result guarantees that we can rely on this common practice to achieve stability without requiring any assumptions.



Histogram of leave-one-out errors for a neural network (red) and a bagged neural network (blue). Leaving one training sample out frequently causes neural network outputs to change dramatically (red histogram), whereas our bagging approach guarantees almost all changes will be very small.

By: Jake A. Soloff (IFDS Postdoc), Rina Foygel Barber (IFDS Co-PI), Rebecca Willett (IFDS Co-PI)

NSF Award NSF DMS-2023109

Reference: https://arxiv.org/abs/2301.12600