Accurate Scaled Summation: Identifying a method that reduces Floating Point Error

Emmet McDonald, Master in Computer Science

University of Dublin, Trinity College, 2024

Supervisor: David Gregg

In any summation problem, Floating Point Error can occur, which reduces the accuracy of the final output. This is obviously undesirable as these outputs are usually used, and their accuracy would lead to better results. This project specifically deals with Scaled Summation problems, where the inputs being summed are weighted by known weights. Such a project is important in regards to the topic of Computer Science as many neural networks can be seen as a series of interconnected Scaled Summation problems, and reducing the Floating Point Error present in them will only lead to more accurate models being produced. In this paper I discuss several summation methods and their effect on the final output's Floating Point Error, and I introduce two key ideas with an aim to reduce this Floating Point Error further. While the first idea, which relies on the Expected Mean of the outcome, generally introduces ∼12-20% more absolute error and ∼30% more relative error, the second idea, which creates a "Hyper"-Sorted permutation of weights for Pairwise Summation, generates only ∼66% of the absolute error and only ∼38% of the relative error that would be generated with an unsorted set of weights..