Verifying Deep Mathematical Properties of AI Systems

Alex Aiken, Stanford University

May 7, 2015

Abstract

It seems likely that any AI system will involve substantial software complexity, will depend on advanced mathematics in both its implementation and justification, and will be naturally flexible and seem to degrade gracefully in the presence of many types of implementation errors. Thus we face a fundamental challenge in developing trustworthy AI: how can we build and maintain complex software systems that require advanced mathematics in order to implement and understand, and which are all but impossible to verify empirically? We believe that it will be possible and desirable to formally state and prove that the desired mathematical properties hold with respect to the underlying programs, and to maintain and evolve such proofs as part of the software artifacts themselves. We propose to demonstrate the feasibility of this methodology by implementing several different certified inference algorithms for probabilistic graphical models, including the Junction Tree algorithm, Gibbs sampling, Mean Field, and Loopy Belief Propagation. Each such algorithm has a very different notion of correctness that involves a different area of mathematics. We will develop a library of the relevant formal mathematics, and then for each inference algorithm, we will formally state its specification and prove that our implementation satisfies it.