Abstract

AI systems do more than execute formal instructions: they write, translate, code and perform cognitive work. Whether or not this is intelligence in the human sense, such operations are now available through a programmable interface. Intelligence is becoming infrastructure, but infrastructure built from data, computation, communication and energy. The distribution of these resources shapes who can build, improve and use it.

Efficiency therefore affects AI's economic and environmental cost and who controls its development. Companies and governments are committing datacenters, energy and supply chains in expectation of lasting value. This investment may concentrate benefits and control; if expectations fail, the consequences extend beyond AI research.

This thesis considers two points where resource distribution becomes an algorithmic problem: training across data owners and deployment on different hardware. During training, users or organizations hold different data, communication is expensive and only some participate. The first contribution analyzes SGD with momentum, a classical deep-learning optimizer, under data heterogeneity and partial participation. It then develops Generalized Heavy-Ball Momentum (GHBM), a generalization that supports reliable and communication-efficient collaboration without centralizing data and computation.

At deployment, one model must instead serve devices with different budgets. The second contribution formalizes this requirement through model elasticity and develops FlexRank, which offers several accuracy–cost operating points from one shared parameterization. The model can adapt to the available hardware without maintaining a separate network for every device. Both contributions adapt learning systems to available resources, broadening where machine intelligence can be produced and used.