In distributed machine learning, data is dispatched to multiple machines for processing. Motivated by the fact that similar data points often belong to the same or similar classes, and more generally, classification rules of high accuracy tend to be "locally simple but globally complex" (Vapnik and Bottou, 1993), we propose data dependent dispatching that takes advantage of such structure. We present an in-depth analysis of this model, providing new algorithms with provable worst-case guarantees, analysis proving existing scalable heuristics perform well in natural non worst-case conditions, and techniques for extending a dispatching rule from a small sample to the entire distribution. We overcome novel technical challenges to satisfy important conditions for accurate distributed learning, including fault tolerance and balancedness. We empirically compare our approach with baselines based on random partitioning, balanced partition trees, and locality sensitive hashing, showing that we achieve significantly higher accuracy on both synthetic and real world image and advertising datasets. We also demonstrate that our technique strongly scales with the available computing power.

This is joint work with Mu Li, Krishna Pillutla, Colin White, Nina Balcan, and Alex Smola.