On the Optimality of Averaging in Distributed Statistical Learning

Jonathan Rosenblatt, Weizmann Institute

A common approach to statistical learning on big data is to randomly split it among m machines and calculate the parameter of interest by averaging their m individual estimates.

Focusing on empirical risk minimization, or equivalently M-estimation, we study the statistical error incurred by this strategy.
We consider two asymptotic settings:
one where the number of samples per machine but the number of parameters p is fixed, and a second high-dimensional regime where both with .

Most previous works provided only moment bounds on the error incurred by splitting the data in the fixed p setting. In contrast, we present for both regimes asymptotically exact distributions for this estimation error. In the fixed-p setting, under suitable assumptions, we thus prove that to leading order, averaging is as accurate as the centralized solution.
In the high-dimensional setting, we show a qualitatively different behavior:
data splitting does incur a first order accuracy loss, which we quantify precisely.
In addition, our asymptotic distributions allow the construction of confidence intervals and hypothesis testing on the estimated parameters.

Our main conclusion is that in both regimes, averaging parallelized estimates is an attractive way to speedup computations and save on memory, while incurring a quantifiable and typically moderate excess error.