The Big Data Bootstrap

Ariel Kleiner
Ameet Talwalkar, Purnamrita Sarkar
Michael I. Jordan

UC Berkeley

The Setting

Observe data $X_1, \ldots, X_n$

Form an estimate $\hat{\theta}_n = \theta(X_1, \ldots, X_n)$
(e.g., θ could be a classifier)

Want to compute an assessment ξ of the quality of $\hat{\theta}_n$
(e.g., ξ could compute a confidence region)

Our Goal

A procedure for quantifying estimator quality which is

accurate
automatic
scalable

The Unachievable Ideal

Ideally, we would

1. Observe many independent datasets of size *n*.
2. Compute $\hat{\theta}_n$ on each.
3. Compute ξ based on these multiple realizations of $\hat{\theta}_n$.

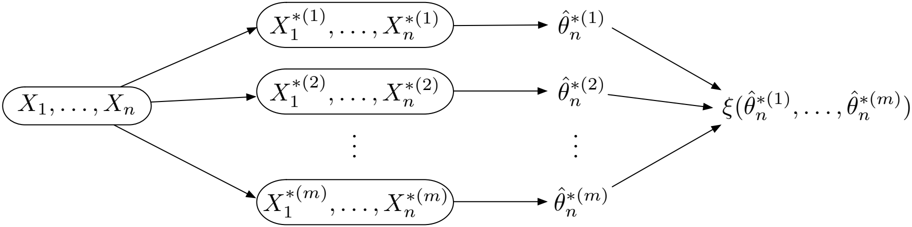
$$
\begin{tikzcd}
X_1^{(1)}, \ldots, X_n^{(1)} \arrow[r] & \hat{\theta}_n^{(1)} \arrow[dr] & \\
X_1^{(2)}, \ldots, X_n^{(2)} \arrow[r] & \hat{\theta}_n^{(2)} \arrow[r] & \xi(\hat{\theta}_n^{(1)}, \ldots, \hat{\theta}_n^{(m)}) \\
\vdots & \vdots & \\
X_1^{(m)}, \ldots, X_n^{(m)} \arrow[r] & \hat{\theta}_n^{(m)} \arrow[uur] &
\end{tikzcd}
$$

But, we only observe *one* dataset of size *n*.

Prior Work: The Bootstrap

Use the observed data to simulate multiple datasets of size n:

1. Repeatedly *resample n* points *with replacement* from the original dataset of size n.
2. Compute $\hat{\theta}_n^*$ on each resample.
3. Compute ξ based on these multiple realizations of $\hat{\theta}_n^*$ as our estimate of ξ for $\hat{\theta}_n$.



Prior Work: The Bootstrap

Computational Issues

- Expected number of distinct points in a bootstrap resample is $\sim 0.632n$.
- Resources required to compute estimate generally scale in number of *distinct* data points.
  - This is true of many commonly used learning algorithms (e.g., SVM, logistic regression, linear regression, kernel methods, general M-estimators, etc.).
  - Use weighted representation of resampled datasets to avoid physical data replication.
  - Example: If original dataset has size 1 TB, then expect resample to have size $\sim$ 632 GB.

Prior Work: The Bootstrap

Computational Issues

Suppose that the original dataset has size 1 TB. The bootstrap does the following:

```
for i ← 1 to 300
      resample ∼ 632 GB of data
      compute θ̂*_n on resample
compute ξ based on the resampled θ̂*_n's
```

Prior Work: The Bootstrap

Advantages

- Accurate for a wide range of estimators.
- Automatic: can compute without knowledge of estimator internals.

Disadvantages

- Must repeatedly compute estimates on $\sim 63\%$ of the data.
- For big data, difficult to parallelize across different estimate computations.

Prior Work: The *b* out of *n* Bootstrap

Compute estimates only on smaller resamples of the data of size $b < n$, and analytically correct our quality assessment.

More favorable computational profile than the bootstrap.

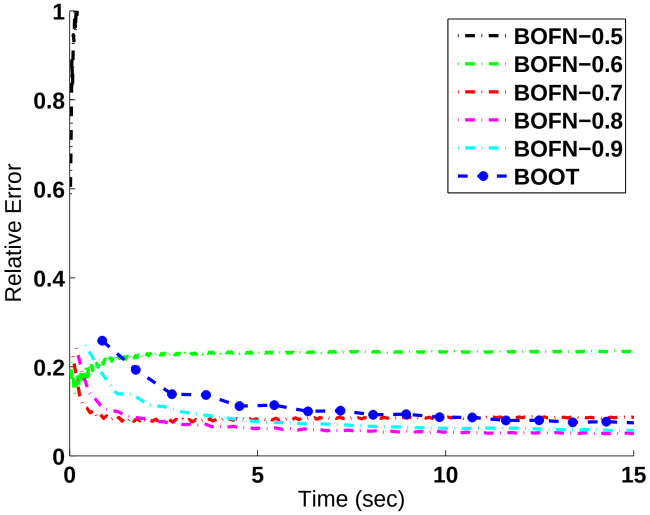
Issues

- Accuracy sensitive to choice of *b*.
- Still fairly automatic, though analytical correction introduces some dependency on estimator internals.

Empirical Results: Bootstrap and *b* out of *n* Bootstrap

- Multivariate linear regression with $d = 100$ and $n = 20,000$ on synthetic data.
- Estimate parameters $\hat{\theta}_n$ via least squares.
- ξ computes confidence intervals.
- Compare widths to ground truth (via relative error).
- For b out of n bootstrap, use $b = n^\gamma$ for various values of γ.

Empirical Results: Bootstrap and *b* out of *n* Bootstrap

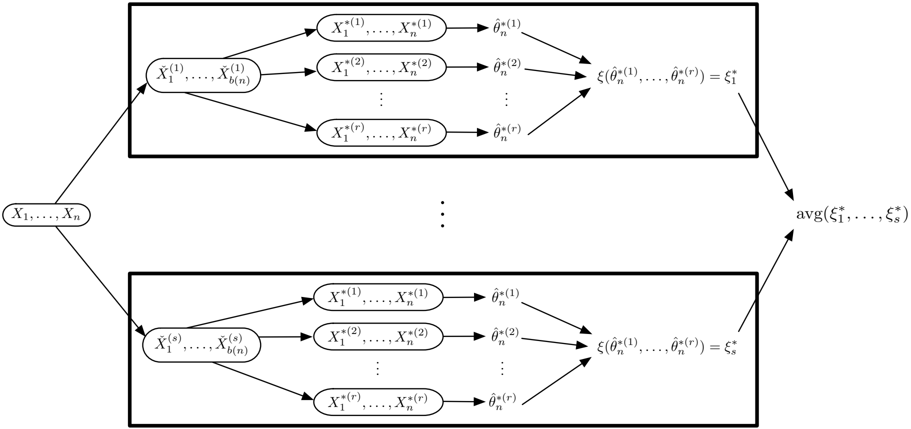


Our Approach: The Bag of Little Bootstraps (BLB)

Use only $b < n$ data points to compute each resample while maintaining robustness to choice of b:

1. Repeatedly *subsample* $b < n$ points *without replacement* from the original dataset of size n.
2. For each subsample do:
   1. Repeatedly *resample* n points *with replacement* from the subsample.
   2. Compute $\hat{\theta}_n^*$ on each resample.
   3. Compute an estimate of ξ based on these multiple resampled realizations of $\hat{\theta}_n^*$.
3. We now have one estimate of ξ per subsample. Output their average as our final estimate of ξ for $\hat{\theta}_n$.

Our Approach: BLB



Our Approach: BLB

Computational Issues

- Recall: resources required to compute estimate generally scale in number of *distinct* data points.
- Each BLB subsample/resample contains at most $b < n$ distinct points.
- Example: if $n = 1,000,000$, data point size is 1 MB, and we take $b = n^{0.6}$, then
  - full dataset has size 1 TB
  - subsamples/resamples contain at most 3,981 distinct data points and have size at most 4 GB
  - (in contrast, bootstrap resamples have size $\sim$ 632 GB)

Our Approach: BLB

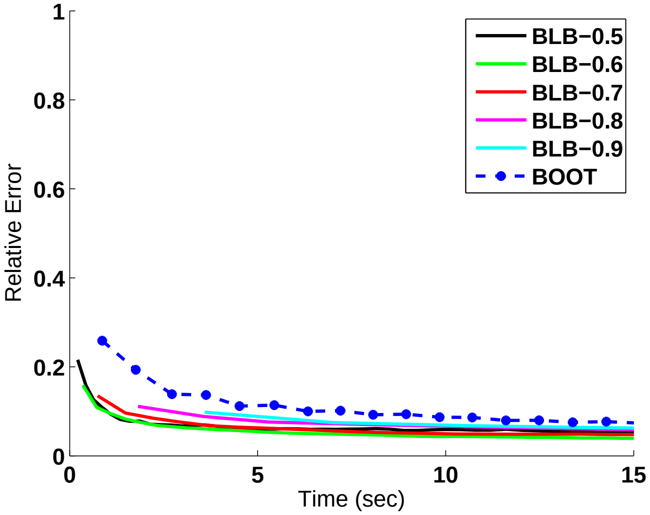
Like the Bootstrap

- Accurate for a wide range of estimators. Shares the bootstrap's consistency and higher-order correctness.
- Automatic: can compute without knowledge of estimator internals.

Beyond the Bootstrap (and *b* out of *n* Bootstrap/Subsampling)

- Can explicitly control b, the amount of data on which we must repeatedly compute estimates; can have $b/n \to 0$ as $n \to \infty$.
- More robust to choice of b, which can be much smaller than n.
- Generally faster than the bootstrap (even if computing serially).
- Easy to parallelize across different estimate computations.

Empirical Results: BLB



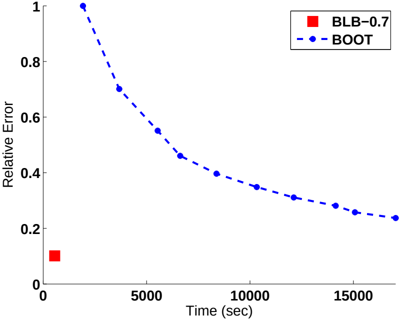
Theoretical Results

BLB shares the bootstrap's favorable statistical properties
(consistency & higher-order correctness)

under the same conditions that have been used in prior analysis of the bootstrap

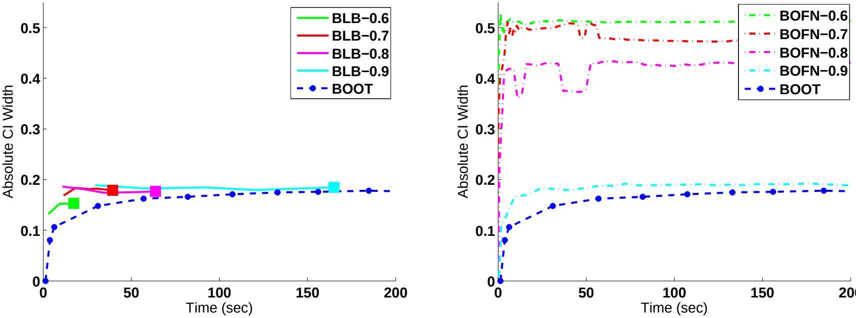
Scalability

10 nodes on Amazon EC2 using Spark; 150 GB of data



Non-Synthetic Data

UCI connect4 dataset: logistic regression, $d = 42, n = 67,557$



More Empirical Results

Logistic Regression