Bootstrapping Big Data

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

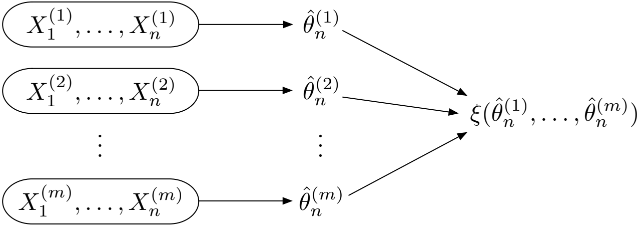
Ideally, we would

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

The Unachievable Ideal

Ideally, we would

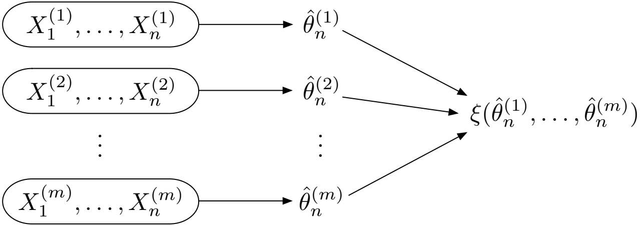
1. Observe many independent datasets of size n
2. Compute $\hat{\theta}_n$ on each
3. Compute ξ based on these multiple realizations of $\hat{\theta}_n$.



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 $\hat{\theta}_n$.



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

Prior Work: The Bootstrap

Use the observed data to simulate multiple datasets 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.

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.

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

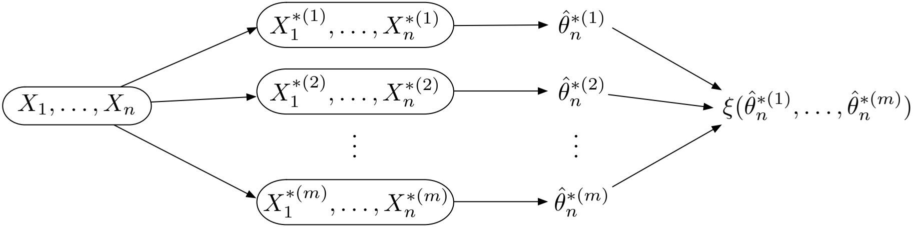
Use the observed data to simulate multiple datasets of size *n*:

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

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 resample is $\sim 0.632n$

Prior Work: The Bootstrap
Computational Issues

- Expected number of *distinct* points in a resample is $\sim 0.632n$
- Resources required to compute θ generally scale in number of *distinct* data points.

- Expected number of *distinct* points in a resample is $\sim 0.632n$
- Resources required to compute θ 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.).

Prior Work: The Bootstrap

Computational Issues

- Expected number of *distinct* points in a resample is $\sim 0.632n$
- Resources required to compute θ 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.

Prior Work: The Bootstrap

Computational Issues

- Expected number of *distinct* points in a resample is $\sim 0.632n$
- Resources required to compute θ 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 θ on resample
compute ξ based on the resampled θ's
```

Prior Work: The Bootstrap

Advantages

- Accurate for a wide range of θ
- Automatic: can compute without knowledge of the internals of θ

Prior Work: The Bootstrap

Advantages

- Accurate for a wide range of θ
- Automatic: can compute without knowledge of the internals of θ

Disadvantages

- Must repeatedly compute θ on $\sim$ 63% of the data
- For big data, difficult to parallelize across different computations of θ (though θ could perhaps be parallelized internally)

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

Compute θ only on smaller resamples of the data of size $b(n) < n$, and analytically correct our uncertainty estimates:

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

Compute θ only on smaller resamples of the data of size $b(n) < n$, and analytically correct our uncertainty estimates:

1. Repeatedly *resample* $b(n) < n$ points *with replacement* from the original dataset of size n.

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

Compute θ only on smaller resamples of the data of size $b(n) < n$, and analytically correct our uncertainty estimates:

1. Repeatedly *resample* $b(n) < n$ points *with replacement* from the original dataset of size n.
2. Compute $\hat{\theta}^*_{b(n)}$ on each resample.

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

Compute θ only on smaller resamples of the data of size $b(n) < n$, and analytically correct our uncertainty estimates:

1. Repeatedly *resample* $b(n) < n$ points *with replacement* from the original dataset of size n.
2. Compute $\hat{\theta}^*_{b(n)}$ on each resample.
3. Compute ξ based on these multiple realizations of $\hat{\theta}^*_{b(n)}$.

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

Compute θ only on smaller resamples of the data of size $b(n) < n$, and analytically correct our uncertainty estimates:

1. Repeatedly *resample* $b(n) < n$ points *with replacement* from the original dataset of size n.
2. Compute $\hat{\theta}^*_{b(n)}$ on each resample.
3. Compute ξ based on these multiple realizations of $\hat{\theta}^*_{b(n)}$.
4. Analytically correct to produce final estimate of ξ for $\hat{\theta}_n$.

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

Compute θ only on smaller resamples of the data of size $b(n) < n$, and analytically correct our uncertainty estimates:

1. Repeatedly *resample* $b(n) < n$ points *with replacement* from the original dataset of size n.
2. Compute $\hat{\theta}^*_{b(n)}$ on each resample.
3. Compute ξ based on these multiple realizations of $\hat{\theta}^*_{b(n)}$.
4. Analytically correct to produce final estimate of ξ for $\hat{\theta}_n$.

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

Compute θ only on smaller resamples of the data of size $b(n) < n$, and analytically correct our uncertainty estimates:

1. Repeatedly *resample* $b(n) < n$ points *with replacement* from the original dataset of size n.
2. Compute $\hat{\theta}^*_{b(n)}$ on each resample.
3. Compute ξ based on these multiple realizations of $\hat{\theta}^*_{b(n)}$.
4. Analytically correct to produce final estimate of ξ for $\hat{\theta}_n$.

Much more favorable computational profile than the bootstrap.

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

Compute θ only on smaller resamples of the data of size $b(n) < n$, and analytically correct our uncertainty estimates:

1. Repeatedly *resample* $b(n) < n$ points *with replacement* from the original dataset of size n.
2. Compute $\hat{\theta}^*_{b(n)}$ on each resample.
3. Compute ξ based on these multiple realizations of $\hat{\theta}^*_{b(n)}$.
4. Analytically correct to produce final estimate of ξ for $\hat{\theta}_n$.

Much more favorable computational profile than the bootstrap.

Issues

- Accuracy sensitive to choice of $b(n)$.
- Still fairly automatic, though analytical correction introduces some dependency on internals of θ.

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

- Multivariate linear regression with $d = 100$ and $n = 20,000$ on synthetic data.

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

- Multivariate linear regression with $d = 100$ and $n = 20,000$ on synthetic data.
- x values sampled independently from coordinate-wise Gamma distributions.

Empirical Results: Bootstrap and b out of n Bootstrap

- Multivariate linear regression with $d = 100$ and $n = 20,000$ on synthetic data.
- x values sampled independently from coordinate-wise Gamma distributions.
- $y = w \cdot x + \epsilon$, where $w \in \mathbb{R}^d$ is a fixed weight vector and ϵ is independent Gamma noise.

Empirical Results: Bootstrap and b out of n Bootstrap

- Multivariate linear regression with $d = 100$ and $n = 20,000$ on synthetic data.
- x values sampled independently from coordinate-wise Gamma distributions.
- $y = w \cdot x + \epsilon$, where $w \in \mathbb{R}^d$ is a fixed weight vector and ϵ is independent Gamma noise.
- Estimate $\hat{\theta}_n = \hat{w} \in \mathbb{R}^d$ via least squares.

Empirical Results: Bootstrap and b out of n Bootstrap

- Multivariate linear regression with $d = 100$ and $n = 20,000$ on synthetic data.
- x values sampled independently from coordinate-wise Gamma distributions.
- $y = w \cdot x + \epsilon$, where $w \in \mathbb{R}^d$ is a fixed weight vector and ϵ is independent Gamma noise.
- Estimate $\hat{\theta}_n = \hat{w} \in \mathbb{R}^d$ via least squares.
- Compute a marginal confidence interval for each component of $\hat{w}$ and assess accuracy via relative mean (across components) absolute deviation from true confidence interval size.

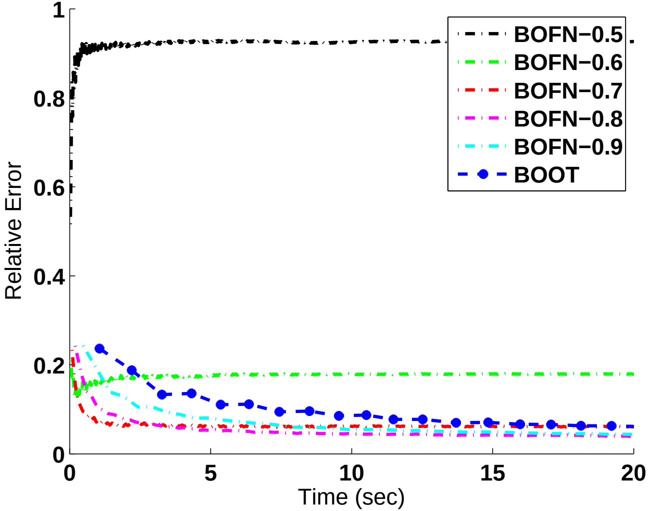
Empirical Results: Bootstrap and b out of n Bootstrap

- Multivariate linear regression with $d = 100$ and $n = 20,000$ on synthetic data.
- x values sampled independently from coordinate-wise Gamma distributions.
- $y = w \cdot x + \epsilon$, where $w \in \mathbb{R}^d$ is a fixed weight vector and ϵ is independent Gamma noise.
- Estimate $\hat{\theta}_n = \hat{w} \in \mathbb{R}^d$ via least squares.
- Compute a marginal confidence interval for each component of $\hat{w}$ and assess accuracy via relative mean (across components) absolute deviation from true confidence interval size.
- For b out of n bootstrap, use $b(n) = n^\gamma$ for various values of γ.

Empirical Results: Bootstrap and b out of n Bootstrap

- Multivariate linear regression with $d = 100$ and $n = 20,000$ on synthetic data.
- x values sampled independently from coordinate-wise Gamma distributions.
- $y = w \cdot x + \epsilon$, where $w \in \mathbb{R}^d$ is a fixed weight vector and ϵ is independent Gamma noise.
- Estimate $\hat{\theta}_n = \hat{w} \in \mathbb{R}^d$ via least squares.
- Compute a marginal confidence interval for each component of $\hat{w}$ and assess accuracy via relative mean (across components) absolute deviation from true confidence interval size.
- For b out of n bootstrap, use $b(n) = n^\gamma$ for various values of γ.
- Similar results obtained with Normal and StudentT data generating distributions, as well as if estimate a misspecified model.

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



Our Approach: The Bag of Little Bootstraps (BLB)

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

Our Approach: The Bag of Little Bootstraps (BLB)

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

1. Repeatedly *subsample* $b(n) < n$ points *without replacement* from the original dataset of size n.

Our Approach: The Bag of Little Bootstraps (BLB)

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

1. Repeatedly *subsample* $b(n) < n$ points *without replacement* from the original dataset of size n.
2. For each subsample do:

Our Approach: The Bag of Little Bootstraps (BLB)

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

1. Repeatedly *subsample* $b(n) < n$ 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.

Our Approach: The Bag of Little Bootstraps (BLB)

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

1. Repeatedly *subsample* $b(n) < n$ 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.

Our Approach: The Bag of Little Bootstraps (BLB)

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

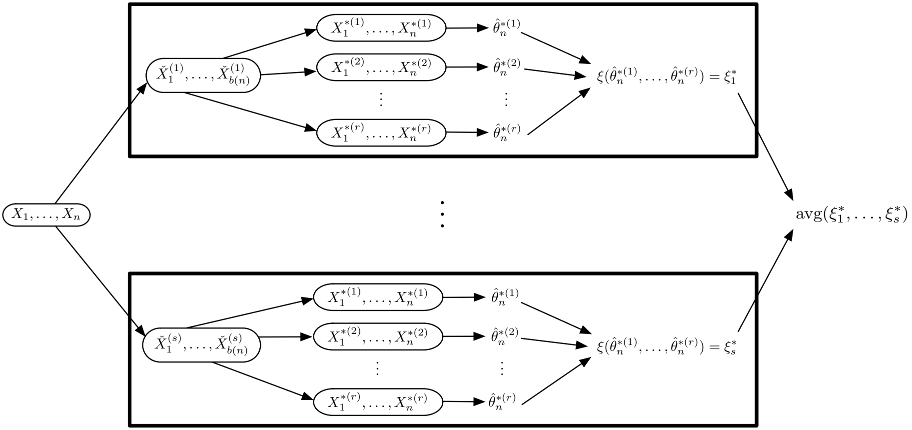
1. Repeatedly *subsample* $b(n) < n$ points *without replacement* from the original dataset of size n.
2. For each subsample do:
    1. Repeatedly *resample* n points *with replacement* from the subsample.
    2. Compute $\hat{\theta}_n^*$ on each resample.
    3. Compute an estimate of ξ based on these multiple resampled realizations of $\hat{\theta}_n^*$.

Our Approach: The Bag of Little Bootstraps (BLB)

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

1. Repeatedly *subsample* $b(n) < n$ 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 θ generally scale in number of *distinct* data points.

Our Approach: BLB

Computational Issues

- Recall: resources required to compute θ generally scale in number of *distinct* data points.
- Each BLB subsample/resample contains at most $b(n) < n$ distinct points.

Our Approach: BLB

Computational Issues

- Recall: resources required to compute θ generally scale in number of *distinct* data points.
- Each BLB subsample/resample contains at most $b(n) < n$ distinct points.
- Example: if $n = 1,000,000$, data point size is 1 MB, and we take $b(n) = n^{0.6}$, then

Our Approach: BLB

Computational Issues

- Recall: resources required to compute θ generally scale in number of *distinct* data points.
- Each BLB subsample/resample contains at most $b(n) < n$ distinct points.
- Example: if $n = 1,000,000$, data point size is 1 MB, and we take $b(n) = n^{0.6}$, then
  - full dataset has size 1 TB

Our Approach: BLB

Computational Issues

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

Our Approach: BLB

Computational Issues

- Recall: resources required to compute θ generally scale in number of *distinct* data points.
- Each BLB subsample/resample contains at most $b(n) < n$ distinct points.
- Example: if $n = 1,000,000$, data point size is 1 MB, and we take $b(n) = n^{0.6}$, then
  - full dataset has size 1 TB
  - subsamples/resamples contain at most 3,981 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 θ. Shares the bootstrap's consistency and higher-order correctness.
- Automatic: can compute without knowledge of the internals of θ

Our Approach: BLB

Like the Bootstrap

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

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

- Can explicitly control $b(n)$, the amount of data on which we must repeatedly compute θ; can have $b(n)/n \to 0$ as $n \to \infty$.

Our Approach: BLB

Like the Bootstrap

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

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

- Can explicitly control $b(n)$, the amount of data on which we must repeatedly compute θ; can have $b(n)/n \to 0$ as $n \to \infty$.
- More robust to choice of $b(n)$, which can be much smaller than n.

Our Approach: BLB

Like the Bootstrap

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

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

- Can explicitly control $b(n)$, the amount of data on which we must repeatedly compute θ; can have $b(n)/n \to 0$ as $n \to \infty$.
- More robust to choice of $b(n)$, which can be much smaller than n.
- Generally faster than the bootstrap (even if computing serially), and requires less total computation.

Our Approach: BLB

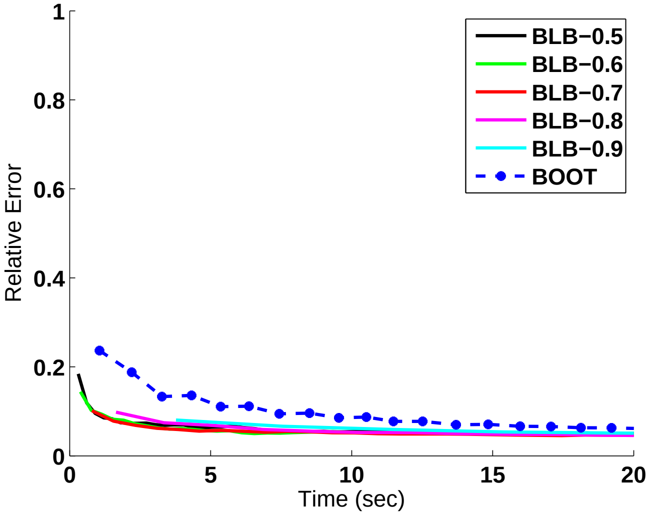
Like the Bootstrap

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

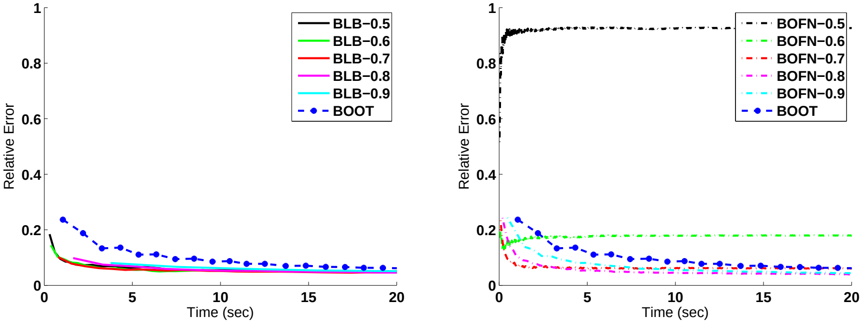
Beyond the Bootstrap (and *b* out of *n* Bootstrap/Subsampling)

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

Empirical Results: BLB



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