Divide and Conquer in Shape Restricted Problems and the Super-Efficiency Phenomenon\textsuperscript{1,2,3}

Moulinath Banerjee\textsuperscript{4}
University of Michigan, Ann Arbor

Edinburgh Workshop
Edinburgh, UK

July 5, 2018

\textsuperscript{1}Joint work with Cecile Durot (U. Paris Ouest) and Bodhisattva Sen (Columbia University)
\textsuperscript{3}Supported by NSF grant DMS-1712962
\textsuperscript{4}
Outline

1. Divide and Conquer: Introduction
2. The Pooled Versus the Global Estimator
3. Details
4. The Super-efficiency Phenomenon
5. Fixing the Super-efficiency Phenomenon
   - Asymptotic Properties
   - Computational Considerations
Divide and Conquer Principle

Sample splitting with large datasets

- **Split** dataset into *sub-datasets*, analyze each *separately*, *combine* the analysis.

- **Statistical parlance**: Compute *estimate* for each *subsample*, then combine estimates from subsamples, say by *averaging*.

Why Sample Splitting?

- Dataset may be too large for computational resources on 1 computer.

- Even if possible on one computer, computational efficiency may result if the underlying algorithm has high complexity. For example\(^a\):

  \[ N^2 = (m \times n)^2 \gg m \times n^2 + \text{complexity of averaging step}. \]

  \(^aN = \text{total sample size}, \ n = \text{size of each subsample}; \ m = \# \text{ subsamples} \)

- What we want: *Precision* of pooled estimator does *not suffer* too much in comparison to that of hard-to-compute *global estimator*. 

- **Statistical inference in massive data sets**, Runze Li and others (2010).
- **A Scalable Bootstrap for Massive Data**, Kleiner, Talwalkar, Sarkar and Jordan (2014).
Existing results typically show that the pooled estimator’s performance matches that of the global estimator in terms of the rate of convergence (risk bounds).

**Question:** Does it always work? In particular, can we do better than the global estimator in some sense?

**Question:** If so, do we pay a price?
Divide and Conquer: Introduction

The Pooled Versus the Global Estimator

Details

The Super-efficiency Phenomenon

Fixing the Super-efficiency Phenomenon
  • Asymptotic Properties
  • Computational Considerations
X_1, \ldots, X_N are i.i.d. random elements with common distribution P driven primarily by a monotone function f of interest

\( \theta \equiv \theta(P) \) is the finite dimensional parameter of interest. For this talk \( \theta(P) \) is either \( f(t_0) \) or \( f^{-1}(a_0) \).

Isotonic estimator \( \hat{\theta} \) (of \( \theta_0 \)) behaves like:

\[
\frac{r_N(\hat{\theta} - \theta_0)}{\sqrt{N}} \overset{d}{\rightarrow} G,
\]

where \( r_N \neq \sqrt{N} \), \( G \) has a scaled Chernoff’s distribution, with mean 0 and variance \( \sigma^2 > 0 \)

Typically \( \sigma^2 \) is difficult to estimate
Examples

\[ r_N(\hat{\theta} - \theta_0) \xrightarrow{d} G \]

where \( r_N \neq \sqrt{N} \), \( G \) is non-normal, has mean 0 and variance \( \sigma^2 > 0 \)

Examples

- Estimating a monotone regression function with additive errors \( Y = \mu(T) + \epsilon \) with \( \epsilon, T \) independent.
- Current Status Model (Case I interval censoring): observe whether a patient is infected or not when they are inspected. Response is \( \Delta = 1(T \leq U) \) where \( T \) is time to infection and \( U \) is inspection time.
- Estimating a monotone density based on i.i.d. observations. (Grenander estimator)
- Estimating a monotone hazard (failure rate).
Recombining by Averaging: Pooled estimator

Assume that $N$ is large and suppose that $N = n \times m$, where $n$ is still large and $m$ small/moderate (e.g., $n = 10K$, $m = 50$, so that $N = 500K$).

We define the *pooled estimator* $\bar{\theta}$ as follows:

1. Divide the set of samples $X_1, \ldots, X_N$ evenly and uniformly at random into $m$ *disjoint* subsets $S_1, \ldots, S_m$.

2. For each $i = 1, \ldots, m$, we compute the estimator $\hat{\theta}_i$ based on the data points in $S_i$.

3. Average together these estimators to obtain our final estimator

$$\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i.$$
Fixed $m$ and $n$ going to infinity

Recall: $\{\hat{\theta}_1, \ldots, \hat{\theta}_m\}$; $\bar{\theta} = \frac{1}{m} \sum_{i=1}^{m} \hat{\theta}_i$ and

$$r_n(\hat{\theta}_i - \theta_0) \xrightarrow{d} G,$$  

as $n \to \infty$.  

(1)

**Lemma**

Suppose (1) holds where $G$ has mean 0 and var. $\sigma^2 > 0$. For fixed $m$,

$$\sqrt{mr_n}(\bar{\theta} - \theta_0) \xrightarrow{d} H := m^{-1/2}(G_1 + G_2 + \ldots + G_m), \text{ as } n \to \infty,$$  

(2)

where $G_i$’s are i.i.d. $G$. Note that the limiting distribution $H$ has mean zero and variance $\sigma^2$.

- Compare this with $r_N(\hat{\theta} - \theta_0) \xrightarrow{d} G$ (global estimator)

- The asymptotic relative efficiency of $\bar{\theta}$ with respect to $\hat{\theta}$:

$$\frac{\sigma^2/r_N^2}{\sigma^2/(mr_n^2)} = \frac{m}{r_N^2} = \frac{mn^2}{m^{2\gamma}n^{2\gamma}} = m^{1-2\gamma} \text{ when } r_n = n^\gamma.$$
The asymptotic relative efficiency of \( \bar{\theta} \) with respect to \( \hat{\theta} \):

\[
\frac{\sigma^2 / r_n^2}{\sigma^2 / (m r_n^2)} = \frac{m r_n^2}{r_N^2} = \frac{m n^2 \gamma}{m^2 \gamma n^2 \gamma} = m^{1-2 \gamma} \text{ when } r_n = n^\gamma.
\]

The pooled estimate outperforms the global if and only if \( \gamma < 1/2 \): i.e., slower than the parametric rate.

For parametric problems, the ARE is 1. Why? Think of estimating the population mean via sample-splitting.

Pooled estimate underperforms when \( \gamma \) is larger than 1/2. For example, change-point problems.
Further Observations:

- As we have *m independent* replicates from the distribution of $\hat{\theta}_j$, $\sigma^2$ can be approximated by
  \[
  \hat{\sigma}^2 := \frac{r_n^2}{m - 1} \sum_{j=1}^{m} (\hat{\theta}_j - \bar{\theta})^2.
  \]

  [under a UI condition]

- Recall: For fixed $m$,
  \[
  \sqrt{mr_n}(\bar{\theta} - \theta_0) \overset{d}{\rightarrow} H := m^{-1/2}(G_1 + G_2 + \ldots + G_m), \quad \text{as } n \rightarrow \infty,
  \]
  where $G_i$’s are i.i.d. $G$.

- For moderately large $m$ (e.g., $m \geq 30$) $H$ in (2) maybe well approximated by $N(0, \sigma^2)$.

- Resulting CI for $\theta$ looks like:
  \[
  \left[ \bar{\theta} - \frac{\hat{\sigma}}{r_n \sqrt{m}} z_{\alpha/2}, \bar{\theta} + \frac{\hat{\sigma}}{r_n \sqrt{m}} z_{\alpha/2} \right]
  \]
  where $z_{\alpha}$ is the $(1 - \alpha)$-th quantile of the standard normal distribution.
Asymptotics when $m = m_n$ goes to $\infty$

**Theorem**

Let $\xi_{n,j} := r_n(\hat{\theta}_{n,j} - \theta_0)$, $j = 1, \ldots, m_n$; $\sigma^2_n := \text{Var}(\xi_{n,j})$. Suppose that $b_n := r_n(\mathbb{E} \hat{\theta}_{n,1} - \theta_0) = o(1)$ as $n \to \infty$. Also, suppose that the sequence $\{\xi^2_{n,1}\}$ is UI. Then, for any $m_n \to \infty$: if $\sqrt{m_n} b_n \to \tau \in \mathbb{R}$ (where $\tau = 0$ if $m_n \ll |b_n|^{-2}$), then

$$\sqrt{m_n} r_n(\bar{\theta}_{m_n,n} - \theta_0) \overset{d}{\to} N(\tau, \sigma^2).$$

So the two key challenges are:

(a) Establishing uniform integrability (UI) as desired above, and,

(b) Establishing an order for the bias $b_n$, since this gives us the right choice for $m_n$.

We get the maximal convergence rate when $m_n \sim b_n^{-2}$.
Isotonic regression

- \{(X_i, Y_i) : i = 1, \ldots, N\} i.i.d. data from the regression model
  \[ Y_i = \mu_0(X_i) + \epsilon_i \]
  where \( X_i \in [0, 1] \), \( X_i \) is independent of \( \epsilon_i \), \( E(\epsilon_i) = 0 \) and \( \mu \) is monotone increasing.

- **Direct estimation:** \( \theta_0 \equiv \mu_0(t_0) \), \( 0 < t_0 < 1 \) and \( \hat{\theta}_N = \hat{\mu}_N(t_0) \), where \( \hat{\mu}_N \) is isotonic regression estimator defined as the minimizer of
  \[ g \mapsto \sum_{i=1}^{N} (Y_i - g(X_i))^2 \]
  over the set of all nondecreasing functions.

- **Inverse estimation:** \( \theta_0 \equiv \mu_0^{-1}(a_0) \) and \( \hat{\theta}_N = \hat{\mu}_N^{-1}(a_0) \) with generalized inverses.
Limiting distribution in Isotonic Regression

Assume that the errors are independent of the regressors with positive variance $\tau^2$, $X$ has a density $f_X$, and $\mu'(t_0) > 0$.

- We have
  \[ N^{1/3}(\hat{\theta}_N - \theta_0) \rightarrow_d C Z, \]
  where $Z \sim$ Chernoff's distribution, is symmetric with $E(Z) = 0$, $\text{Var}(Z) = .26$ (approximate); and $\sigma^2 = C^2 \times .26$.
  For direct estimation, $C = (4 \tau^2 \mu'_0(t_0)/f_X(t_0))^{1/3}$.

- With $r_n = n^{1/3}$ and $m = m_n \rightarrow \infty$, if the assumptions of the general theorem hold (uniform integrability and control of $m_n$ by the bias), then
  \[ \sqrt{m_n}r_n(\bar{\theta}_{n,m} - \theta_0) \overset{d}{\rightarrow} N(\tau, \sigma^2), \text{ as } n \rightarrow \infty. \]
  and $\bar{\theta}_{n,m}$ outperforms $\hat{\theta}_N$. 
Checking the Assumptions (1)

To apply the general Theorem, need:

(a) the uniform integrability of \( \left\{ n^{2/3}(\hat{\theta}_n - \theta_0)^2 \right\}_{n \geq 1} \)

(b) analytical expression for the bias \( b_n \equiv n^{1/3}(E(\hat{\theta}_n) - \theta_0) \).

Theorem (Uniform integrability)

Assume in addition

- \( \mu_0 \) is differentiable on [0, 1],
- \( \mu'_0 \) and \( f_X \) are bounded away from zero and infinity,
- there exists \( \alpha > 0 \) such that for all \( \theta \in \mathbb{R} \),

\[
\mathbb{E}(e^{\theta \epsilon}) \leq \exp(\theta^2 \alpha).
\]

Then for any \( p \geq 1 \), \( \mathbb{E} \left( n^{p/3}|\hat{\theta}_n - \theta_0|^p \right) = O(1). \)
Theorem (Bias)

Assume in addition that the derivative \( \mu'_0 \) has the following Hölder smoothness property: there exist \( C > 0 \) and \( s > 3/4 \) such that

\[
|\mu'_0(u) - \mu'_0(v)| \leq C|u - v|^s \quad \text{for all } u, v \in [0, 1].
\]

Then for inverse estimation we have \( b_n = o(n^{-1/6}) \) and for direct estimation, if \( s = 1 \), then with \( \zeta > 0 \) arbitrary, \( b_n = O(n^{-2/15+\zeta}) \).

In both cases, the pooled-by-averaging estimator with \( m_n \sim b_n^{-2} \) outperforms the global estimator and converges to a Gaussian law.
The Inverse Problem

- For the *inverse function* estimation problem: \( \theta_0 = \mu_0^{-1}(a) \) and \( \hat{\theta}_N = \hat{\mu}_N^{-1}(a) \), we also have:

  \[
  N^{1/3} \left( \hat{\theta}_N - \theta_0 \right) \overset{d}{\rightarrow} \text{constant} \times \mathbb{Z}.
  \]

- The UI condition is satisfied as in the ‘forward’ problem.

- Our calculations yield a *lower bias* under similar assumptions: namely,

  \[
  b_n = \mathbb{E}[n^{1/3}(\hat{\mu}_n^{-1}(a) - \mu_0^{-1}(a))] = o(n^{-1/6}).
  \]

- Choosing \( m_n = n^{2\phi} \) with \( \phi = 1/6 \) (i.e., \( m_n = O(c_n^2) \)) and noting that \( \sqrt{m_n} b_n \to 0 \), we conclude

  \[
  N^{3/8} (\bar{\theta}_{m_n} - \theta_0) \overset{d}{\rightarrow} N(0, \text{Variance}).
  \]
Divide and Conquer: Introduction

The Pooled Versus the Global Estimator

Details

The Super-efficiency Phenomenon

Fixing the Super-efficiency Phenomenon
  - Asymptotic Properties
  - Computational Considerations
**Variance reduction** accomplished by sample-splitting for estimating a fixed monotone function at a given point comes at a *price*.

A larger number of splits ($m$) brings about greater reduction in the variance for a fixed function.

But the performance of the pooled estimator in a uniform sense, over an appropriately large class of functions, *deteriorates* in comparison to the global estimator with increasing $m$.

A super-efficiency phenomenon: a trade-off between *point-wise* performance and performance in a *uniform sense*. 
Fix a continuous monotone (non-increasing) function $\mu_0$ on $[0, 1]$ that is continuously differentiable on $[0, 1]$ with $0 < c < |\mu_0'(t)| < d < \infty$ for all $t \in [0, 1]$. Let $t_0 \in (0, 1)$.

Define a neighborhood $\mathcal{M}_0$ of $\mu_0$ as the class of all continuous non-increasing functions on $[0, 1]$ that are continuously differentiable on $[0, 1]$, that coincide with $\mu_0$ outside of $(t_0 - \epsilon_0, t_0 + \epsilon_0)$ for some (small) $\epsilon_0 > 0$ and such that $0 < c < |\mu'(t)| < d < \infty$ for all $t \in [0, 1]$.

Consider $N$ i.i.d. observations $\{Y_i, T_i\}_{i=1}^n$ from the model:

$$Y = \mu_0(T) + \epsilon,$$

where $T \sim \text{Uniform}(0, 1)$ is independent of $\epsilon \sim N(0, \tau^2)$. 
Recall

- We have adequate uniform integrability, and can show that:
  \[ \mathbb{E}_{\mu_0} \left[ N^{2/3} ( (\hat{\mu}_N(t_0) - \mu_0(t_0))^2 \right] \to \text{Var}(G), \quad \text{as } N \to \infty, \]

- While,
  \[ \mathbb{E}_{\mu_0} \left[ N^{2/3} ( \bar{\mu}_N(t_0) - \mu_0(t_0))^2 \right] \to m^{-1/3} \text{Var}(G), \quad \text{as } N \to \infty, \]
  noting that \( G \) and \( (G_1 + G_2 + \ldots + G_m)/\sqrt{m} \) have the same variance.

- Hence, for estimating \( \mu_0 \) at the point \( t_0 \), the pooled estimator \textit{outperforms} the isotonic regression estimator.
We now compare the performance of the estimators over the class $\mathcal{M}_0$.

**Theorem**

Let

$$E := \limsup_{N \to \infty} \sup_{\mu \in \mathcal{M}_0} \mathbb{E}_\mu \left[ N^{2/3} (\hat{\mu}_N(t_0) - \mu(t_0))^2 \right], \quad \text{and}$$

$$E_m := \liminf_{N \to \infty} \sup_{\mu \in \mathcal{M}_0} \mathbb{E}_\mu \left[ N^{2/3} (\overline{\mu}_N(t_0) - \mu(t_0))^2 \right].$$

Then,

$$E < \infty, \quad \text{while} \quad E_m \geq m^{2/3} c_0 \quad \text{for some } c_0 > 0.$$

- In the case that $m = m_n \to \infty$,

  $$\liminf_{N \to \infty} \sup_{\mu \in \mathcal{M}_0} \mathbb{E}_\mu \left[ N^{2/3} (\overline{\mu}_N(t_0) - \mu(t_0))^2 \right] = \infty.$$

- Thus, the *better off* we are in a *point-wise* sense with the pooled estimator, the *worse-off* we are in the *uniform sense* over $\mathcal{M}_0$.
Consider a fixed model against a sequence of models varying with $n$.

**Null model:** $Y = T + \epsilon$, $T \sim \text{Unif}(0, 1)$, $\epsilon \sim \mathcal{N}(0, 0.2^2)$

**Alternative (varying with $n$) models:**

$$Y = T + n^{-1/3} B(n^{1/3}(T - t_0)) + \epsilon, \quad t_0 = 0.5$$

other parameters remain the same and

$$B(u) = 2^{-1}(1 - (|u| - 1)^2)^2 1_{\{|u| \leq 2\}}.$$
Null model: \( Y = T + \epsilon, \ T \sim \text{Unif}(0, 1), \ \epsilon \sim N(0, 0.2^2). \)

<table>
<thead>
<tr>
<th>(( n, m ))</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>45</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.64</td>
<td>2.01</td>
<td>2.49</td>
<td>2.55</td>
<td>2.66</td>
<td>2.44</td>
</tr>
<tr>
<td>200</td>
<td>1.49</td>
<td>2.21</td>
<td>2.83</td>
<td>3.47</td>
<td>2.87</td>
<td>3.02</td>
</tr>
<tr>
<td>500</td>
<td>1.57</td>
<td>2.34</td>
<td>2.92</td>
<td>3.61</td>
<td>3.68</td>
<td>3.88</td>
</tr>
<tr>
<td>1000</td>
<td>1.57</td>
<td>2.22</td>
<td>2.99</td>
<td>3.18</td>
<td>4.09</td>
<td>4.18</td>
</tr>
<tr>
<td>3000</td>
<td>1.77</td>
<td>2.50</td>
<td>3.20</td>
<td>3.66</td>
<td>3.80</td>
<td>4.53</td>
</tr>
<tr>
<td>10000</td>
<td>1.59</td>
<td>2.63</td>
<td>3.05</td>
<td>3.67</td>
<td>3.74</td>
<td>4.25</td>
</tr>
</tbody>
</table>

Alternative models: \( Y = T + n^{-1/3} B(n^{1/3}(T - t_0)) + \epsilon, \ t_0 = 0.5. \)

<table>
<thead>
<tr>
<th>(( n, m ))</th>
<th>5</th>
<th>15</th>
<th>30</th>
<th>45</th>
<th>60</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1.21</td>
<td>1.22</td>
<td>1.26</td>
<td>1.20</td>
<td>1.11</td>
<td>1.04</td>
</tr>
<tr>
<td>200</td>
<td>1.17</td>
<td>1.30</td>
<td>1.11</td>
<td>1.09</td>
<td>1.05</td>
<td>0.94</td>
</tr>
<tr>
<td>500</td>
<td>1.15</td>
<td>1.18</td>
<td>1.12</td>
<td>1.01</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>1000</td>
<td>1.14</td>
<td>1.17</td>
<td>1.07</td>
<td>1.03</td>
<td>0.93</td>
<td>0.84</td>
</tr>
<tr>
<td>3000</td>
<td>1.14</td>
<td>1.15</td>
<td>1.02</td>
<td>1.00</td>
<td>0.95</td>
<td>0.85</td>
</tr>
<tr>
<td>10000</td>
<td>1.17</td>
<td>1.08</td>
<td>1.05</td>
<td>0.98</td>
<td>0.91</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Table: The ratio of MSEs \( \frac{E[(\hat{\mu}_N(t_0) - \mu(t_0))^2]}{E[(\mu(t_0) - \mu(t_0))^2]} \) of global to pooled increases with \( m \) for the fixed model but starts falling under the alternative sequence.
Super-Efficiency (in the inverse problem): Demonstration via Simulation

- The same setting as in the forward problem.
- **Null model:** \( Y = x + \epsilon, \ X \sim \text{Unif}(0, 1), \ \epsilon \sim \mathcal{N}(0, 0.2^2). \)
- **Alternative (varying with \( n \)) models:**
  \( Y = x + n^{-1/3} B(n^{1/3}(x - x_0)) + \epsilon, \) other parameters remain the same and
  \[
  B(u) = 2^{-1}(1 - (|u| - 1)^2)^2 1_{\{|u| \leq 2\}}.
  \]
- In the next slide, we present ratios of the (estimated) mean squared errors
  \[
  \frac{\mathbb{E}[(\hat{\mu}_N^{-1}(0.5) - \mu^{-1}(0.5))^2]}{\mathbb{E}[(\bar{\theta}_m(x_0) - \mu^{-1}(0.5))^2]} \]
  comparing the performance of the pooled estimator \( \bar{\theta}_m \) with the global estimator \( \hat{\mu}_N^{-1}(0.5) \) as \( n \) and \( m \) change for these models.
### Table: The same phenomenon as in the forward problem.
More generally, this phenomenon can be established a variety of monotone function problems, e.g. regression models under more general assumptions on errors, current status/case-1 interval censoring model, estimation of a monotone density (Grenander estimator), estimation of a monotone hazard rate; with the same recurring convergence rates.

The differentiability assumption on the regression function is critical. Without differentiability, the pooled estimator can fail even for a fixed model (fixed $\mu_0$).

Can we sharpen the bias calculations?
**Goal:** construct an estimator in the isotonic regression problem that does not suffer from super-efficiency, and has the same limiting behaviour as the global estimator.

**Broad regression setting with heterogeneity in data:** The pairs \( \{(X_i, Y_i)\}_{i=1}^N \), where \( X_i \in [0, 1] \), are independent and come from \( m \) different sub-populations with the pairs in each sub-population being i.i.d. The sub-populations are linked by the common mean function \( \mu_0 \) of interest: \( E(Y_i|X_i) = \mu_0(X_i) \) for all \( i \), for an increasing function \( \mu_0 \).

**D&C setting:** The \( N \) pairs are distributed arbitrarily across \( L \) servers. The number \( L \) of different servers can grow as \( N \to \infty \).

**Parameters of interest:** \( \theta_0 \equiv \mu_0(t_0) \) where \( t_0 \in [\delta, 1 - \delta] \), for some \( \delta > 0 \), or \( \theta_0 \equiv \mu_0^{-1}(a_0) \) for \( a_0 \in (\mu_0(0), \mu_0(1)) \).
The Estimator

We define the Smooth-and-Isotonize estimator as follows:

1. Let $K \in \mathbb{N}$ and $I_k = ((k - 1)/K, k/K]$ for all $k \in \{1, \ldots, K\}$.
2. On each server $\ell = 1, \ldots, L$, for all $k = 1, \ldots, K$, compute $T_{\ell k}$, the sum of all $Y_i$ on server $\ell$ with corresponding $X_i \in I_k$, $C_{\ell k}$, the number of observations on server $\ell$ with $X_i \in I_k$.
3. Transfer these statistics to a central server.
4. For each $k \in \{1, \ldots, K\}$, compute

$$\bar{y}_k = \frac{1}{\sum_{\ell=1}^L C_{\ell k}} \sum_{\ell=1}^L T_{\ell k}.$$ 

Our final estimator of $(\mu_0(1/K), \ldots, \mu_0(K/K))^T$, is

$$\hat{y} = \arg\min_{h \in \mathbb{R}^K: h_1 \leq \cdots \leq h_K} \sum_{k=1}^K w_k (\bar{y}_k - h_k)^2,$$

where $w_k = N^{-1} \sum_{\ell=1}^L C_{\ell k}$. 
The final estimator of $\mu_0$ on $[0, 1]$ is obtained by piecewise-constant interpolation in between the points $1/K, 2/K, \ldots, K/K$.

The final estimator of $\mu_0^{-1}$ is defined as generalized inverse.

We denote by $\hat{\theta}_N$ the estimator of the parameter of interest $\theta_0$. 
Assumptions

Let $F_X := \sum_{j=1}^{m} \frac{n_j}{N} F_{Xj}$ be the mixing distribution function, with $n_j$ the number of observations from the $j$-th sub-population and $F_{Xj}$ the common distribution function.

Assume:

1. The density $f_X$ of $F_X$ is bounded from above and away from zero by positive numbers $C_1$, $C_2$ independent of $N$.

2. There exists $\sigma > 0$ such that $\mathbb{E}[(Y_i - \mu(X_i))^2|X_i] \leq \sigma^2$ for all $i$, with probability one.

3. There exist positive numbers $C_3$ and $C_4$ such that

   $$C_3 < \left| \frac{\mu_0(t) - \mu_0(x)}{t - x} \right| < C_4 \text{ for all } t \neq x \in [0, 1],$$

4. $K^{-1} = o(N^{-1/3})$ and there exists $\lambda \in (0, 1]$ such that

   $$\min_{1 \leq j \leq m} \frac{n_j}{N} \geq \lambda > 0 \text{ and } \liminf_{N \to \infty} N^{1/3} \lambda (\log N)^{-3} = \infty.$$
Remarks on the assumptions

1. The estimator and assumptions do not depend on the way the observations were stored across different servers,

2. assuming $K^{-1} = o(N^{-1/3})$ ensures that the isotonic algorithm operating on these averages at the central machine can still recover the $N^{-1/3}$ convergence rate,

3. Since

$$1 = \sum_{j=1}^{m} \frac{n_j}{N} \geq m \min_{1 \leq j \leq m} \frac{n_j}{N},$$

the conditions imply that the number $m$ of different sub-populations cannot grow too fast: we must have $m \ll N^{1/3}(\log N)^{-3}$. 
Let $\mathcal{F}_1$ be the class of non-decreasing functions $\mu$ on $[0, 1]$ that satisfy

$$C_3 < \left| \frac{\mu(t) - \mu(x)}{t - x} \right| < C_4 \quad \text{for all } t \neq x \in [0, 1],$$

and $\sup_t |\mu(t)| \leq C_5$, where $C_5 > 0$ is a positive number.

**Theorem (Direct and inverse estimation problems)**

There exists $C > 0$ that depends only on $\sigma^2$, $C_1$, $C_2$, $C_3$, $C_4$, $C_5$, $\delta$ such that for all $a \in \mathbb{R}$,

$$\limsup_{N \to \infty} \sup_{\mu \in \mathcal{F}_1} N^{2/3} \mathbb{E}_\mu (\hat{\theta}_N - \theta_0(\mu))^2 \leq C.$$
We next make the following further technical assumptions.

\( \tilde{A}_0 \). The densities \( \{f_j\} \) are uniformly bounded in \( j \) on \([0, 1]\).

\( \tilde{A}_1 \). With \( \sigma_j^2(u) = \mathbb{E}[(Y - \mu_0(X))^2 | X = u] \) in the \( j \)-th sub-population, as \( \delta \to 0 \) we have

\[
\sup_j \max\{ \sup_{|u-v| \leq \delta} |\sigma_j^2(u) - \sigma_j^2(v)|, \sup_{|u-v| \leq \delta} |f_j(u) - f_j(v)| \} \to 0.
\]

\( \tilde{A}_2 \). \( f_X \) converges pointwise on \([0, 1]\) as \( N \to \infty \) to a continuous function \( f_{\infty} \) that is bounded away from zero.

\( \tilde{A}_3 \). With \( \sigma_X^2(u) := \sum_{j=1}^m \frac{n_j}{N} \sigma_j^2(u)f_j(u) \), \( \sigma_X^2 \) converges pointwise on \([0, 1]\) to a continuous function \( \sigma_{\infty}^2 \), bounded away from 0, as \( N \to \infty \).

\( \tilde{A}_4 \). There exist \( \sigma > 0 \) and \( p > 2 \) such that for all \( t \) and all sub-populations, \( \mathbb{E}[|Y - \mu_0(X)|^p | X = t] \leq \sigma^p \).

\( \tilde{A}_5 \). \( \mu_0 \) is differentiable on \([0, 1]\) with \( \inf_{u \in [0,1]} |\mu'_0(u)| > 0 \).
Limiting distribution

Theorem (Direct and inverse estimation problems)

We have

$$N^{1/3}(\hat{\theta}_N - \theta_0) \rightarrow_d C \mathbb{Z} \text{ as } N \rightarrow \infty,$$

where $\mathbb{Z}$ has the Chernoff's distribution.

- In the inverse problem, $C = \left( \frac{2\sigma_\infty(t)}{|\mu'_0(t)|f_\infty(t)} \right)^{2/3}$.
- In the direct problem, $C = \left( \frac{4\sigma_\infty^2(t)|\mu'_0(t)|}{f_\infty^2(t)} \right)^{1/3}$.
- In both problems, the limiting behavior is the same as that of the global estimator.
Computational Considerations

To estimate $\theta_0$ requires
- Global: $O(N \log N)$ elementary computations.
- Pooled-by-Averaging with $m$ splits:
  - $O(N \log N)$ elementary computations,
  - $O(m)$ transfers of numbers.
- Smooth and Isotonize with $K \sim N^\zeta$ for some $1/3 < \zeta < 1$:
  - $O(N \log N \lor LK)$ elementary computations,
  - $O(LK)$ transfers.

More general cube-root problems: Shi, Lu and Song (to appear in JASA) considered the pooled-by-averaging estimator in general cube root problems of the Kim and Pollard (1990) type. We believe that similar strategies to the one presented here can be used to fix the super-efficiency problem that also arises in their setting.