Divide and Conquer in Shape Restricted Problems and the Super-Efficiency Phenomenon[1,2,3]

Moulinath Banerjee[4]
University of Michigan, Ann Arbor

Edinburgh Workshop
Edinburgh, UK

July 5, 2018

[1] Joint work with Cecile Durot (U. Paris Ouest) and Bodhisattva Sen (Columbia University)

[2] Available at http://arxiv.org/abs/1605.04446, http://arxiv.org/abs/1806.08542

[3] Supported by NSF grant DMS-1712962

[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]:

$$N^2 = (m \times n)^2 >> m \times n^2 + \text{complexity of averaging step.}$$

[a] N = total sample size, n = size of each subsample; m = # subsamples

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

Divide and Conquer (some references)

- **Statistical inference in massive data sets**, Runze Li and others (2010).
- **Divide and Conquer Kernel Ridge Regression**, Zhang, Duchi, Wainwright (2013).
- **A Divide-and-Conquer Solver for Kernel Support Vector Machines,** Hsieh, Si and Dhillon (2014).
- **A Partially Linear Framework for Massive Heterogeneous Data,** Zhao, Cheng and Liu (2014).
- **A Scalable Bootstrap for Massive Data,** Kleiner, Talwalkar, Sarkar and Jordan (2014).
- **A Massive Data Framework for M-Estimators with Cubic-Rate,** Shi, Lu and Song (2016).

Divide and Conquer

- 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?

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

Basic Framework

- $X_1, \dots, X_N$ 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 θ_0) behaves like:

$$r_N(\hat{\theta} - \theta_0) \overset{d}{\to} G,$$

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

- Typically σ^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 ϵ, 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

$$r_n(\hat{\theta}_i - \theta_0) \xrightarrow{d} G, \qquad \text{as } n \to \infty. \qquad (1)$$

Lemma

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

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

where G_i's are i.i.d. G. Note that the limiting distribution H has mean zero and variance σ^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 $\hat{\theta}$:

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

Improving Efficiency and the Convergence Rate

- 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 γ is *larger* than $1/2$. For example, change-point problems.

Further Observations:

- As we have m independent replicates from the distribution of $\hat{\theta}_j$, σ^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*,
$\sqrt{m} r_n(\bar{\theta} - \theta_0) \xrightarrow{d} H := m^{-1/2}(G_1 + G_2 + \ldots + G_m), \quad$ as $n \to \infty$,
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 θ 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_α is the $(1-\alpha)$-th quantile of the standard normal distribution.

Asymptotics when $m = m_n$ goes to ∞

Theorem

Let $\xi_{n,j} := r_n(\hat{\theta}_{n,j} - \theta_0), \qquad j = 1, \ldots, m_n; \quad \sigma_n^2 := Var(\xi_{n,j})$. *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* $\{\xi_{n,1}^2\}$ *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*

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

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 ϵ_i, $E(\epsilon_i) = 0$ and μ 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 τ^2, X has a density f_X, and $\mu'(t_0) > 0$.

- We have
$$N^{1/3}(\hat{\theta}_N - \theta_0) \to_d C\,\mathbb{Z}\,,$$
where $\mathbb{Z} \sim$ Chernoff's distribution, is symmetric with $E(\mathbb{Z}) = 0$, $\mathrm{Var}(\mathbb{Z}) = .26$ (appx); 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 $m = m_n \to \infty$, if the assumptions of the general theorem hold (uniform integrability and control of m_n by the bias), then
$$\sqrt{m_n} r_n(\bar{\theta}_{n,m} - \theta_0) \xrightarrow{d} N(\tau, \sigma^2), \text{ as } n \to \infty.$$
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

- μ_0 *is differentiable on* $[0, 1]$,
- μ_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 $p \geq 1$, $\mathbb{E}\left(n^{p/3}|\hat{\theta}_n - \theta_0|^p\right) = O(1)$.

Checking the Assumptions (2)

Theorem (Bias)

Assume in addition that the derivative μ_0' has the following Hölder smoothness property: there exist $C > 0$ and $s > 3/4$ such that

$$|\mu_0'(u) - \mu_0'(v)| \leq C|u - v|^s \text{ for all } u, v \in [0, 1].$$

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:

$$N^{1/3}\,(\hat{\theta}_N - \theta_0) \overset{d}{\to} \text{constant} \times \mathbb{Z}\,.$$

- 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

$$N^{3/8}\,(\overline{\theta}_{m_n} - \theta_0) \overset{d}{\to} N(0, \text{Variance})\,.$$

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

Sample Splitting and Super-Efficiency

- *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*.

Super-efficiency

- Fix a continuous monotone (non-increasing) function μ_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 μ_0 as the class of all continuous non-increasing functions on $[0, 1]$ that are continuously differentiable on $[0, 1]$, that coincide with μ_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:

$$\mathbb{E}_{\mu_0}\left[N^{2/3}((\hat{\mu}_N(t_0)-\mu_0(t_0))^2\right]\to \mathrm{Var}(G), \qquad \text{as } N\to\infty,$$

- While,

$$\mathbb{E}_{\mu_0}\left[N^{2/3}(\overline{\mu}_N(t_0)-\mu_0(t_0))^2\right]\to m^{-1/3}\,\mathrm{Var}(G), \qquad \text{as } N\to\infty,$$

  noting that G and $(G_1+G_2+\ldots+G_m)/\sqrt{m}$ have the same variance.

- Hence, for estimating μ_0 at the point t_0, the pooled estimator *outperforms* the isotonic regression estimator.

Super-efficiency

We now compare the performance of the estimators over the class $\mathcal{M}_0$.

Theorem

Let

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

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

Then,

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

- 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 $\mathcal{M}_0$.

Super-efficiency: Demonstration via Simulation

- Consider a fixed model against a sequence of models varying with n.
- **Null model:** $Y = T + \epsilon$, $T \sim \text{Unif}(0,1)$, $\epsilon \sim N(0, 0.2^2)$
- **Alternative (varying with n) models:**

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

other parameters remain the same and

$$B(u) = 2^{-1}(1 - (|u| - 1)^2)^2 \mathbf{1}_{\{|u| \le 2\}}\,.$$

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

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

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

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

Table: The ratio of MSEs $\frac{\mathbb{E}\left[(\hat{\mu}_N(t_0) - \mu(t_0))^2\right]}{\mathbb{E}\left[(\overline{\mu}_N(t_0) - \mu(t_0))^2\right]}$ 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:** $Y = x + \epsilon$, $X \sim \text{Unif}(0,1)$, $\epsilon \sim N(0, 0.2^2)$.
- **Alternative (varying with n) models:** $Y = x + n^{-1/3} B(n^{1/3}(x - x_0)) + \epsilon$, other parameters remain the same and
$$B(u) = 2^{-1}(1 - (|u| - 1)^2)^2 \mathbf{1}_{\{|u| \le 2\}} \,.$$
- In the next slide, we present ratios of the (estimated) mean squared errors $\frac{\mathbb{E}\left[(\hat{\mu}_N^{-1}(0.5) - \mu^{-1}(0.5))^2\right]}{\mathbb{E}\left[(\overline{\theta}_m(x_0) - \mu^{-1}(0.5))^2\right]}$ comparing the performance of the pooled estimator $\overline{\theta}_m$ with the global estimator $\hat{\mu}_N^{-1}(0.5)$ as n and m change for these models.

Super-Efficiency: Demonstration via Simulation

(n, m)	5	10	15	30	45	60	90
50	1.67	1.71	1.90	1.66	1.57	1.65	1.17
100	1.31	1.76	2.21	2.29	2.16	2.46	2.33
200	1.75	2.06	2.42	2.81	2.58	3.16	3.39
500	1.70	2.13	2.12	2.80	3.16	3.59	4.11
1000	1.46	2.04	2.46	2.88	3.60	3.51	4.31
3000	1.63	2.12	2.33	3.11	4.15	3.84	3.69
10000	1.75	2.11	2.70	2.86	3.31	5.08	5.18

5	10	15	30	45	60	90	
50	1.47	1.21	0.94	0.70	0.55	0.54	0.39
100	1.04	0.97	0.90	0.59	0.47	0.40	0.31
200	1.03	0.94	0.76	0.68	0.42	0.38	0.29
500	1.01	0.90	0.69	0.54	0.44	0.34	0.24
1000	1.16	0.88	0.66	0.52	0.36	0.34	0.24
3000	1.09	0.87	0.75	0.43	0.40	0.31	0.21
10000	0.94	0.79	0.80	0.43	0.33	0.31	0.23

Table: The same phenomenon as in the forward problem.

Related Problems

- 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 μ_0).
- Can we sharpen the bias calculations?

Framework

- 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 μ_0 of interest: $E(Y_i|X_i) = \mu_0(X_i)$ for all i, for an increasing function μ_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 ℓ with corresponding $X_i \in I_k$, $C_{\ell k}$, the number of observations on server ℓ with $X_i \in I_k$,
3. Transfer these statistics to a central server.
4. For each $k \in \{1, \ldots, K\}$, compute

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

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

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

where $w_k = N^{-1} \sum_{\ell=1}^{L} C_{\ell k}$.

The Estimator

- The final estimator of μ_0 on $[0, 1]$ is obtained by piecewise-constant interpolation in between the points $1/K,\ 2/K, \dots, K/K$.
- The final estimator of μ_0^{-1} is defined as generalized inverse.
- We denote by $\hat{\theta}_N$ the estimator of the parameter of interest θ_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 probablity 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 to fast: we must have $m \ll N^{1/3}(\log N)^{-3}$.

Uniform Bounds

Let $\mathcal{F}_1$ be the class of non-decreasing functions μ on $[0,1]$ that satisfy

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

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 $a \in \mathbb{R}$,

$$\limsup_{N\to\infty} \sup_{\mu\in\mathcal{F}_1} N^{2/3}\mathbb{E}_\mu(\hat{\theta}_N - \theta_0(\mu))^2 \leq C.$$

Limiting Distribution

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

$$\sup_{j \geq 1} \max\{ \sup_{|u-v| \leq \delta} |\sigma_j^2(u) - \sigma_j^2(v)|, \sup_{|u-v| \leq \delta} |f_j(u) - f_j(v)|\} \to 0.$$

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

$\tilde{A}_3$. With $\sigma_X^2(u) := \sum_{j=1}^m \frac{n_j}{N} \sigma_j^2(u) f_j(u)$, σ_X^2 converges pointwise on $[0, 1]$ to a continuous function σ_∞^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, $\mathbb{E}[|Y - \mu_0(X)|^p | X = t] \leq \sigma^p$.

$\tilde{A}_5$. μ_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 θ_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 $1/3 < \zeta < 1$:
  - $O(N \log N \vee LK)$ 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.