A Massive Data Framework for M-Estimators with Cubic-Rate

Chengchun Shi, Wenbin Lu and Rui Song*

Abstract

The divide and conquer method is a common strategy for handling massive data. In this article, we study the divide and conquer method for cubic-rate estimators under the massive data framework. We develop a general theory for establishing the asymptotic distribution of the aggregated M-estimators using a weighted average with weights depending on the subgroup sample sizes. Under certain condition on the growing rate of the number of subgroups, the resulting aggregated estimators are shown to have faster convergence rate and asymptotic normal distribution, which are more tractable in both computation and inference than the original M-estimators based on pooled data. Our theory applies to a wide class of M-estimators with cube root convergence rate, including the location estimator, maximum score estimator and value search estimator. Empirical performance via simulations and a real data application also validate our theoretical findings.

Keywords: Cubic rate asymptotics; divide and conquer; M-estimators; massive data.

*Chengchun Shi is graduate student (E-mail: cshi4@ncsu.edu), Wenbin Lu is Professor (E-mail: lu@stat.ncsu.edu), and Rui Song is Associate Professor (rsong@ncsu.edu), Department of Statistics, North Carolina State University, Raleigh, NC 27695. The authors would like to thank an associate editor and three referees for their thoughtful and constructive comments, which help to improve an earlier version of the paper. This work was partly supported by a NIH grant P01 CA142538.

1 Introduction

In a world of explosively large data, effective estimation procedures are needed to deal with the computational challenge arisen from analysis of massive data. The divide and conquer method is a commonly used approach for handling massive data, which divides data into several groups and aggregate all subgroup estimators by a simple average to lessen the computational burden. A number of problems have been studied for the divide and conquer method, including variable selection (Chen and Xie, 2014), nonparametric regression (Zhang et al., 2013; Zhao et al., 2016) and bootstrap inference (Kleiner et al., 2014), to mention a few. Most papers establish that the aggregated estimators achieve the oracle result, in the sense that they possess the same nonasymptotic error bounds or limiting distributions as the pooled estimators, which are obtained by fitting all the data in a single model. This implies that the divide and conquer scheme can not only maintain efficiency, but also obtain a feasible solution for analyzing massive data.

In addition to the computational advantages for handling massive data, the divide and conquer method, somewhat surprisingly, can lead to aggregated estimators with improved efficiency over pooled estimators with slower than the usual $n^{1/2}$ convergence rate. A recent independent work of Banerjee et al. (2016) studied the divide and conquer principle in the monotone regression setting where the estimator converges at $n^{1/3}$ rate. In particular, they showed the aggregated estimator obtained by averaging all subgroup estimators converges much faster than the pooled estimator based on all observations and is asymptotically normal. This phenomenon is expected to hold under many other cube-root estimation problems. For example, Chernoff (1964) studied a cubic-rate estimator for estimating the mode. It was shown therein that the estimator converges in distribution to the argmax of a Brownian motion minus a quadratic drift. Kim and Pollard (1990) systematically studied a class of cubic-rate M-estimators and established their limiting distributions as the argmax of a general Gaussian process minus a quadratic form. These results were extended to a more general class of M-estimators using modern empirical process results (van der Vaart and Wellner, 1996; Kosorok, 2008). In this paper, we study a class of M-estimators with cubic-rate and develop a general inference framework for the aggregated estimators obtained by the divide and conquer method. Our theory states that the aggregated estimators can achieve a faster convergence rate than the pooled estimators and have asymptotic normal distributions when the number of groups diverges at a proper rate as the sample size of each group grows. This enables a simple way for estimating the covariance matrix of the aggregated estimators.

When establishing the asymptotic properties of the aggregated estimators, a major technical challenge is to quantify the accumulated bias. Different from estimators with standard $n^{1/2}$ convergence rate, M-estimators with $n^{1/3}$ convergence rate generally do not have a nice linearization representation and the magnitude of the associated biases is difficult to quantify. One way to obtain the magnitude of the bias is by establishing a coupling inequality for the cubic-rate estimator. For example, Banerjee et al. (2016) derived a nonasymptotic bound for the biases of the isotonic estimator in a monotone regression model and its inverse, based on the coupling inequality of the isotonic estimator (see Lemma 8.10 in that paper, and also Equation (29) in Durot (2002)). Groeneboom et al. (1999) provided a coupling inequality for the inverse process of the Grenander estimator. Their results can be used to establish the bias of the Grenander estimator. While such strategy is useful for studying the bias of some one-dimensional cubic-rate estimators, it is not suitable for multi-dimensional estimators. On one hand, these coupling inequalities are all based on Komos-Major-Tusnady (KMT) approximation (Komlós et al., 1975) and its extensions (cf. Csörgő et al., 1985; Sakhanenko, 2006) that only apply to the empirical distribution or the quantile process. There are extensions of the KMT approximation for more general empirical process (cf. Rio, 1994; Koltchinskii, 1994). However, the rate of the approximation will depend on the dimension of the parameter and decays fast as the dimension increases. On the other hand, proofs of these coupling inequalities all rely on the properties of the argmax of a Brownian motion process with a parabolic drift (cf. Proposition 1 in Durot (2002) and the discussions therein), and are not applicable to cubic-rate estimators that converge to the argmax of a more general Gaussian process minus a quadratic term. Here, we propose a novel approach to derive an upper bound for the bias, without establishing the coupling inequalities. To the best of our knowledge, this is the first time that a nonasymptotic error bound for the bias of a general cubic-rate estimator is provided.

A key innovation in our analysis is to introduce a linear perturbation in the empirical objective function. In that way, we transform the problem of quantifying the bias into comparison of the expected supremum of the empirical objective function and that of its limiting Gaussian process. To bound the difference of these expected suprema, we adopt similar techniques that have been recently studied by Chernozhukov et al. (2013) and Chernozhukov et al. (2014). Specifically, they compared a function of the maximum for sum of mean-zero Gaussian random vectors with that of multivariate mean-zero random vectors with the same covariance function, and provided an associated coupling inequality. We improve their arguments by providing more accurate approximation results (Lemma A.3) for the identity function of maximums as needed in our applications.

Another major contribution of this paper is to provide a tail inequality for cubic-rate M-estimators (Theorem 5.1). This helps us to construct a truncated estimator with bounded second moment, which is essential to apply Lyapunov's central limit theorem for establishing the normality of the aggregated estimator. Under some additional tail assumptions on the underlying empirical process, our results can be viewed as a generalization of empirical process theories that establish consistency and $n^{1/3}$ convergence rate for the M-estimators. Based on the results, we show that the asymptotic variance of the aggregated estimator can be consistently estimated by the sample variance of individual M-estimators in each group, which largely simplifies the inference procedure for M-estimators.

The rest of the paper is organized as follows. We describe the divide and conquer method for M-estimators and state the major central limit theorem (Theorem 2.1) in Section 2. Three examples for the location estimator, maximum score estimator and value search estimator are presented in Section 3 to illustrate the application of Theorem 2.1. In Section 4, we demonstrate the empirical performance of the aggregated estimators using both simulation studies and an application to the Yahoo! Front Page Today Module user click log dataset. Section 5 studies a tail inequality that are needed to prove Theorem 2.1, followed by a Discussion Section. All the technical proofs are provided in the Appendix.

2 Method

The divide and conquer scheme for M-estimators is described as follows. In the first step, the data are randomly divided into several groups. For the jth group, consider the following M-estimator

$$\hat{\theta}^{(j)} = \arg\max_{\theta \in \Theta} \mathbb{P}_{n_j}^{(j)} m(\cdot, \theta) \equiv \arg\max_{\theta \in \Theta} \frac{1}{n_j} \sum_{i=1}^{n_j} m(X_i^{(j)}, \theta), \qquad j = 1, \ldots, S,$$

where $(X_1^{(j)}, \ldots, X_{n_j}^{(j)})$ denote the data for the jth group, n_j is the number of observations in the jth group, S is the number of groups, $m(\cdot, \cdot)$ is the objective function and θ is a d-dimensional vector of parameters that belong to a compact parameter space Θ. In the second step, the aggregated estimator $\hat{\theta}_0$ is obtained as a weighted average of all subgroup estimators,

$$\hat{\theta}_0 = \sum_{j=1}^{S} \omega_j \hat{\theta}^{(j)} = \frac{\sum_{j=1}^{S} n_j^{2/3} \hat{\theta}^{(j)}}{\sum_{j=1}^{S} n_j^{2/3}}. \tag{1}$$

Remark 2.1 *The weights ω_j's are chosen such that $\hat{\theta}_0$ achieves the smallest asymptotic covariance matrix among the class of linearly aggregated estimators $\{\theta_\omega = \sum_j \omega_j \hat{\theta}^{(j)} | \sum_j \omega_j = 1, \omega_j \geq 0, \forall j = 1, \ldots, S\}$ (see Section F in the supplementary appendix for detailed illustrations). When $n_1 = n_2 = \cdots = n_S$, $\hat{\theta}_0$ reduces to a simple average of all $\hat{\theta}^{(j)}$'s.*

We assume that all the $X_i^{(j)}$'s are independent and identically distributed across i and j. Here, we only consider M-estimation with non-smooth functions $m(\cdot, \theta)$ of θ, and the resulting M-estimators $\hat{\theta}^{(j)}$'s have a convergence rate of $O_p(n_j^{-1/3})$. Such cubic-rate M-estimators have been widely studied in the literature, for example, the location estimator and maximum score estimator as demonstrated in the next section. Define $N = \sum_j n_j$ and $n = N/S$. The main goal of this paper is to establish the convergence rate and asymptotic normality of $\hat{\theta}_0$ under suitable conditions for S and n_j's.

Before introducing our main results, we first provide an intuitive explanation here why the divide and conquer method can improve the efficiency in cubic-rate M-estimation problems. Assume for now, $n_1 = n_2 = \cdots = n_S = n$ and S is fixed. Following Kim and Pollard

(1990), we can show that

$$n^{1/3}(\hat{\theta}^{(j)} - \theta_0) \xrightarrow{d} h_0,$$
$$N^{1/3}(\tilde{\theta}_0 - \theta_0) \xrightarrow{d} h_0,$$

where $\tilde{\theta}_0$ is the pooled estimator, i.e, $\tilde{\theta}_0 = \arg\max_{\theta \in \Theta} \sum_{i,j} m(X_i^{(j)}, \theta)$, θ_0 is the unique maximizer of $\mathrm{E}\{m(\cdot, \theta)\}$ and $h_0 = \arg\max_h Z(h)$ with

$$Z(h) = G(h) - \frac{1}{2} h^T V h. \tag{2}$$

Here G is a mean-zero Gaussian process and $V = \partial^2 \mathrm{E}\{m(\cdot, \theta)\}/\partial\theta\theta^T|_{\theta=\theta_0}$ is a positive definite matrix.

Assume $||N^{1/3}(\tilde{\theta}_0 - \theta_0)||_2^2$ and $||n^{1/3}(\hat{\theta}^{(j)} - \theta_0)||_2^2$ are uniformly integrable. Then, we have

$$N^{2/3}\mathrm{E}(\tilde{\theta}_0 - \theta_0)(\tilde{\theta}_0 - \theta_0)^T \to \mathrm{cov}(h_0), \quad \text{as } N \to \infty \tag{3}$$
$$n^{2/3}\mathrm{E}(\hat{\theta}^{(j)} - \theta_0)(\hat{\theta}^{(j)} - \theta_0)^T \to \mathrm{cov}(h_0), \quad \text{as } N \to \infty.$$

Under equal allocation, $\hat{\theta}^{(j)}$'s are independent and identical. We have

$$\begin{aligned} & N^{2/3}\mathrm{E}\{(\hat{\theta}_0 - \theta_0)(\hat{\theta}_0 - \theta_0)^T\} \\ = \;& N^{2/3}\frac{1}{S^2}\sum_{j=1}^{S}\mathrm{E}\{(\hat{\theta}^{(j)} - \theta_0)(\hat{\theta}^{(j)} - \theta_0)^T\} + N^{2/3}\frac{1}{S^2}\sum_{j \neq k}\mathrm{E}\{(\hat{\theta}^{(j)} - \theta_0)\mathrm{E}(\hat{\theta}^{(k)} - \theta_0)^T\} \\ = \;& \frac{n^{2/3}}{S^{1/3}}\mathrm{E}\{(\hat{\theta}^{(1)} - \theta_0)(\hat{\theta}^{(1)} - \theta_0)^T\} + b_n b_n^T S^{2/3}(S-1)/S \to S^{-1/3}\mathrm{cov}(h_0), \end{aligned} \tag{4}$$

where $b_n = n^{1/3}\mathrm{E}(\hat{\theta}^{(j)} - \theta_0) = o(1)$ is the bias of $n^{1/3}\hat{\theta}^{(j)}$. Comparing (3) with (4), we can see that the aggregated estimator is more efficient than the pooled estimator in the fixed S scenario.

Now let S grow with N. As long as S satisfies $S = O(1/(||b_n||_2^2))$, we have

$$b_n b_n^T S^{2/3}(S-1)/S \to O(S^{-1/3}),$$

and hence $N^{2/3}\mathrm{E}\{(\hat{\theta}_0 - \theta_0)(\hat{\theta}_0 - \theta_0)^T\} = O(S^{-1/3})$. In view of (3), this implies that the aggregated estimator can have a faster convergence rate than the pooled estimator.

2.1 Main results

We assume the dimension d is fixed, while the number of groups $S \to \infty$ as $N \to \infty$. Let $|| \cdot ||_2$ denote the Euclidean norm for vectors or induced matrix L_2 norm for matrices. We first introduce some conditions.

(A1.) There exists a small neighborhood $N_\delta = \{\theta : ||\theta - \theta_0||_2 \le \delta\}$ in which $\mathrm{E}m\{(\cdot,\theta)\}$ is twice continuously differentiable with the Hessian matrix $-V(\theta)$, where $V(\theta)$ is positive definite in N_δ. Moreover, assume $\mathrm{E}\{m(\cdot,\theta_0)\} > \sup_{\theta \in N_\delta^c} \mathrm{E}\{m(\cdot,\theta)\}$.
(A2.) For any $\theta_1, \theta_2 \in N_\delta$, we have $\mathrm{E}\{|m(\cdot,\theta_1) - m(\cdot,\theta_2)|^2\} \le K||\theta_1 - \theta_2||_2$ for a constant K that is independent of θ_1 and θ_2.
(A3.) There exists some positive constant ω such that $|m(x,\theta)| \le \omega$ for all x and θ.
(A4.) The envelope function $M_R(\cdot) \equiv \sup_\theta \{|m(\cdot,\theta)| : ||\theta - \theta_0||_2 \le R\}$ satisfies $\mathrm{E}M_R^2 = O(R)$ when $R \le \delta$.
(A5.) The set of functions $\{m(\cdot,\theta) | \theta \in \Theta\}$ has Vapnik-Chervonenkis (VC) index $1 \le v < \infty$.
(A6.) For any $\theta \in N_\delta$, $||V(\theta) - V||_2 = O(||\theta - \theta_0||_2)$, where $V = V(\theta_0)$.
(A7.) Let $L(\cdot)$ denote the variance process of $G(\cdot)$ satisfying $L(h) > 0$ whenever $h \neq 0$. (i) The function $L(\cdot)$ is symmetric and continuous, and has the rescaling property: $L(kh) = kL(h)$ for $k > 0$. (ii) For any $h_1, h_2 \in \mathbb{R}^d$ satisfying $||h_1||_2 \le n^{1/3}\delta$ and $||h_2||_2 \le n^{1/3}\delta$, we have

$$\left| L(h_1 - h_2) - n^{1/3}\mathrm{E}\left\{m(\cdot,\theta_0 + n^{-1/3}h_1) - m(\cdot,\theta_0 + n^{-1/3}h_2)\right\}^2 \right| = O\left(\frac{(||h_1|| + ||h_2||)^2}{n^{1/3}}\right).$$

(A8.) Let $c_j = n_j/n$. Assume there exists some constant $\bar{c} > 1$ such that $1/\bar{c} \le c_j \le \bar{c}$ for all j.

Theorem 2.1 *Under Conditions (A1)-(A8), if $S = o(n^{1/6}/\log^{5/6} n)$ and $S \to \infty$ as $n \to \infty$, we have*

$$\sqrt{c_1^{2/3} + \cdots + c_S^{2/3}}\, n^{1/3}(\hat{\theta}_0 - \theta_0) \xrightarrow{d} N(0, A), \tag{5}$$

for some positive definite matrix A.

Remark 2.2 *Under Condition A8, Theorem 2.1 suggests that $\hat{\theta}_0$ converges at a rate of $O_p(S^{-1/2}n^{-1/3})$. In contrast, the original M-estimator obtained based on pooled data has a convergence rate of $O_p(S^{-1/3}n^{-1/3})$. This implies that we can gain efficiency by adopting the split and conquer scheme for cubic-rate M-estimators. Such result is interesting as most aggregated estimators in the divide and conquer literature share the same convergence rates as the original estimators based on pooled data.*

Remark 2.3 *The constraints on S suggest that the number of group cannot diverge too fast. A main reason as we showed in the proof of Theorem 2.1 is that if S grows too fast, the asymptotic normality of $\hat{\theta}_0$ will fail due to accumulation of bias in the aggregation of subgroup estimators. Given a data of size N, we can take $S \approx N^l$, $n = N/S \approx N^{1-l}$ with $l < 1/7$ to fulfill this requirement. It turns out that this requirement on S can be relaxed under some special cases. In particular, when $d = 1$, i.e, θ_0 is a scalar, we show in the supplementary appendix that the aggregated estimator is asymptotically normal as long as $S \leq N^l$ with $l < 4/13$. Details can be found in Section A.5 of the supplementary appendix.*

Remark 2.4 *Conditions A1 - A5 and A7 (i) are similar to those in Kim and Pollard (1990) and are used to establish the cubic-rate convergence of the M-estimator in each group. Conditions A6 and A7 (ii) are used to establish the normality of the aggregated estimator. In particular, Condition A7 (ii) implies that the Gaussian process $G(\cdot)$ has stationary increments, i.e. $E[\{G(h_1) - G(h_2)\}^2] = L(h_1 - h_2)$ for any $h_1, h_2 \in \mathbb{R}^d$, which is used to control the bias of the aggregated estimator. Condition A8 automatically holds when $n_1 = \cdots = n_S$.*

In the rest of this section, we give a sketch for the proof of Theorem 2.1. The details of the proof are given in Section 5 and Section A in the supplementary appendix. Let $\hat{h}^{(j)} = n_j^{1/3}(\hat{\theta}^{(j)} - \theta_0)$. By definition, it is equivalent to show

$$\frac{1}{\sqrt{c_1^{2/3} + \cdots + c_S^{2/3}}} \sum_{j=1}^{S} c_j^{1/3} \hat{h}^{(j)} \xrightarrow{d} N(0, A). \tag{6}$$

When S diverges, intuitively, (6) follows by a direct application of central limit theorem for triangular arrays (cf. Theorem 11.1.1, Athreya and Lahiri, 2006). However, a few challenges remain. First, the estimator $\hat{h}^{(j)}$ may not possess finite second moment. Analogous to Kolmogorov's 3-series theorem (cf. Theorem 8.3.5, Athreya and Lahiri, 2006), we handle this by first defining $\tilde{h}^{(j)}$, which is a truncated version of $\hat{h}^{(j)}$ with $||\tilde{h}^{(j)}||_2 \leq \delta_{n_j}$ for some $\delta_{n_j} > 0$, such that $\sum_j \hat{h}^{(j)}$ and $\sum_j \tilde{h}^{(j)}$ are tail equivalent, i.e.

$$\lim_k \Pr\left(\bigcap_{n\geq k}\left\{\sum_{j=1}^{S(n)} c_j^{1/3}\hat{h}^{(j)} = \sum_{j=1}^{S(n)} c_j^{1/3}\tilde{h}^{(j)}\right\}\right) = 1.$$

Using Borel-Cantelli lemma, it suffices to show

$$\sum_n \Pr\left(\sum_{j=1}^{S(n)} c_j^{1/3}\hat{h}^{(j)} \neq \sum_{j=1}^{S(n)} c_j^{1/3}\tilde{h}^{(j)}\right) < \infty. \tag{7}$$

Now it remains to show

$$\frac{1}{\sqrt{\sum_j c_j^{2/3}}}\sum_{j=1}^{S} c_j^{1/3}\tilde{h}^{(j)} = \frac{1}{\sqrt{\sum_j c_j^{2/3}}}\sum_{j=1}^{S}\left\{\tilde{h}^{(j)} - \mathrm{E}(\tilde{h}^{(j)})\right\} + \frac{1}{\sqrt{\sum_j c_j^{2/3}}}\sum_j \mathrm{E}c_j^{1/3}\tilde{h}^{(j)} \xrightarrow{d} N(0, A).$$

The second challenge is to control the accumulated bias in the aggregated estimator, i.e. showing

$$\frac{1}{\sqrt{\sum_j c_j^{2/3}}}\sum_j c_j^{1/3}\mathrm{E}(\tilde{h}^{(j)}) \to 0,$$

or

$$\sqrt{S}\sup_j |\mathrm{E}(\tilde{h}^{(j)})| \to 0, \tag{8}$$

by Assumption A8. Finally, it remains to show that the second and third moments of $\tilde{h}^{(j)}$ satisfies

$$\sup_j |\mathrm{E}(a^T\tilde{h}^{(j)})^2 - a^T A a| \to 0, \tag{9}$$

$$\sup_j \mathrm{E}||\tilde{h}^{(j)}||_2^3 < \infty, \tag{10}$$

for any $a \in \mathbb{R}^d$. When (7), (8), (9) and (10) are established, Theorem 2.1 follows by Lyapunov's central limit theorem (cf. Corollary 11.1.4 Athreya and Lahiri, 2006). Section 5 is devoted to verifying (7), (9) and (10), while Section A in the supplementary appendix is devoted to proving (8).

3 Applications

In this section, we illustrate our main theorem (Theorem 2.1) with three applications including simple one-dimensional location estimator (Example 3.1) and more complicated multi-dimensional estimators with some constraints, such as maximum score estimator (Example 3.2) and value-search estimator (Example 3.3).

3.1 Location estimator

Let $X_i^{(j)}$ $(i = 1, \dots, n; j = 1, \dots, S)$ be i.i.d. random variables on the real line, with a continuous density p. In each subgroup j, consider the location estimator

$$\hat{\theta}^{(j)} = \arg\max_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} I(\theta - 1 \le X_i \le \theta + 1).$$

It was shown in Example 3.2.13 of van der Vaart and Wellner (1996) and Example 6.1 of Kim and Pollard (1990) that each $\hat{\theta}^{(j)}$ has a cubic-rate convergence. We assume that $\Pr(X \in [\theta - 1, \theta + 1])$ has a unique maximizer at θ_0. When the derivative of p exists and is continuous, $p'(\theta_0 - 1) - p'(\theta_0 + 1) > 0$ implies that the second derivative of $\Pr(X \in [\theta - 1, \theta + 1])$ is negative for all θ within some small neighbor N_δ around θ_0. Therefore, Condition (A2) holds, since

$$\begin{aligned} & \mathrm{E}|I(\theta_1 - 1 \le X \le \theta_1 + 1) - I(\theta_2 - 1 \le X \le \theta_2 + 1)|^2 \\ = \; & \Pr(\theta_1 - 1 \le X \le \theta_2 - 1) + \Pr(\theta_1 + 1 \le X \le \theta_2 + 1) \\ \le \; & \sup_{\theta \in N_\delta} \{p'(\theta - 1) + p'(\theta + 1)\} |\theta_1 - \theta_2|, \end{aligned}$$

for $\theta_1 \le \theta_2$ and $|\theta_1 - \theta_2| < 0.5$. Moreover, if we further assume that p has continuous second derivative in the neighborhood N_δ, Condition (A6) is satisfied.

The class of functions $\{|I(\theta - 1 \le X \le \theta + 1)| : \theta \in \Theta\}$ is bounded by 1 and belongs to VC class. In addition, we have

$$\begin{aligned} & \sup_{|\theta - \theta_0| < \epsilon} |I(\theta - 1 \le X \le \theta + 1) - I(\theta_0 - 1 \le X \le \theta_0 + 1)| \\ \le \; & I(\theta_0 - 1 - \epsilon \le X \le \theta_0 - 1 + \epsilon) + I(\theta_0 + 1 - \epsilon \le X \le \theta_0 + 1 + \epsilon), \end{aligned}$$

for small ϵ. The $L_2(P)$ norm of the function on the second line is $O(\sqrt{\epsilon})$. Hence, Conditions (A4) and (A5) hold.

Next, we claim that Condition (A7) holds for function $L(h) \equiv 2p(\theta_0 + 1)|h|$, or equivalently $\{p(\theta_0 - 1) + p(\theta_0 + 1)\}|h|$, since $p(\theta_0 - 1) = p(\theta_0 + 1)$. Obviously, $L(\cdot)$ is symmetric and satisfies the rescaling property. For any h_1, h_2 such that $\max(|h_1|, |h_2|) \le n^{1/3}\delta$, we define $\theta_1 = \theta_0 + n^{-1/3}h_1 \in N_\delta$ and $\theta_2 = \theta_0 + n^{-1/3}h_2 \in N_\delta$. Let $[a, b]$ denote the indicator function $I(a \le X \le b)$. Assume $h_1 \le h_2$. We have

$$\begin{aligned} & n^{1/3}\mathrm{E}\,|[\theta_1 - 1, \theta_1 + 1] - [\theta_2 - 1, \theta_2 + 1]|^2 = n^{1/3}\mathrm{E}[\theta_1 - 1, \theta_2 - 1] + n^{1/3}\mathrm{E}[\theta_1 + 1, \theta_2 + 1] \\ = \;& n^{1/3}\int_{\theta_1 - 1}^{\theta_2 - 1} p(\theta)d\theta + n^{1/3}\int_{\theta_1 + 1}^{\theta_2 + 1} p(\theta)d\theta = \{p(\theta_0 + 1) + p(\theta_0 - 1)\}(h_2 - h_1) + R, \end{aligned}$$

where the remainder term R is bounded by

$$\begin{aligned} & \sup_{\theta_1 \le \theta \le \theta_2} \left(|p(\theta - 1) - p(\theta_0 - 1)| + |p(\theta + 1) - p(\theta_0 + 1)|\right)(h_2 - h_1) \\ \le \;& \sup_{\theta \in N_\delta} 4n^{-1/3}|p'(\theta)|(h_2 - h_1)\max(|h_1|, |h_2|) \le \sup_{\theta \in N_\delta} 4n^{-1/3}|p'(\theta)|(|h_1| + |h_2|)^2, \end{aligned}$$

using a first order Taylor expansion. The case when $h_1 > h_2$ can be similarly discussed. Therefore, Condition (A7) holds. Theorem 2.1 then follows.

3.2 Maximum score estimator

Consider the regression model $Y_i^{(j)} = {X_i^{(j)}}^T\beta_0 + e_i^{(j)}$, $, j = 1, \cdots, S$, where $X_i^{(j)}$ is a d-dimensional vector of covariates and $e_i^{(j)}$ is the random error. Assume that $(X_i^{(j)}, e_i^{(j)})$'s are i.i.d. copies of (X, e). The maximum score estimator is defined as

$$\hat{\beta}^{(j)} = \arg\max_{||\beta||_2 = 1} \sum_{i=1}^{n} \{I(Y_i^{(j)} \ge 0, {X_i^{(j)}}^T\beta \ge 0) + I(Y_i^{(j)} < 0, {X_i^{(j)}}^T\beta < 0)\},$$

where the constraint $||\beta||_2 = 1$ is to guarantee the uniqueness of the maximizer.

Assume $||\beta_0|| = 1$, otherwise we can define $\beta^\star = \beta_0/||\beta_0||_2$ and establish the asymptotic distribution of $\hat{\beta}_0 - \beta^\star$ instead. It was shown in Example 6.4 of Kim and Pollard (1990) that $\hat{\beta}^{(j)}$ has a cubic-rate convergence, when (i) median$(e|X) = 0$; (ii) X has a bounded, continuously differentiable density p; and (iii) the angular component of X has a bounded continuous density with respect to the surface measure on $\mathbb{S}^{d-1}$, which corresponds to the unit sphere in $\mathbb{R}^d$.

Theorem 2.1 is not directly applicable to this example since Assumption (A1) is violated. The Hessian matrix

$$V = -\frac{\partial^2 \mathrm{E}\{I(Y_i^{(j)} \geq 0, {X_i^{(j)}}^T\beta \geq 0) + I(Y_i^{(j)} < 0, {X_i^{(j)}}^T\beta < 0)\}}{\partial\beta\beta^T}|_{\beta_0}$$

is not positive definite. One possible solution is to use the arguments from the constrained M-estimator literature (e.g. Geyer, 1994) to approximate the set $||\beta||_2 = 1$ by a hyperplane $(\beta - \beta_0)^T\beta = 0$, and obtain a version of Theorem 2.1 for the constrained cubic-rate M-estimators. We adopt an alternative approach here, and consider a simple reparameterization to make Theorem 2.1 applicable.

By Gram-Schmidt orthogonalization, we can obtain an orthogonal matrix $[\beta_0, U_0]$ with U_0 being a $\mathbb{R}^{d\times(d-1)}$ matrix subject to the constraint $U_0^T\beta_0 = 0$. Define

$$\beta(\theta) = \sqrt{1 - ||\theta||_2^2}\beta_0 + U_0\theta, \tag{11}$$

for all $\theta \in \mathbb{R}^{d-1}$ and $||\theta||_2 \leq 1$. Take Θ to be the unit ball B_2^{d-1} in $\mathbb{R}^{d-1}$. Define

$$\hat{\theta}^{(j)} = \arg\max_{\theta\in\Theta} \sum_{i=1}^n [I(Y_i^{(j)} \geq 0, {X_i^{(j)}}^T\beta(\theta) \geq 0) + I(Y_i^{(j)} < 0, {X_i^{(j)}}^T\beta(\theta) < 0)].$$

Note that under the assumption median$(e|X) = 0$, we have $\theta_0 = 0$.

Let $m(y, x, \beta) = I(y \geq 0, x^T\beta \geq 0) + I(y < 0, x^T\beta < 0)$. Define

$$\kappa(x) = \mathrm{E}\{I(e + X^T\beta_0 \geq 0) - I(e + X^T\beta_0 < 0)|X = x\}.$$

It is shown in Kim and Pollard (1990) that

$$\frac{\partial\mathrm{E}\{m(\cdot,\cdot,\beta)\}}{\partial\beta} = ||\beta||_2^{-2}\beta^T\beta_0(I + ||\beta||_2^{-2}\beta\beta^T)\int_{x^T\beta_0=0}\kappa(T_\beta x)p(T_\beta x)d\sigma, \tag{12}$$

where

$$T_\beta = (I - ||\beta||_2^{-2}\beta\beta^T)(I - \beta_0\beta_0^T) + ||\beta||_2^{-1}\beta\beta_0^T,$$

and σ is the surface measure on the line $x^T\beta_0 = 0$.

Note that $\partial\beta(\theta)/\partial\theta$ has finite derivatives for all orders as long as $||\theta||_2 < 1$. Assume that κ and p have twice continuous derivatives. This together with (12) implies that $\mathrm{E}\{m(\cdot,\cdot,\beta(\theta))\}$ has third continuous derivative as a function of θ in a small neighborhood N_δ ($\delta < 1$) around 0. This verifies (A6). Moreover, for any θ_1, $\theta_2 \in N_\delta$ with $||\theta_1 - \theta_2||_2 \leq \epsilon$, we have

$$\begin{aligned} & ||\beta(\theta_1) - \beta(\theta_2)||_2^2 = ||\theta_1 - \theta_2||_2^2 + \left(\sqrt{1 - ||\theta_1||_2^2} - \sqrt{1 - ||\theta_2||_2^2}\right)^2 \\ = \;& ||\theta_1 - \theta_2||_2^2 + \frac{(1 - ||\theta_1||_2^2 - 1 + ||\theta_2||_2^2)^2}{\left(\sqrt{1 - ||\theta_1||_2^2} + \sqrt{1 - ||\theta_2||_2^2}\right)^2} \leq \frac{2||\theta_1 - \theta_2||_2^2}{1 - \delta^2}. \end{aligned} \tag{13}$$

Kim and Pollard (1990) showed that $\mathrm{E}\{|m(\cdot,\cdot,\beta_1) - m(\cdot,\cdot,\beta_2)|\} = O(||\beta_1 - \beta_2||_2)$ near β_0. This together with (13) implies

$$\mathrm{E}\{|m(\cdot,\cdot,\beta(\theta_1)) - m(\cdot,\cdot,\beta(\theta_2))|^2\} \leq 2\mathrm{E}\{|m(\cdot,\cdot,\beta(\theta_1)) - m(\cdot,\cdot,\beta(\theta_2))|\} = O(||\theta_1 - \theta_2||_2).$$

Therefore, (A2) is satisfied and (A3) trivially holds since $|m| \leq 1$.

It was also shown in Kim and Pollard (1990) that the envelope M_ϵ of the class of functions $\{m(\cdot,\cdot,\beta) - m(\cdot,\cdot,\beta_0) : ||\beta - \beta_0||_2 \leq \epsilon\}$ satisfies $\mathrm{E}M_\epsilon^2 = O(\epsilon)$. Using (13), we can show that the envelope $\tilde{M}_\epsilon$ of the class of functions $\{m(\cdot,\cdot,\beta(\theta)) - m(\cdot,\cdot,\beta_0) : ||\theta||_2 \leq \epsilon\}$ also satisfies $\mathrm{E}\tilde{M}_\epsilon^2 = O(\epsilon)$. Thus, (A4) is satisfied. Moreover, since the class of functions $m(\cdot,\cdot,\beta)$ over all β belongs to the VC class, so does the class of function $m(\cdot,\cdot,\beta(\theta))$. This verifies (A5).

Finally, we establish (A7). For any $\theta_1, \theta_2 \in N_\delta$, define $h_1 = n^{1/3}\theta_1$ and $h_2 = n^{1/3}\theta_2$. We have

$$\begin{aligned} & n^{1/3}\mathrm{E}\left\{\left|m(Y,X,\beta(h_1/n^{1/3})) - m(Y,X,\beta(h_2/n^{1/3}))\right|^2\right\} \\ = \;& n^{1/3}\mathrm{E}\left\{\left|I(X^T\beta(h_1/n^{1/3}) \geq 0) - I(X^T\beta(h_2/n^{1/3}) \geq 0)\right| I(Y \geq 0)\right\} \\ + \;& n^{1/3}\mathrm{E}\left\{\left|I(X^T\beta(h_1/n^{1/3}) < 0) - I(X^T\beta(h_2/n^{1/3}) < 0)\right| I(Y < 0)\right\} \\ = \;& n^{1/3}\mathrm{E}\left\{\left|I(X^T\beta(h_1/n^{1/3}) \geq 0) - I(X^T\beta(h_2/n^{1/3}) \geq 0)\right|\right\}. \end{aligned} \tag{14}$$

We write X as $r\beta_0 + z$ with z orthogonal to β_0. Equation (14) can be written as

$$n^{1/3}\mathrm{E}\left\{\left|I\left(r\sqrt{1 - \left|\left|\frac{h_1}{n^{1/3}}\right|\right|_2^2} + z^T U\frac{h_1}{n^{1/3}} \geq 0\right) - I\left(r\sqrt{1 - \left|\left|\frac{h_2}{n^{1/3}}\right|\right|_2^2} + z^T U\frac{h_2}{n^{1/3}} \geq 0\right)\right|\right\}. \tag{15}$$

Define $\omega = n^{1/3}r$. Equation (15) can be expressed as

$$\int\int I(-z^TUh_1(1-n^{-2/3}||h_1||_2^2)^{-1/2} > \omega \geq -z^TUh_2(1-n^{-2/3}||h_2||_2^2)^{-1/2})p\left(\frac{\omega}{n^{1/3}}, z\right) d\omega dz.$$

Assume that $p(r, z)$ is differentiable with respect to r and $|\partial p(r, z)/\partial r| \leq q(z)$ for some function q. Then, (15) is equal to

$$\begin{aligned} & \int |z^TU\{h_1(1-n^{-2/3}||h_1||_2^2)^{-1/2} - h_2(1-n^{-2/3}||h_2||_2^2)^{-1/2}\}|p(0, z)dz + R_1 \\ = & \int |z^TU(h_1 - h_2)|p(0, z)dz + R_1 + R_2, \end{aligned}$$

where the remainders $|R_1|$ and $|R_2|$ are bounded by

$$|R_1| \leq \int n^{-1/3}\{(z^TUh_1)^2 + (z^TUh_2)^2\}q(z)dz = O(n^{-1/3}\{||h_1||_2^2 + ||h_2||_2^2\}),$$

and

$$\begin{aligned} |R_2| \ \leq \ & |(1-n^{-2/3}||h_1||_2^2)^{-1/2} - 1| \int |z^TUh_1|p(0, z)dz \\ + \ & |(1-n^{-2/3}||h_2||_2^2)^{-1/2} - 1| \int |z^TUh_2|p(0, z)dz \\ \leq \ & n^{-1/3}(||h_1||_2 + ||h_2||) \int (|z^TUh_1| + |z^TUh_2|)p(0, z)dz = O(n^{-1/3}\{||h_1||_2^2 + ||h_2||_2^2\}), \end{aligned}$$

under suitable moment assumptions on functions $p(0, z)$ and $q(z)$. This verifies (A7).

An application of Theorem 2.1 implies

$$\frac{1}{\sqrt{S}} \sum_{j=1}^{S} n^{1/3}\hat{\theta}^{(j)} \xrightarrow{d} N(0, A),$$

for some positive definite matrix $A \in \mathbb{R}^{(d-1)\times(d-1)}$. Hence

$$\frac{1}{\sqrt{S}} \sum_{j=1}^{S} n^{1/3}U\hat{\theta}^{(j)} \xrightarrow{d} N(0, UAU^T). \tag{16}$$

By the definition of $\hat{\theta}^{(j)}$ and $\hat{\beta}^{(j)}$, we have

$$\begin{aligned} & \left|\frac{1}{\sqrt{S}} \sum_{j=1}^{S} n^{1/3}(\hat{\beta}^{(j)} - \beta_0 - U\hat{\theta}^{(j)})\right| \leq \left|\frac{1}{\sqrt{S}} \sum_{j=1}^{S} n^{1/3}|\sqrt{1-||\hat{\theta}^{(j)}||_2^2} - 1|\right| \\ \leq \ & \left|\frac{1}{\sqrt{S}} \sum_{j=1}^{S} n^{1/3} \frac{|1-||\hat{\theta}^{(j)}||_2^2 - 1|}{|\sqrt{1-||\hat{\theta}^{(j)}||_2^2} + 1|}\right| \leq \frac{n^{1/3}}{\sqrt{S}} \sum_{j} ||\hat{\theta}^{(j)}||_2^2. \end{aligned}$$

With probability at least $1-S/n \to 1$, the last expression is equal to $O(\sqrt{S}n^{1/3}n^{-2/3}\log^{2/3} n) = o(1)$, which is implied by the tail inequality for $\hat{\theta}^{(j)}$ established in Theorem 5.1. Combining this together with (16), we have

$$\frac{1}{\sqrt{S}}\sum_{j=1}^{S} n^{1/3}(\hat{\beta}^{(j)} - \beta_0) \xrightarrow{d} N(0, UAU^T).$$

3.3 Value search estimator

The value search estimator was introduced by Zhang et al. (2012) for estimating the optimal treatment regime. The data can be summarized as i.i.d. triples $\{O_i^{(j)} = (X_i^{(j)}, A_i^{(j)}, Y_i^{(j)}), i = 1, \ldots, n; j = 1, \ldots, S\}$, where $X_i^{(j)} \in \mathbb{R}^d$ denote patient's baseline covariates, $A_i^{(j)}$ is the treatment received by the patient taking the value 0 or 1, and $Y_i^{(j)}$ is the response, the larger the better by convention. Consider the following model

$$Y_i^{(j)} = \mu(X_i^{(j)}) + A_i^{(j)} C(X_i^{(j)}) + e_i^{(j)}, \tag{17}$$

where $\mu(\cdot)$ is the baseline mean function, $C(\cdot)$ is the contrast function, and $e_i^{(j)}$ is the random error with $E\{e_i^{(j)} | A_i^{(j)}, X_i^{(j)}\} = 0$. The optimal treatment regime is defined in the potential outcome framework. Specifically, let $Y_i^{(j)\star}(0)$ and $Y_i^{(j)\star}(1)$ be the potential outcomes that would be observed if the patient received treatment 0 or 1, accordingly. For a treatment regime d that maps $X_i^{(j)}$ to $\{0, 1\}$, define the potential outcome

$$Y_i^{(j)\star}(d) = d(X_i^{(j)}) Y_i^{(j)\star}(1) + \{1 - d(X_i^{(j)})\} Y_i^{(j)\star}(0).$$

The optimal regime d^{opt} is defined as the rule that maximizes the expected potential outcome, i.e, the value function, $\mathrm{E}\{Y_i^{(j)\star}(d)\}$. Under the stable unit treatment value assumption (SUTVA) and no unmeasured confounders assumption (Splawa-Neyman, 1990), the optimal treatment regime under model (17) is given by $d^{opt}(x) = I\{C(x) > 0\}$.

The true contrast function $C(\cdot)$ can be complex. As suggested by Zhang et al. (2012), in practice we can find the restricted optimal regimen within a class of decision rules, such as linear treatment decision rules $d(x, \beta) = I(\beta_1 + x_1\beta_2 + \cdots + x_d\beta_{d+1} > 0)$ indexed by $\beta \in \mathbb{R}^{d+1}$, where the subscript k denotes the kth element in the vector. Let $\beta^\star = \arg\max_\beta V(\beta)$, where $V(\beta) = \mathrm{E}\{Y_i^{(j)\star}(d(X_i^{(j)}, \beta))\}$. To make $\beta^\star$ identifiable, we assume

$\beta_1^\star = -1$. Define $\theta^\star = (\beta_2^\star, \cdots, \beta_{d+1}^\star)^T$. The restricted optimal treatment regime is given by $\tilde{d}^{opt}(x, \theta^\star) = I(x^T\theta^\star > 1)$ and the value function is defined by $V(\theta) = \mathrm{E}\{Y_i^{(j)\star}(\tilde{d}(X_i^{(j)}, \theta))\}$ with $\tilde{d}(x, \theta) = I(x^T\theta > 1)$. Zhang et al. (2012) proposed an inverse propensity score weighted estimator of the value function $V(\theta)$ and the associated value search estimator by maximizing the estimated value function. Specifically, for each group j, the value search estimator is defined as

$$\hat{\theta}^{(j)} = \arg\max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{d}(X_i^{(j)}, \theta)A_i^{(j)} + \{1 - \tilde{d}(X_i^{(j)}, \theta)\}(1 - A_i^{(j)})}{\pi_i^{(j)}A_i^{(j)} + (1 - \pi_i^{(j)})(1 - A_i^{(j)})} Y_i^{(j)}, \tag{18}$$

where $\pi_i^{(j)} = \Pr(A_i^{(j)} = 1|X_i^{(j)})$ is the propensity score and known in a randomized study. Here, for illustration purpose, we assume that $\pi_i^{(j)}$'s are known.

Define $m(O_i^{(j)}, \theta) = \xi_i^{(j)}\tilde{d}(X_i^{(j)}, \theta)$, where

$$\xi_i^{(j)} = \frac{A_i^{(j)}}{\pi_i^{(j)}}C(X_i^{(j)}) + \frac{A_i^{(j)} - \pi_i^{(j)}}{\pi_i^{(j)}(1 - \pi_i^{(j)})}\left\{\mu(X_i^{(j)}) + e_i^{(j)}\right\} = \left(\frac{A_i^{(j)}}{\pi_i^{(j)}} - \frac{1 - A_i^{(j)}}{1 - \pi_i^{(j)}}\right)Y_i^{(j)}.$$

With some algebra, we can show that $\hat{\theta}^{(j)}$ also maximizes $\mathbb{P}_n^{(j)}m(\cdot, \theta)$, where $\mathbb{P}_n^{(j)}$ is the empirical measure for data in group j. Unlike the previous two examples, here the function m is not bounded. To fulfill (A3), we need $||\xi_i^{(j)}||_{\psi_1} < \infty$. This holds when $0 < \gamma_1 < \pi_i^{(j)} < \gamma_2 < 1$ for some constants γ_1 and γ_2, $||C(X_i^{(j)})||_{\psi_1} < \infty$, $||\mu(X_i^{(j)})||_{\psi_1} < \infty$ and $||e_i^{(j)}||_{\psi_1} < \infty$.

To show (A1) and (A6), we evaluate the integral

$$\Gamma(\theta) = \mathrm{E}\{\xi\tilde{d}(X, \theta)\} = \mathrm{E}\{C(X)\tilde{d}(X, \theta)\} = \int_{x^T\theta > 1} C(x)p(x)dx, \tag{19}$$

where $p(x)$ is the density function of $X_i^{(j)}$. Consider the transformation

$$T_\theta = (I - ||\theta||_2^{-2}\theta\theta^T) + ||\theta||_2^{-2}\theta(\theta^\star)^T,$$

which maps the region $\{x^T\theta^\star > 1\}$ onto $\{x^T\theta > 1\}$, and $\{x^T\theta^\star = 1\}$ onto $\{x^T\theta = 1\}$. We exclude the trivial case with $\theta^\star = 0$. The above definition is meaningful when θ is taken over a small neighborhood N_δ of $\theta^\star$. We assume that functions p and C are continuously differentiable. Note that

$$\frac{\partial T_\theta x}{\partial \theta} = -\frac{\{\theta^T x - (\theta^\star)^T x\}}{||\theta||_2^2}I - \frac{\theta x^T}{||\theta||_2^2} + \frac{2\theta\theta^T(x^T\theta - x^T\theta^\star)}{||\theta||_2^4}.$$

Using some differential geometry arguments similarly as in Section 5 of Kim and Pollard (1990), we can show that the integral (19) can be represented as

$$\Gamma(\theta) = \int_{x^T\theta^\star>1} \left[-\frac{1}{||\theta||_2^2}\theta^T \frac{\partial C(x)p(x)}{\partial x} x + \frac{\{\theta^T x - (\theta^\star)^T x\}}{||\theta||_2^4}\theta^T \frac{\partial C(x)}{\partial x}\theta - \frac{\theta^T x - (\theta^\star)^T x}{||\theta||_2^2}\frac{\partial C(x)p(x)}{\partial x} \right] dx,$$

which is thrice differentiable under certain conditions on $C(x)$, $p(x)$ and their derivatives.

To show (A7), we assume that the conditional density $p(x|y)$ of X given $Y = 1 - X^T\theta^\star$ exists and is continuously differentiable with respect to y. Similarly assume that the density $q(y)$ of Y exists and is continuously differentiable. Let $g(X) = \mathrm{E}(\xi^2|X)$. For any $h_1, h_2 \in \mathbb{R}^d$, we have

$$\begin{aligned} & n^{1/3}\mathrm{E}\left\{\xi^2 \left|I(X^T\theta^\star + n^{-1/3}X^T h_1 > 1) - I(X^T\theta^\star + n^{-1/3}X^T h_2 > 1)\right|^2\right\} \\ = \; & n^{1/3}\int g(x)\left|I(n^{-1/3}x^T h_1 > y) - I(n^{-1/3}x^T h_2 > y)\right| p(x|y)q(y)dxdy. \end{aligned}$$

Let $y = n^{-1/3}z$. The last expression in the above equation can be written as

$$\begin{aligned} & \int g(x)\left|I(x^T h_1 > z) - I(x^T h_2 > z)\right| p(x|0)q(0)dxdz + R \\ = \; & \int g(x)|x^T(h_1 - h_2)|p(x|0)q(0)dx + R, \end{aligned}$$

with the remainder term

$$R = \int g(x)\left|I(x^T h_1 > z) - I(x^T h_2 > z)\right| \{p(x|n^{-1/3}z)q(n^{-1/3}z) - p(x|0)q(0)\}dxdz,$$

which is $O(n^{-1/3}(||h_1||_2^2 + ||h_2||_2^2))$ under certain conditions on $q(x)$ and $p(x|\cdot)$. Conditions (A2) and (A4) can be similarly verified. Since the class of functions $\{g(x)I(x^T\theta > 1) : \theta \in \mathbb{R}^d\}$ has finite VC index, Condition (A5) also holds. Theorem 2.1 then follows.

4 Numerical studies

In this section, we examine the numerical performance of the aggregated M-estimator for the three examples studied in the previous section and compare it with the M-estimator based on pooled data, denoted as the pooled estimator.

4.1 Location estimator

The data X_j ($j = 1, \ldots, N$) were independently generated from the standard normal distribution. The true parameter θ_0 that maximizes $\mathrm{E}\{I(\theta - 1 \le X_j \le \theta + 1)\}$ was set to be 0. Let $\tilde{\theta}_0$ and $\hat{\theta}_0$ denote the pooled estimator and the aggregated estimator, respectively. To obtain $\hat{\theta}_0$, we randomly divided the data into S blocks with equal size $n = N/S$.

We took $N = 2^i$ for $i = 14, 16, 18, 20$, and choose $S = 2^j$ such that $0.2 \le j/i < 0.625$ when $N = 2^i$. For each combination of N and S, we estimated the standard error of $\hat{\theta}_0$ by

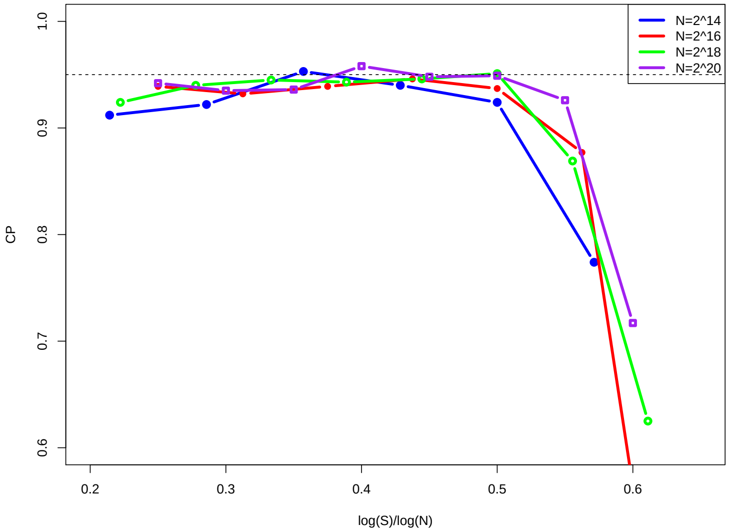
$$\widehat{SE}(\hat{\theta}_0) = \frac{1}{\sqrt{S}} \left\{ \frac{1}{S-1} \sum_{l=1}^{S} \left(\hat{\theta}^{(l)} - \hat{\theta}_0 \right)^2 \right\}^{1/2},$$

where $\hat{\theta}^{(l)}$ denotes the M-estimator for the lth group. For each scenario, we conducted 1000 simulation replications and plot the coverage probabilities of 95% predictive intervals in Figure 1. We also report the bias and sample standard deviation (denoted as SD) of estimators $\tilde{\theta}_0$ and $\hat{\theta}_0$, and mean of estimated standard errors and coverage probability (denoted as CP) of Wald-type 95% confidence interval for $\hat{\theta}_0$ in Table 1 of the supplementary appendix, for some of the scenarios where $N = 2^i$ for $i = 14, 16, 18, 20$, and $S = 2^j$ for $j = 4,5,6,7$. Unlike $\hat{\theta}_0$, $\tilde{\theta}_0$ doesn't converge to a tractable limiting distribution and it doesn't have a convenient variance estimator. Hence, in Table 1, we didn't provide the standard errors and confidence intervals for $\tilde{\theta}_0$.

From Figure 1, it is clear that for this specific application, the coverage probabilities are approximately 95% when $S \le S^*$ where $S^* \approx N^{0.55}$. In this example, the cubic rate estimator is one dimensional and according to Theorem A.1 and the discussion in Section A.5.1 in the supplementary appendix, the aggregated estimator is asymptotically normal when $S = O(N^l)$ for $0 < l < 4/13$. This is consistent with our numerical findings. Based on the results in Table 1, it can be seen that the aggregated estimator $\hat{\theta}_0$ has much smaller standard deviation than the pooled estimator $\tilde{\theta}_0$, indicating the efficiency gain by the divide and conquer scheme as shown in our theory. In addition, the bias of $\hat{\theta}_0$ generally becomes bigger and the standard deviation of $\hat{\theta}_0$ generally becomes smaller when S and N increase, and the normal approximation becomes more accurate when S increases. This demonstrates the bias-variance trade off for aggregated estimators. With properly chosen

S, the estimated standard error of $\hat{\theta}_0$ is close to its standard deviation and the coverage probability is close to the nominal level.

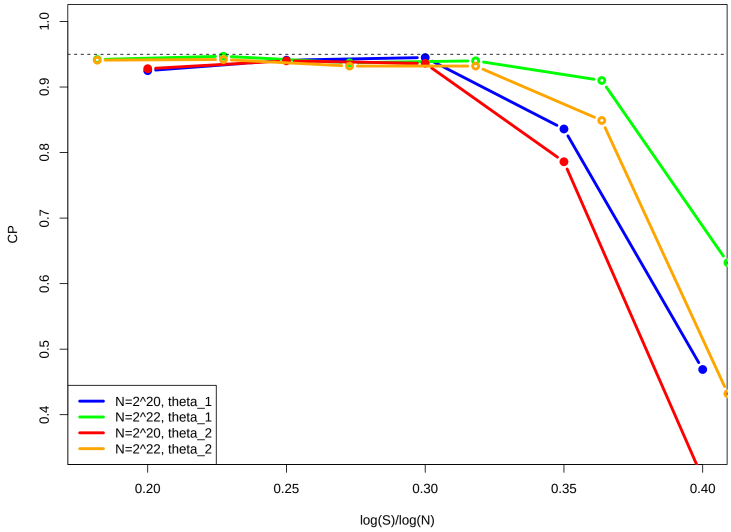
Figure 1: Coverage probability of 95% predictive interval with different choices of N and S, for the location estimator



4.2 Maximum score estimator

Consider the model $Y_i = 1.5X_{i1} - 1.5X_{i2} + 0.5e_i$, $i = 1, \cdots, N$, where X_{i1}, X_{i2} and e_i were generated independently from the standard normal. Hence, $\theta_0 = (\theta_1, \theta_2)^T = (1.5, -1.5)^T$. Let $\tilde{\theta}_0 = (\tilde{\theta}_1, \tilde{\theta}_2)^T$ denote the pooled estimator and $\hat{\theta}_0 = (\hat{\theta}_1, \hat{\theta}_2)^T$ the aggregated estimator. We set $N = 2^{20}, 2^{22}$ and $S = 2^j$ such that $0.18 \leq j/i \leq 0.42$ when $N = 2^i$. The coverage probabilities of 95% confidence intervals for $\hat{\theta}_1$ and $\hat{\theta}_2$ are plotted in Figure 2 based on 1000 replications. They are close to the nominal level when $S \leq S^*$ where $S^* \approx N^{0.32}$. This example can also be regarded as a one-dimensional cubic rate estimation problem since $\hat{\theta}_1$ and $\hat{\theta}_2$ satisfy the constraint: $\hat{\theta}_1^2 + \hat{\theta}_2^2 = 1$. Therefore, similar to the discussions in Section A.5.2, we can show $\hat{\theta}_1$ and $\hat{\theta}_2$ are asymptotically normal when $S = O(N^l)$ for $0 < l < 4/13$. This upper bound is close to S^* since $4/13 \approx 0.308$. Other results are given in Table 2 of the supplementary appendix. The findings are very similar to those for the location estimator in the previous example.

Figure 2: Coverage probability of 95% predictive interval with different choices of N and S, for the maximum score estimator



4.3 Value search estimator

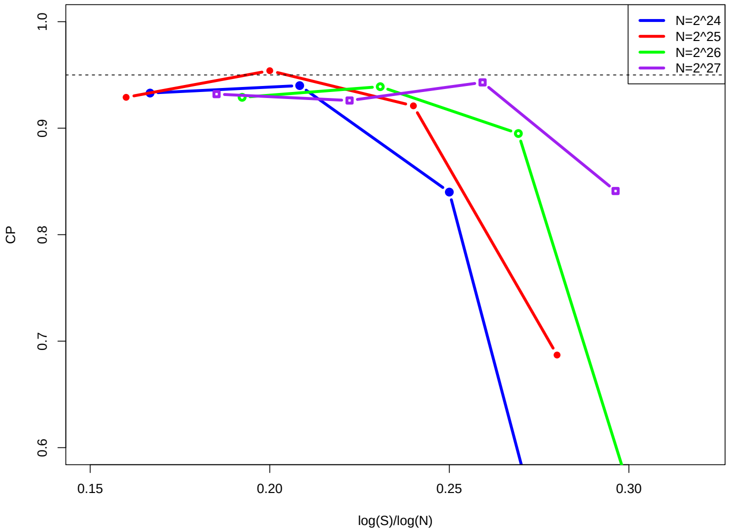
Consider the model $Y_i = 1 + A_i(2X_i - 1) + e_i$, $i = 1, \cdots, N$, where $X_i \sim N(0, 1)$, $e_i \sim N(0, 0.25)$, and $\Pr(A_j = 1) = 0.5$. Under this model assumption, the optimal treatment rule takes the form,

$$d^{opt}(x) = I(2x > 1),$$

and hence $\beta^\star = 2$.

We take $N = 2^{24}, 2^{25}, 2^{26}$ and 2^{27}. When $N = 2^{24}$ and 2^{25}, we choose $S = 2^j$ for $j = 4, 5, 6, 7$. When $N = 2^{26}$ and 2^{27}, we choose $S = 2^j$ for $j = 5, 6, 7, 8$. This gives a total of 16 scenarios. We plot the coverage probabilities of 95% predictive intervals for $\hat{\theta}_0$ in Figure 3, with these combinations of S and N. When $S \leq S^* \approx N^{0.27}$, the coverage probabilities are close to 95%. This is also a one-dimensional problem. Note that in this application, the rate 0.27 in the practical upper bound is slightly smaller than the theoretical upper bound $4/13 \approx 0.308$. However, it is noted that the theoretical upper bound is up to a scaling constant. When N becomes larger, the ratio $\log S^*/\log N$ should be close to or larger than 0.308. Details about the bias and the sample standard deviations of the aggregated estimator are given in Table 3 of the supplementary appendix.

Figure 3: Coverage probability of 95% predictive interval with different choices of N and S, for the value search estimator



4.4 Yahoo! Today Module user click log dataset

Online content recommendation services have received extensive attention both in the machine learning and statistics literature. These online services strive to make recommendations of advertisements or news articles to individual users by making use of both the content and user information. In this subsection, we apply the proposed method to a Yahoo! Today Module user click log dataset, which contains 45,811,883 user visits to the Today Module, during the first ten days in May 2009. Given such a large number of observations, it is extremely difficult to analyze the entire data on a single computer. This makes the divide and conquer method as an emerging need to deal with such large datasets.

For the ith visit, the dataset contains a binary response variable Y_i, an ID of the recommended article and a 6 dimensional feature vector of the user. Due to sensitivity and privacy concerns, feature definitions and article names were not included in the data. Here, $Y_i = 1$ means the user clicked the recommended article and $Y_i = 0$ means the user didn't click. The last element in the feature vector is always 1, and the first five sums to 1. Therefore, we took the first three and the fifth elements in the feature vector to form the covariates X_i. For illustration, we only consider a subset of data that contains visits on May 1st where the recommended article ID is either 109510 or 109520. There were a total of 50 candidate articles on May 1st. We chose these two articles since they were being recommended most on that day. This gives us a total of 405888 visits. On the reduced dataset, define $A_i = 1$ if the recommended article is 109510 and $A_i = 0$ otherwise. In this example, the online recommendation problem can be formulated as follows. Denoted by $\mathcal{D}$ a given set of functions that maps the covariate space to the space of article ID's. Our aim is to find the optimal recommendation strategy to maximize user's click through rate. We consider estimating the optimal recommendation rule among the set of linear decision functions $\mathcal{D} = \{I(x^T\theta > 1) : \forall\theta \in \mathbb{R}^4\}$. Hence, estimating the optimal recommendation strategy is similar to the problem of estimating the optimal treatment regime as described in Section 3.3. Specifically, we divide the data randomly into S pieces: $\{(X_i^{(j)}, A_i^{(j)}, Y_i^{(j)}) : i = 1, \dots, n_j\}_{j=1,\dots,S}$ and obtain

$$\begin{aligned}\hat{\theta}^{(j)} = \arg\max_{\theta\in\mathbb{R}^4} \frac{1}{n_j}\sum_{i=1}^{n_j}\Bigg\{ &\left(\frac{A_i^{(j)}}{\hat{\pi}_i^{(j)}} I(\theta^T X_i^{(j)} > 1) + \frac{1-A_i^{(j)}}{1-\hat{\pi}_i^{(j)}} I(\beta^T X_i^{(j)} \le 1)\right) Y_i^{(j)} \qquad (20)\\ + &\left(\frac{A_i^{(j)}}{\hat{\pi}_i^{(j)}} I(\theta^T X_i^{(j)} > 1) + \frac{1-A_i^{(j)}}{1-\hat{\pi}_i^{(j)}} I(\theta^T X_i^{(j)} \le 1) - 1\right) \{\hat{h}_{0i}^{(j)} I(\theta^T X_i^{(j)} \le 1) + \hat{h}_{1i}^{(j)} I(\theta^T X_i^{(j)} > 1)\}\Bigg\},\end{aligned}$$

as the subgroup estimator, where $\hat{\pi}_i^{(j)}$, $\hat{h}_{0i}^{(j)}$, $\hat{h}_{1i}^{(j)}$ are estimators of $\Pr(A_i^{(j)} = 1|X_i^{(j)})$, $\Pr(Y_i^{(j)} = 1|A_i^{(j)} = 0, X_i^{(j)})$ and $\Pr(Y_i^{(j)} = 1|A_i^{(j)} = 1, X_i^{(j)})$ respectively. The estimators $\hat{\pi}_i^{(j)}$, $\hat{h}_{0i}^{(j)}$, $\hat{h}_{1i}^{(j)}$ are obtained by logistic regressions. We chose n_j such that $\max_j n_j - \min_j n_j \le 1$. The estimated optimal recommendation strategy is given as $I(x^T\hat{\theta}_0 > 1)$ where $\hat{\theta}_0 = \sum_j \hat{\theta}^{(j)}/S$.

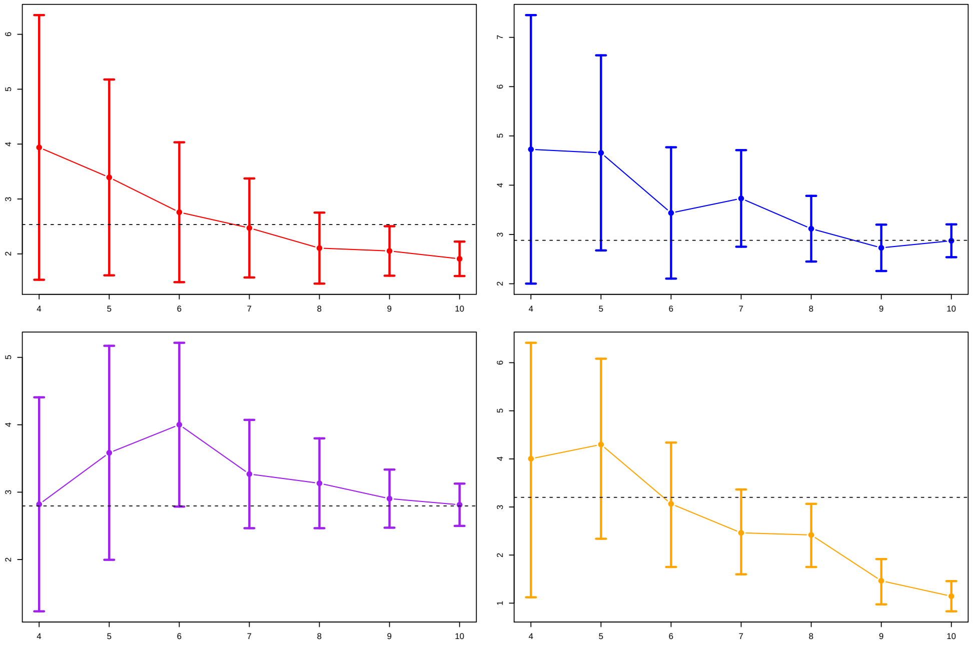
Remark 4.1 *Compared to the value search estimator defined in* (18)*, here we obtain the subgroup estimator by maximizing an augmented version of the inverse propensity score weighted estimator. The resulting estimator also converges at a rate of* $n^{-1/3}$ *but is more efficient than the original one in* (18)*.*

Due to data confidentiality agreement, we are not able to use the raw data. Here, we generate pseudo responses $\tilde{Y}_i^{(j)}$ given $X_i^{(j)}$ and $A_i^{(j)}$ from the Yahoo data, and use the dataset

$\{(X_i^{(j)}, A_i^{(j)}, \tilde{Y}_i^{(j)}) : i = 1, \ldots, n_j, j = 1, \ldots, S\}$ in our application. The generated variables $\tilde{Y}_i^{(j)}$'s are similar to the original responses $Y_i^{(j)}$'s. For example, we have $\sum_{i,j} Y_i^{(j)} / \sum_j n_i \approx 4.71\%$ while $\sum_{i,j} \tilde{Y}_i^{(j)} / \sum_j n_i \approx 4.73\%$. Besides, under our data generating process, the population limit of $\hat{\theta}^{(j)}$ in (20) can be explicitly calculated as $\theta_0 = (\theta_{0,1}, \theta_{0,2}, \theta_{0,3}, \theta_{0,4})^T = (2.534, 2.881, 2.796, 3.200)^T$ for any j. Hence, θ_0 is also the population limit of $\hat{\theta}_0$ when S does not diverge too fast. Detailed descriptions of generating $\tilde{Y}_i^{(j)}$'s are given in Section I of the supplementary appendix.

We choose $S = 2^j$ for $j = 4, 5, \ldots, 10$. Under a given S, denoted by $\hat{\theta}_0^{(S)} = (\hat{\theta}_{0,1}^{(S)}, \hat{\theta}_{0,2}^{(S)}, \hat{\theta}_{0,3}^{(S)}, \hat{\theta}_{0,4}^{(S)})^T$ the corresponding aggregated estimator. For each S, we use sample variance to estimate the variance of the aggregated estimator. Based on these estimates, we plot the estimators $\hat{\beta}_{0,i}^{(S)}$ and the Wald-type 95% confidence intervals of $\theta_{0,i}$ in Figure 4, for $i = 1, \ldots, 4$ with different choices of S.

Figure 4: 95% confidence intervals of $\theta_{0,1}$, $\theta_{0,2}$, $\theta_{0,3}$ and $\theta_{0,4}$ from top to bottom and from left to right, against $\log(S)/\log(2)$. Dash lines are the corresponding $\theta_{0,i}$'s.



It is clear from Figure 4 that the variance of $\hat{\theta}_0^{(S)}$ decreases as S increases, since the width of confidence intervals decreases as S increases. Moreover, when S is extremely large, some of the parameters are not covered in the 95% confidence intervals. For example, from the top left plot in Figure 4, $\theta_{0,1}$ is not covered in the confidence intervals of $\hat{\theta}_{0,1}^{(S)}$ when $S = 2^9$ and 2^{10}. Such phenomenon is due to the large bias of $\hat{\theta}_0^{(S)}$. These empirical results demonstrate the bias-variance trade off for the aggregated estimator, and are consistent with our theoretical findings.

5 Tail inequality for $\hat{h}^{(j)}$

In this section, we establish tail inequalities for $\hat{\theta}^{(j)}$ and $\hat{h}^{(j)}$, which are used to construct $\tilde{h}^{(j)}$, a truncated version of $\hat{h}^{(j)}$ with tail equivalence.

Theorem 5.1 *Under Conditions (A1)-(A5), for sufficiently large n_j, there exists some constant C_0, such that*

$$Pr(\hat{\theta}^{(j)} \notin N_\delta) \leq 2\exp(-C_0 n_j). \tag{21}$$

Moreover, for sufficiently large n_j, there exist some constants $C_1, C_2 > 0$ and $N_0 \geq 2$, such that

$$Pr(||\hat{h}^{(j)}||_2 \geq x | \hat{\theta}^{(j)} \in N_\delta) \leq C_2 \exp(-C_1 x^3), \tag{22}$$

for any $N_0 \leq x \leq n_j^{1/3}\delta$.

Remark 5.1 (21) *and* (22) *can be viewed as generalization of the consistency and rate of convergence results established for cube root estimators (cf. Corollary 4.2 in Kim and Pollard, 1990). The tail probability of $||\hat{h}^{(j)}||_2$ is obtained based on the subexponential tail Assumption (A3) for $m(\cdot, \theta)$.*

We represent $\hat{h}^{(j)}$ as

$$\hat{h}^{(j)} = \arg\max_{h \in H_{n_j}} M_{n_j,j}(h) \equiv \arg\max_{h \in H_{n_j}} \left\{ n_j^{1/6} \mathbb{G}_{n_j}^{(j)}(m_h^{(j)}) + n_j^{2/3} \mathrm{E}(m_h^{(j)}) \right\},$$

where $H_{n_j} = \{h \in \mathbb{R}^d : n_j^{-1/3} h + \theta_0 \in \Theta\}$, $\mathbb{G}_{n_j}^{(j)} = n_j{}^{1/2}(\mathbb{P}_{n_j}^{(j)} - \mathrm{E})$ and $m_h^{(j)}(\cdot) = m(\cdot, \theta_0 + n_j^{-1/3} h) - m(\cdot, \theta_0)$. Similarly define

$$\tilde{h}^{(j)} = \arg\max_{h \in H_{n_j} \cap H_{\delta_n}} M_{n_j,j}(h) = \arg\max_{h \in H_{n_j} \cap H_{\delta_n}} \left\{ n_j^{1/6} \mathbb{G}_n^{(j)}(m_h^{(j)}) + n_j^{2/3} \mathrm{E}(m_h^{(j)}) \right\},$$

where $H_{\delta_n} = \{h : ||h||_2 \leq \delta_n\}$. By its definition, we have $||\tilde{h}^{(j)}||_2 \leq \delta_n$. The following Corollaries are immediate applications of Theorem 5.1.

Corollary 5.1 *Assume $\delta_n \leq n_j^{1/3}\delta$. Under Conditions (A1)-(A5), for sufficiently large n_j, there exist some constants $N_0 \geq 2$, C_4 and C_5, such that*

$$Pr(||\tilde{h}^{(j)}||_2 > x) \leq C_5 \exp(-C_4 x^3), \qquad \forall x \geq N_0. \tag{23}$$

The proof is straightforward by noting that for any $x \leq n_j^{1/3}\delta$,

$$\begin{aligned}\Pr(||\tilde{h}^{(j)}||_2 > x) &\leq \Pr(||\tilde{h}^{(j)}||_2 > x|\hat{\theta}^{(j)} \in N_\delta)\Pr(\hat{\theta}^{(j)} \in N_\delta) + \Pr(\hat{\theta}^{(j)} \notin N_\delta)\\ &\leq C_2 \exp(-C_1 x^3) + 2\exp(-C_0 n_j) \leq C_5 \exp(-C_4 x^3).\end{aligned}$$

Remark 5.2 *Corollary 5.1 suggests that $\tilde{h}^{(j)}$ has finite moments of all orders. For any $a \in \mathbb{R}^d$ and positive integer k, this implies that the sequence of random variables $|a^T\tilde{h}^{(j)}|^k$ are uniformly integrable. This result is useful in establishing the convergence for moments of $\tilde{h}^{(j)}$ (see Corollary 5.3).*

Corollary 5.2 *Under Conditions (A1)-(A5) and (A8), taking $\delta_n = \max(3^{1/3}, 3^{1/3}/C_1^{1/3})\log^{1/3} n_j$ where C_1 is defined in Theorem 5.1, then $\tilde{h}^{(j)}$ and $\hat{h}^{(j)}$ are tail equivalent. If $S = o(n^3)$, then $\sum_{j=1}^S \tilde{h}^{(j)}$ and $\sum_{j=1}^S \hat{h}^{(j)}$ are also tail equivalent.*

Tail equivalence of $\tilde{h}^{(j)}$ and $\hat{h}^{(j)}$ follows by

$$\Pr\left(\tilde{h}^{(j)} \neq \hat{h}^{(j)}\right) = \Pr\left(||\hat{h}^{(j)}||_2 > \delta_n\right) \leq \frac{C_2}{n_j^3} + 2\exp(-C_0 n_j) \leq \frac{C_2\bar{c}^3}{n^3} + 2\exp\left(-\frac{C_0 n}{\bar{c}}\right), \tag{24}$$

where the first inequality is implied by Theorem 5.1 and the last inequality is due to Condition (A8). The second assertion follows by an application of Bonferroni's inequality.

Corollary 5.2 proves (7). From now on, we take $\delta_{n_j} = \max(3^{1/3}, 3^{1/3}/C_1^{1/3})\log^{1/3} n_j$. By (24), Slutsky's Theorem implies $\tilde{h}^{(j)} \xrightarrow{d} h_0$. Applying Skorohod's representation Theorem (cf. Section 9.4 in Athreya and Lahiri, 2006), we have that there exist random vectors $\tilde{h}^{(j)\star} \overset{d}{=} \tilde{h}^{(j)}$ and $h_0^\star \overset{d}{=} h_0$ such that $\tilde{h}^{(j)\star} \to h_0^\star$, almost surely. This together with the uniform integrability of $||\tilde{h}^{(j)}||_2^k$ gives the following Corollary.

Corollary 5.3 *Under Conditions (A1)-(A5), for any $a \in \mathbb{R}^d$ and integer $k \geq 1$, we have $E\{(a^T\tilde{h}^{(j)})^k\} \rightarrow E\{(a^T h_0)^k\}$ as $n_j \rightarrow \infty$.*

Remark 5.3 *Due to the i.i.d assumption of $X_i^{(j)}$, $E\{(a^T\tilde{h}^{(j)})^k\}$ is a function of n_j only. Under Condition (A8), Corollary* (5.3) *implies*

$$\sup_j |E\{(a^T\tilde{h}^{(j)})^k\} - E\{(a^T h_0)^k\}| \rightarrow 0, \quad as \ \ n \rightarrow \infty.$$

Taking $k = 2$, it proves (9)*. Taking $k = 3$, it proves* (10)*. Moreover, Corollary 5.3 suggests a simple scheme for estimating the covariance matrix $A \equiv cov(h_0)$ given in* (5)*. For any vector a, by law of large numbers, we obtain*

$$\frac{1}{S}\sum_{j=1}^{S}(a^T\tilde{h}^{(j)})^2 - \frac{1}{S}\sum_{j=1}^{S}E(a^T\tilde{h}^{(j)})^2 \overset{a.s.}{\rightarrow} 0.$$

This together with tail equivalence between $\tilde{h}^{(j)}$ and $\hat{h}^{(j)}$, and (9) *implies that $\sum_j (a^T\hat{h}^{(j)})^2/S$ converges to $a^T Aa$.*

6 Discussion

In this paper, we provide a general inference framework for aggregated M-estimators with cubic rates obtained by the divide and conquer method. Our results demonstrate that the aggregated estimators have faster convergence rate than the original M-estimators based on pooled data and achieve the asymptotic normality when the number of groups S does not grow too fast with respect to n, the average sample size of each group.

6.1 Rate of the bias

For a general cubic-rate estimator with sample size n, we showed its bias can be bounded by $O((n/\log n)^{-5/12})$. In comparison, Banerjee et al. (2016) obtained a sharper bound in the specific setting of monotone regression and showed that the bias of the isotonic estimator can be bounded by $O(n^{-7/15+\zeta})$ for any $\zeta > 0$ and the bias of its inverse bounded by $o(n^{-1/2})$ (see Theorem 4.3 and 4.4 in that paper). As commented before, this is because we work on a more general setting and their techniques cannot be easily generalized to other cubic-rate M-estimation problems.

However, it is possible to sharpen the bound for some special cases. In particular, when the parameter is one-dimensional, we show in Theorem A.1 (see also Corollary A.1) in the supplementary appendix that the bias of the estimator can be bounded by $O(n^{-5/9}\log^{9/14} n)$ based on the KMT approximation. Note that this bound is even sharper than those in Theorem 4.3 and 4.4 in Banerjee et al. (2016). This is because we assume a stronger assumption on the Lipschitz continuity of the Hessian matrix (see Assumption A6 and Equation (4.3) in Banerjee et al., 2016). Under Assumption A8, Theorem A.1 implies the asymptotic normality holds for the aggregated estimator as long as the number of machines satisfies $S = O(N^l)$ for some $l < 4/13$, where N is the total number of observations. Again this upper bound on S may still be conservative, however it improves a lot compared to Theorem 2.1. We further apply our theorem to the location estimator (see Section A.5.1) and the one-dimensional value search estimator (see Section A.5.2) for illustration.

For the bias of a general cubic-rate M-estimator, our proof relies on the Gaussian approximation of the suprema of empirical processes (cf. Chernozhukov et al., 2013, 2014) and the Sudakov-Fernique type error bound (Chatterjee, 2005). The proofs for these theorems are based on smooth approximation of the supremum function. It remains unknown whether the rates of these error bounds are optimal and whether they can be improved using other techniques. This is an interesting problem that needs further investigation.

6.2 The super-efficiency phenomenon

In the context of isotonic regression, Banerjee et al. (2016) showed that the faster convergence rate of the aggregated estimator of the inverse function for a fixed model comes at a price, that is, the maximal risk over a class of models in a neighborhood of the given model remains bounded for the pooled estimator but diverges to infinity for the aggregated estimator (see Theorem 6.1 in Banerjee et al., 2016). This is referred to as the super-efficiency phenomenon, which is seen in nonparametric function estimation as well (cf. Brown et al., 1997).

We believe such super-efficiency phenomenon holds for many other cubic-rate M-estimation problems as well. In the supplementary appendix, we mathematically formalize the notion of the super-efficiency phenomenon for general M-estimation problems, and establish such phenomenon for the location estimator (see Section B.1) and the value search estimator (see Section B.2). The super-efficiency phenomenon is essentially due to that the maximal bias over a large class of models for the aggregated estimator will diverge to infinity. We suspect this is because the condition on the Lipschitz continuity of the Hessian matrix (Assumption A6) cannot hold uniformly for all models in such a class. We discuss this in details in the supplementary appendix.

6.3 Other issues

In the current setup, we assume all $X_i^{(j)}$'s are independently and identically distributed. It will be interesting to generalize Theorem 2.1 to the setting where $X_i^{(j)}$'s are independent, but not identically distributed. However, the meaning of the aggregated estimator may become unclear in some applications, such as the value search estimator, and the derivation of the asymptotic properties of the resulting aggregated estimator becomes much more involved. This needs further investigation.

References

Athreya, K. B. and S. N. Lahiri (2006). *Measure theory and probability theory.* Springer Texts in Statistics. Springer, New York.

Banerjee, M., C. Durot, and B. Sen (2016). Divide and conquer in non-standard problems and the super-efficiency phenomenon. *arXiv preprint arXiv:1605.04446*.

Brown, L. D., M. G. Low, and L. H. Zhao (1997). Superefficiency in nonparametric function estimation. *Ann. Statist. 25*(6), 2607–2625.

Chatterjee, S. (2005). An error bound in the sudakov-fernique inequality. *arXiv preprint math/0510424*.

Chen, X. and M.-g. Xie (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica 24*(4), 1655–1684.

Chernoff, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math. 16*, 31–41.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist. 41*(6), 2786–2819.

Chernozhukov, V., D. Chetverikov, and K. Kato (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist. 42*(4), 1564–1597.

Csörgő, M., S. Csörgő, L. Horváth, and P. Révész (1985). On weak and strong approximations of the quantile process. In *Proceedings of the seventh conference on probability theory (Braşov, 1982)*, pp. 81–95. VNU Sci. Press, Utrecht.

Durot, C. (2002). Sharp asymptotics for isotonic regression. *Probab. Theory Related Fields 122*(2), 222–240.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Ann. Statist. 22*(4), 1993–2010.

Groeneboom, P., G. Hooghiemstra, and H. P. Lopuhaä (1999). Asymptotic normality of the L_1 error of the Grenander estimator. *Ann. Statist. 27*(4), 1316–1347.

Kim, J. and D. Pollard (1990). Cube root asymptotics. *Ann. Statist. 18*(1), 191–219.

Kleiner, A., A. Talwalkar, P. Sarkar, and M. I. Jordan (2014). A scalable bootstrap for massive data. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 76*(4), 795–816.

Koltchinskii, V. I. (1994). Komlos-Major-Tusnady approximation for the general empirical process and Haar expansions of classes of functions. *J. Theoret. Probab. 7*(1), 73–118.

Komlós, J., P. Major, and G. Tusnády (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete 32*, 111–131.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer Series in Statistics. Springer, New York.

Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields 98*(1), 21–45.

Sakhanenko, A. I. (2006). Estimates in the invariance principle in terms of truncated power moments. *Sibirsk. Mat. Zh. 47*(6), 1355–1371.

Splawa-Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statist. Sci. 5*(4), 465–472.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2012). A robust method for estimating optimal treatment regimes. *Biometrics 68*(4), 1010–1018.

Zhang, Y., J. Duchi, and M. Wainwright (2013). Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pp. 592–617.

Zhao, T., G. Cheng, and H. Liu (2016). A partially linear framework for massive heterogeneous data. *Ann. Statist. 44*(4), 1400–1437.