arXiv:2105.14577v1 [math.ST] 30 May 2021

The HulC: Confidence Regions from Convex Hulls

Arun Kumar Kuchibhotla, Sivaraman Balakrishnan, and Larry Wasserman

{arunku, siva, larry}@stat.cmu.edu

Department of Statistics & Data Science, Carnegie Mellon University

Abstract

We develop and analyze the HulC, an intuitive and general method for constructing confidence sets using the convex hull of estimates constructed from subsets of the data. Unlike classical methods which are based on estimating the (limiting) distribution of an estimator, the HulC is often simpler to use and effectively bypasses this step. In comparison to the bootstrap, the HulC requires fewer regularity conditions and succeeds in many examples where the bootstrap provably fails. Unlike subsampling, the HulC does not require knowledge of the rate of convergence of the estimators on which it is based. The validity of the HulC requires knowledge of the (asymptotic) median-bias of the estimators. We further analyze a variant of our basic method, called the Adaptive HulC, which is fully data-driven and estimates the median-bias using subsampling. We show that the Adaptive HulC retains the aforementioned strengths of the HulC. In certain cases where the underlying estimators are pathologically asymmetric the HulC and Adaptive HulC can fail to provide useful confidence sets. We propose a final variant, the Unimodal HulC, which can salvage the situation in cases where the distribution of the underlying estimator is (asymptotically) unimodal. We discuss these methods in the context of several challenging inferential problems which arise in parametric, semi-parametric, and non-parametric inference. Although our focus is on validity under weak regularity conditions, we also provide some general results on the width of the HulC confidence sets, showing that in many cases the HulC confidence sets have near-optimal width.

1 Introduction

Estimation and uncertainty quantification are two of the most fundamental aspects of statistical analysis. The theory of point estimation is very well-studied starting from the principle of maximum likelihood estimation (Stigler, 2007; Pfanzagl, 2011; Lehmann and Casella, 2006). Relatively more recent frameworks of parametric efficiency (van der Vaart, 2000, Chapters 4–8) and semiparametric influence functions (Bickel et al., 1993) provide general methods of constructing "good" estimators. Uncertainty quantification, for instance testing a statistical hypothesis or constructing a confidence set, most often follows from studying the asymptotic distribution of the estimator. In many cases this approach requires *estimating* the asymptotic distribution of the estimator. Even in favorable cases, when this asymptotic distribution is mean zero Gaussian, one needs to further estimate the asymptotic variance of the estimator in order to construct a valid confidence set. As a consequence, in practice, methods which yield uncertainty quantification while using only a method for point estimation are often favored.

Generic techniques to obtain uncertainty quantification that do not require any more than the estimation method are bootstrap and subsampling (Efron, 1979; Politis and Romano, 1994; Shao and Tu, 2012; Hall, 2013). The bootstrap however requires that the estimator be Hadamard differentiable; see Dümbgen (1993) and Fang and Santos (2019, Section 3.2). Subsampling is more general, but requires knowing the rate of convergence of the estimator. Bertail et al. (1999) provides a scheme to estimate the unknown rate of convergence, but this method is hard to implement and cannot estimate the slowly varying components of the rate (such as $\log n$ factors); see Sherman and Carlstein (1997, page 3) for details.

In this paper, we propose a new method, the HulC (Hull based Confidence) that does not require variance estimation and is applicable in many examples where the bootstrap and subsampling are not. The HulC does not require knowing the rate of convergence of the estimator. In many cases, the HulC does not involve any tuning parameters. Besides being asymptotically valid, the HulC is *eventually finite sample* (EFS) meaning that the coverage is exact for all samples $n \geqslant n_0$ for some finite n_0.

The basis for the HulC is an assumption that the estimators on which it is based are not pathologically asymmetric: their distributions do not place all their mass to one side of the target parameter. We measure the asymmetry in terms of the median bias of the estimator. This makes the method widely applicable and easy to use. Our method has some similarity to the typical values approach of Hartigan (1969, 1970). See, in particular, point 5 in Section 7 of Hartigan (1970).

Mean unbiasedness is a popular criterion for "good" estimators and mean bias reduction is well-studied in the statistics literature (Firth, 1993; Kosmidis and Firth, 2009; Kim, 2016). However, as noted in (Pfanzagl, 2017) the fact that an estimator is mean unbiased does not naturally aid in uncertainty quantification. In contrast, median unbiasedness implies that the estimator is equally likely to underestimate and overestimate the target of interest. As will be shown in this article, this property can lead to a simple method for constructing confidence intervals. Median unbiasedness and median bias reduction are not as widely known as the mean unbiasedness and mean bias reduction, but we will develop their implications for inference. We refer the reader to Pfanzagl (2011) for details regarding median unbiased estimation and to Kenne Pagui et al. (2017); Kosmidis et al. (2020) for median bias reduction methods.

Inspired by the practical success of resampling methods like the bootstrap and subsampling, the HulC directly exploits our relatively strong understanding of point estimation to address challenging inferential problems. As with these methods, the width of the intervals we construct are naturally related to the accuracy of the underlying estimators, i.e. the HulC based on a very accurate estimator will lead to small confidence sets. On the other hand, in contrast to these methods the HulC uses sample-splitting to avoid strong regularity conditions, and its validity relies instead on a relatively mild assumption. This follows a line of recent work by the authors (for instance, Wasserman et al. (2020); Chakravarti et al. (2019); Rinaldo et al. (2019)), and more classical work Bickel (1982), where sample-splitting eases the challenges of statistical inference, often at a surprisingly small price.

The remainder of this article is organized as follows. In Section 2, we describe our assumptions and the HulC method for constructing confidence regions for univariate and multivariate parameters. We compare the proposed confidence interval to Wald confidence intervals based on asymptotic Normality in terms of their widths. We also compare to the bootstrap and subsampling in terms of applicability. In Section 3, we discuss the applicability of the HulC to some standard examples where limiting distributions are well-understood but constructing valid confidence sets can still be challenging; the examples we consider include mean and median estimation, Binomial proportion estimation, and parameter estimation in exponential families. Our method involves an assumption on the median bias of the estimators under consideration. In Section 4, we

describe the ADAPTIVE HULC which estimates the median bias using subsampling. Interestingly, in contrast to directly using subsampling for constructing a confidence set, the ADAPTIVE HULC does not require knowledge of the rate of convergence. In Section 5, we provide some applications of the ADAPTIVE HULC to nonparametric models including shape constrained regression. In Section 6, we provide an extension, called the UNIMODAL HULC, based on the assumption of unimodality. Between our asymmetry assumption and unimodality assumption, we believe that many challenging confidence set construction problems based on independent observations are solved. Finally, in Section 7, we summarize the article and discuss some future directions. Throughout the article, we focus on the pointwise validity (as in Politis and Romano (1994)) of our confidence region and uniform validity will be the focus of a companion article.

The proofs of all the main results are provided in the supplementary material. The sections and equations of the supplementary file are prefixed with "S." and "E.", respectively, for convenience. We provide the code to reproduce the figures in the paper, including an implementation of our methods in R together with Jupyter notebooks illustrating their application at `https://github.com/Arun-Kuchibhotla/HulC`.

2 The HulC: Hull based Confidence Regions

In this section, we describe the HULC and compare it to classical asymptotic Normality based confidence intervals. We present several results for the HULC, and in order to aid readability we provide a brief roadmap here:

1. Focusing first on univariate parameters, in Theorem 1, we show that when the median bias of the estimators is known to be at most Δ the HULC (as described in Algorithm 1) has guaranteed coverage of at least $1-\alpha$. We also show that, under some mild additional conditions, if the underlying estimators have median bias exactly Δ then the HULC has coverage exactly $1-\alpha$.

2. In Proposition 1, we investigate properties of a (slightly) conservative variant of the HULC, showing that the HULC when provided with the *asymptotic* median bias still ensures finite-sample $1-\alpha$ coverage, for sufficiently large sample sizes. This setting is practically useful because in many cases we know the limiting distribution of our estimates is Normal (say) and in these cases the asymptotic median bias is known to be 0.

3. In Theorem 2 and Remark 2.2, we show that the guarantees of the (non-conservative) HULC erode gracefully, i.e. if we run the HULC with a parameter Δ but the true median bias is at most $\widetilde{\Delta}$ then the HULC has coverage which degrades from the nominal level (multiplicatively) as a function of $|\Delta - \widetilde{\Delta}|$.

4. In (23) and (25), we provide two simple analyses of the width of the HULC intervals. In (23) we show that in the classical setting where the estimates have an asymptotic Normal distribution, the width of the HULC interval is the same as that of the corresponding Wald interval upto a slowly growing factor of $\sqrt{\log_2(\log_2(2/\alpha))}$. In (25), we show that under much more generality the HULC based on B^* splits yields a variance-sensitive confidence interval whose expected width is upper bounded by $2\sigma B^*/\sqrt{n}$ where σ is the standard deviation of the estimators on which the HULC is based.

5. Turning our attention to multivariate parameters, in Lemma 2, we analyze the HULC based on the convex hull of the underlying estimates and on the rectangular hull of the underlying estimates, providing coverage guarantees as a function of the median bias. We further compare the multivariate HULC intervals to those based on multivariate CLTs, highlighting several advantages of the former.

2.1 HulC for univariate parameters

Suppose $\theta_0 \in \mathbb{R}$ is a parameter or functional of interest. Let $X_1, \ldots, X_n$ be independent random variables from some measurable space $\mathcal{X}$. For $B \geqslant 1$, let $\widehat{\theta}_1, \ldots, \widehat{\theta}_B$ be independent estimators of θ_0. These can be obtained by splitting the data $X_1, \ldots, X_n$ into B batches and computing an estimate from each batch. Define the median bias of the estimator $\widehat{\theta}_j$ for θ_0 as

$$\text{Med-Bias}_{\theta_0}(\widehat{\theta}_j) \;:=\; \left(\frac{1}{2} - \min\left\{\mathbb{P}(\widehat{\theta}_j - \theta_0 \geqslant 0), \mathbb{P}(\widehat{\theta}_j - \theta_0 \leqslant 0)\right\}\right)_+, \tag{1}$$

where $(x)_+ = \max\{x, 0\}$ for any $x \in \mathbb{R}$. Using the independence of the estimators $\widehat{\theta}_j, 1 \leqslant j \leqslant B$, we obtain the following result (proved in Section S.2 of the supplementary file).

Lemma 1. *If $\widehat{\theta}_j, 1 \leqslant j \leqslant B$ are independent random variables and*

$$\Delta \;:=\; \max_{1 \leqslant j \leqslant B} \text{Med-Bias}_{\theta_0}(\widehat{\theta}_j), \tag{2}$$

then

$$\mathbb{P}\left(\theta_0 \notin \left[\min_{1 \leqslant j \leqslant B} \widehat{\theta}_j, \max_{1 \leqslant j \leqslant B} \widehat{\theta}_j\right]\right) \leqslant \left(\frac{1}{2} - \Delta\right)^B + \left(\frac{1}{2} + \Delta\right)^B.$$

An estimator $\widehat{\theta}$ is said to **median unbiased** for θ_0 if $\text{Med-Bias}_{\theta_0}(\widehat{\theta}) = 0$ (Pfanzagl, 2011). It is worth noting that median unbiasedness does not imply that the estimator is symmetric. The non-strict inequality in the definition (1) is important: it allows for $\mathbb{P}(\widehat{\theta}_j - \theta_0 \geqslant 0)$ and $\mathbb{P}(\widehat{\theta}_j - \theta_0 \leqslant 0)$ to be equal to 1 or be larger than 1/2. This is useful in cases where θ_0 is on the boundary or $\widehat{\theta}_j$ has a discrete distribution and puts non-zero mass at θ_0. An estimator $\widehat{\theta}_n$ based on n observations is **asymptotically median unbiased** if $\lim_{n\to\infty} \text{Med-Bias}_{\theta_0}(\widehat{\theta}_n) \to 0$.

For any $B \geqslant 1$ and $\Delta \geqslant 0$, set the miscoverage probability from Lemma 1 as

$$P(B; \Delta) \;:=\; \left(\frac{1}{2} - \Delta\right)^B + \left(\frac{1}{2} + \Delta\right)^B. \tag{3}$$

If $\Delta \geqslant 0$ is known, then choosing $B := B_{\alpha,\Delta} \geqslant 1$ such that $P(B; \Delta) \leqslant \alpha$, we conclude that

$$\mathbb{P}\left(\theta_0 \notin \left[\min_{1 \leqslant j \leqslant B} \widehat{\theta}_j, \max_{1 \leqslant j \leqslant B} \widehat{\theta}_j\right]\right) \leqslant \alpha.$$

In words, the smallest rectangle containing $B_{\alpha,\Delta}$ independent estimators of θ_0 has a coverage of at least $1 - \alpha$. Because B is an integer, $P(B; \Delta)$ decreases in steps as B changes over positive integers and this can lead to conservative coverage. This issue can be resolved easily by randomizing the choice of B. Most often Δ is unknown. This issue will be resolved in Section 4 where we show how to estimate Δ.

Algorithm 1 gives the steps to find a randomized confidence interval with $1-\alpha$ coverage when the median bias Δ is known.

There are no restrictions on the input $\mathcal{A}(\cdot)$ in Algorithm 1 except that it produces an estimate with median bias bounded by Δ. Its rate of convergence and variance play a role only in the width properties of the resulting confidence interval, not in the validity guarantee. A better estimator will lead to a smaller confidence interval. Here are two examples of the estimation procedure $\mathcal{A}(\cdot)$:

- If $X_1, \ldots, X_n$ are identically distributed and $\theta_0 = \mathbb{E}[X_1]$, then one can take $\widehat{\theta}_j = \mathcal{A}(\{X_i : i \in S_j\}) = |S_j|^{-1} \sum_{i \in S_j} X_i$. In general, the median bias of the sample mean is unknown, but typically tends to zero as $|S_j| \to \infty$. If the observations are symmetrically distributed around θ_0, then $\text{Med-Bias}_{\theta_0}(\widehat{\theta}_j) = 0$.

Algorithm 1: Confidence Interval with Known Δ (HulC)

Input: data $X_1, \ldots, X_n$, coverage probability $1-\alpha$, a value Δ, and an estimation procedure $\mathcal{A}(\cdot)$ that takes as input observations and returns an estimator with a median bias of at most Δ.

Output: A confidence interval $\widehat{\mathrm{CI}}_{\alpha,\Delta}$ such that $\mathbb{P}(\theta_0 \in \widehat{\mathrm{CI}}_{\alpha,\Delta}) \geqslant 1-\alpha$.

1 Find the smallest integer $B = B_{\alpha,\Delta} \geqslant 1$ such that $P(B;\Delta) \leqslant \alpha$. Recall $P(B;\Delta)$ from (3).

2 Generate a random variable U from the Uniform distribution on $[0,1]$ and set

$$\tau_{\alpha,\Delta} := \frac{\alpha - P(B_{\alpha,\Delta};\Delta)}{P(B_{\alpha,\Delta}-1;\Delta) - P(B_{\alpha,\Delta};\Delta)} \quad \text{and} \quad B^* := \begin{cases} B_{\alpha,\Delta}-1, & \text{if } U \leqslant \tau_{\alpha,\Delta}, \\ B_{\alpha,\Delta}, & \text{if } U > \tau_{\alpha,\Delta}. \end{cases} \tag{4}$$

3 Randomly split the data $X_1, \ldots, X_n$ into B^* disjoint sets $\{\{X_i : i \in S_j\} : 1 \leqslant j \leqslant B^*\}$. These need not be equal sized sets, but having approximately equal sizes yields good width properties.

4 Compute estimators $\widehat{\theta}_j := \mathcal{A}(\{X_i : i \in S_j\})$, for $1 \leqslant j \leqslant B^*$

5 **return** the confidence interval

$$\widehat{\mathrm{CI}}_{\alpha,\Delta} := \left[\min_{1\leqslant j\leqslant B^*} \widehat{\theta}_j, \max_{1\leqslant j\leqslant B^*} \widehat{\theta}_j\right].$$

- If $X_1, \ldots, X_n$ are random variables generated from a parametric model p_{θ_0} that belongs to the parametric family $\{p_\theta : \theta \in \Theta\}$, then one can take $\widehat{\theta}_j = \mathcal{A}(\{X_i : i \in S_j\})$ as the maximum likelihood estimator (MLE) of θ_0 based on the observations $X_i, i \in S_j$. Under standard regularity conditions, MLE has an asymptotic Normal distribution and hence the median bias of $\widehat{\theta}_j$ converges to zero.

The following result (proved in Section S.3 of the supplementary file) proves that the confidence interval from Algorithm 1 has a coverage of at least $1-\alpha$.

Theorem 1. *If $X_1, \ldots, X_n$ are independent random variables and the estimation procedure $\mathcal{A}(\cdot)$ in Algorithm 1 returns estimates that have a median bias of at most Δ, then the confidence interval $\widehat{\mathrm{CI}}_{\alpha,\Delta}$ returned by Algorithm 1 satisfies*

$$\mathbb{P}\left(\theta_0 \in \widehat{\mathrm{CI}}_{\alpha,\Delta}\right) \geqslant 1-\alpha. \tag{5}$$

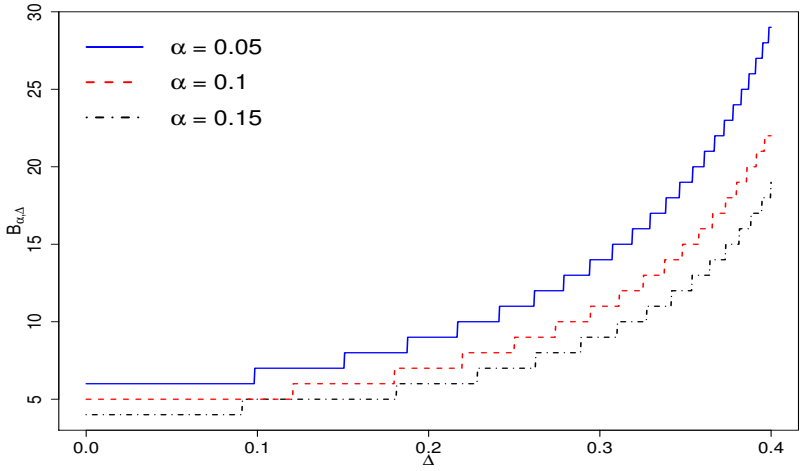
Further, if $\mathbb{P}(\widehat{\theta}_j = \theta_0) = 0$ for all j and the estimation procedure $\mathcal{A}(\cdot)$ in Algorithm 1 returns estimates that have a median bias of exactly Δ, then

$$\mathbb{P}\left(\theta_0 \in \widehat{\mathrm{CI}}_{\alpha,\Delta}\right) = 1-\alpha. \tag{6}$$

In Algorithm 1, it is implicitly assumed that B^* defined in step 2 is smaller than the sample size n so that the estimation procedure can be applied on B^* splits of the data. Recall $P(B;\Delta)$ from (3) and that $B_{\alpha,\Delta}$ is the smallest integer such that $P(B;\Delta) \leqslant \alpha$. It is easy to prove that $P(B;\Delta)$ is an increasing function of $\Delta \in [0, 1/2]$ and hence we obtain that $\max\{(1/2+\Delta)^B, 2^{-B+1}\} \leqslant P(B;\Delta) \leqslant 2(1/2+\Delta)^B$. Therefore, $B_{\alpha,\Delta}$ satisfies

$$\max\left\{\left\lceil\frac{\log(1/\alpha)}{\log(2/(1+2\Delta))}\right\rceil, \left\lceil\frac{\log(2/\alpha)}{\log(2)}\right\rceil\right\} \leqslant B_{\alpha,\Delta} \leqslant \left\lceil\frac{\log(2/\alpha)}{\log(2/(1+2\Delta))}\right\rceil. \tag{7}$$

Here $\lceil x\rceil$, for any real x, denotes the smallest integer larger than x. It is easy to verify that, $B_{\alpha,\Delta} \to \infty$ as $\Delta \to 0.5$. Figure 1 shows the plot of $B_{\alpha,\Delta}$ as Δ varies from 0 to 0.4 and α varies between $0.05, 0.1, 0.15$.



Δ \ α	0.15	0.1	0.05
0	4	5	6
0.05	4	5	6
0.1	5	5	7
0.15	5	6	7
0.2	6	7	9
0.25	7	9	11
0.3	9	11	14
0.35	12	15	19
0.4	19	22	29

Figure 1: Some example values of $B_{\alpha,\Delta}$ for different values of $\alpha \in \{0.05, 0.1, 0.15\}$ and $\Delta \in [0, 0.4]$. Left panel: the plot as Δ changes continuously. Right panel: values of $B_{\alpha,\Delta}$ as Δ changes from 0.0 to 0.4 in increments of 0.05.

2.2 HulC when asymptotic median bias is known

Algorithm 1 requires knowledge of the median bias of the estimators. In some settings, estimation procedures can be constructed so as to ensure median unbiasedness (i.e., $\Delta = 0$). When $\Delta = 0$, $B_{\alpha,0} = \lceil \log_2(2/\alpha) \rceil$. These examples are discussed in Section 3.

Because $B_{\alpha,\Delta}$ is a piecewise constant function in Δ, we do not need to know Δ exactly. This observation implies that for estimators that are asymptotically symmetric around θ_0, one can take Δ to be zero in Algorithm 1 and still retain (asymptotic) validity. Formally, if $\text{Med-Bias}_{\theta_0}(\widehat{\theta}_j) \to 0$ as $|S_j| \to \infty$, then the convex hull of $B_{\alpha,0} = \lceil \log_2(2/\alpha) \rceil$ estimators has an asymptotic coverage of at least $1 - \alpha$. Furthermore, the convex hull is *eventually finite sample* valid, meaning that there is a sample size n_0 such that the coverage is at least $1 - \alpha$ for all $n \geqslant n_0$. Now, we provide more details.

Proposition 1 proved in Section S.4 of the supplementary file formally proves that $B_{\alpha,\Delta}$ is a piecewise constant function of Δ (as illustrated in Figure 1).

Proposition 1. *For* $\widetilde{\Delta}, \Delta \in [0, 1/2)$ *and* $\alpha \in (0, 1)$, *if*

$$2B_{\alpha,\Delta}|\Delta - \widetilde{\Delta}| \;\leqslant\; B_{\alpha,\Delta}\left[\min\left\{\left(\frac{\alpha}{P(B_{\alpha,\Delta}; \Delta)}\right)^{1/B_{\alpha,\Delta}}, \left(\frac{P(B_{\alpha,\Delta} - 1; \Delta)}{\alpha}\right)^{1/B_{\alpha,\Delta}}\right\} - 1\right]. \tag{8}$$

then $B_{\alpha,\widetilde{\Delta}} = B_{\alpha,\Delta}$. *Moreover, if*

$$\frac{B_{\alpha,0}(B_{\alpha,0} - 1)}{2}\widetilde{\Delta}^2 \;\leqslant\; \frac{\alpha}{P(B_{\alpha,0}; 0)} - 1. \tag{9}$$

then $B_{\alpha,\widetilde{\Delta}} = B_{\alpha,0}$.

Remark 2.1 Note that the right hand side of (8) is non-zero if and only if $\tau_{\alpha,\Delta} \neq 0$ in (4). In a typical application, one would take Δ as the hypothesized (or asymptotic) value of the median bias and $\widetilde{\Delta}$ is the true

median bias. Hence, the right hand side of (8) can be computed exactly for any user choice of $\alpha \in (0, 1)$. As a practical matter, the user can change α by a tiny amount to increase the right hand side of (8). Figure 2 shows the behavior of the right hand side of (8). In the most common setting of asymptotic Normality, $\Delta = 0$, and consequently the requirement becomes more relaxed as in (9); this relaxation stems from the fact that $\Delta \mapsto P(B; \Delta)$ has zero first derivative at $\Delta = 0$. The right hand side of (9) is shown in Figure 3. $\diamond$

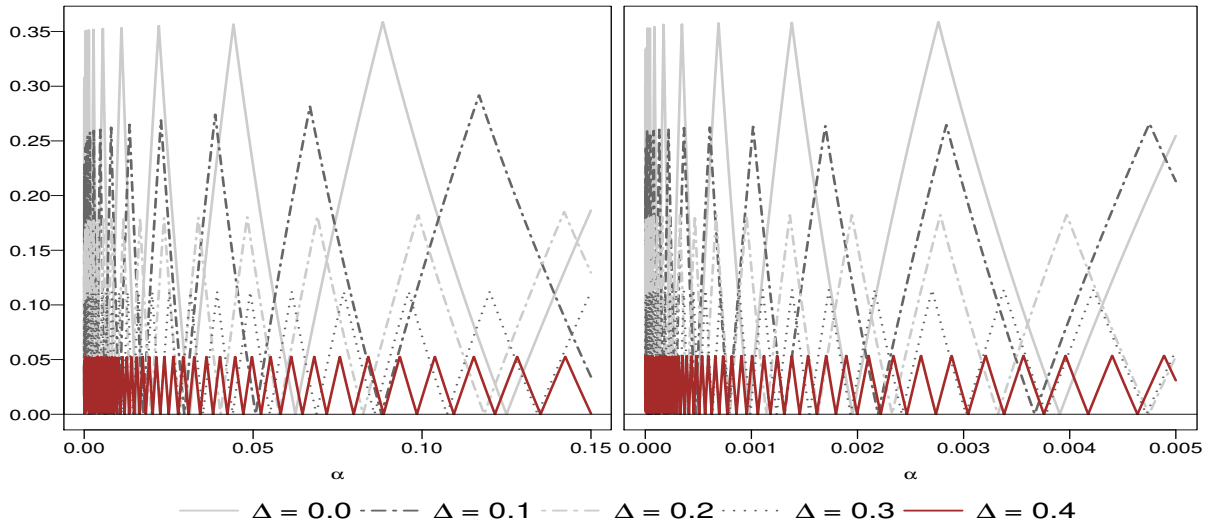


Figure 2: The plots show the right hand side of (8) on the y-axis as α changes from 0 to 0.15 and $\Delta \in \{0, 0.1, 0.2, 0.3, 0.4\}$. In the left panel, we show the plot for $\alpha \in (0, 0.15)$ and in the right panel, we show the plot for $\alpha \in (0, 0.005)$. The y-axis limits remain the same for both plots.

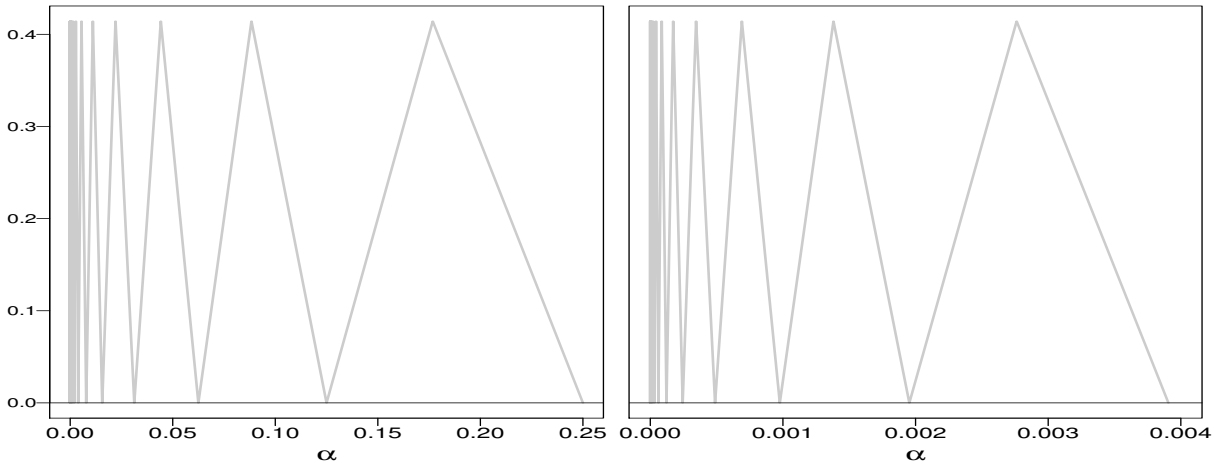


Figure 3: The plots show the right hand side of (9) on the y-axis as α changes from 0 to 0.15. In the left panel, we show the plot for $\alpha \in (0, 0.15)$ and in the right panel, we show the plot for $\alpha \in (0, 0.004)$. The y-axis limits remain the same for both plots.

Recall from the calculation surrounding (3) that the smallest interval containing $B_{\alpha,\tilde{\Delta}}$ estimators has a coverage of at least $1-\alpha$, if the estimators have a median bias of at most $\tilde{\Delta}$. Proposition 1 implies that one need not know the median bias $\tilde{\Delta}$ of the estimators exactly in order to find $B_{\alpha,\tilde{\Delta}}$. Suppose the estimators $\widehat{\theta}_j, j \geqslant 1$ have a known asymptotic median bias of Δ. Recall $\tau_{\alpha,\Delta}$ defined in (4). Proposition 1 implies that for every $\alpha \in (0,1)$ satisfying $\tau_{\alpha,\Delta} \neq 0$ there exists $N_\alpha \geqslant 1$ such that for all $n \geqslant N_\alpha$,

$$\mathbb{P}\left(\theta_0 \in \left[\min_{1 \leqslant j \leqslant B_{\alpha,\Delta}} \widehat{\theta}_j, \max_{1 \leqslant j \leqslant B_{\alpha,\Delta}} \widehat{\theta}_j\right]\right) \geqslant 1-\alpha. \tag{10}$$

Inequality (10) is obvious from Lemma 1 with $B_{\alpha,\tilde{\Delta}}$ estimators. Proposition 1 along with asymptotic median bias of Δ implies that $B_{\alpha,\tilde{\Delta}} = B_{\alpha,\Delta}$ for $n \geqslant N_\alpha$. The threshold sample size N_α depends on how fast $|\tilde{\Delta}-\Delta|$ converges to zero and how big the right hand side of (8), (9) are. The coverage guarantee (10) can be compared to the coverage guarantee for Wald, bootstrap, and subsampling intervals. None of these intervals have a guarantee of at least $1-\alpha$ coverage even for large sample sizes; the coverage only converges to $1-\alpha$ with sample size.

In most cases including parametric and semiparametric models, Berry–Esseen type bounds are available that provide bounds of the form,

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\frac{|S_j|^{1/2}(\widehat{\theta}_j - \theta_0)}{\sigma} \leqslant t \right) - \Phi(t) \right| \;\leqslant\; \frac{\mathfrak{C}_X}{|S_j|^{1/2}}, \tag{11}$$

where $|S_j|$ is the number of observations in the j-th split of the sample based on which $\widehat{\theta}_j$ is computed. Here $\mathfrak{C}_X$ is a constant that depends on the true distribution of the data. (If $\widehat{\theta}_j$ is the sample mean, then $\mathfrak{C}_X$ can be bounded in terms of the skewness of the random variables $X_i, i \geqslant 1$.) For results of this type, see Pfanzagl (1971, 1973a), Bentkus et al. (1997); Bentkus (2005), and Pinelis (2017). (In semi/non-parametric models as in Zhang and Liang (2011) and Han and Kato (2019), the rate of convergence may be slower than $|S_j|^{-1/2}$.) In this case, assuming $|S_j| \asymp \sqrt{n/B_{\alpha,0}}$ (i.e., data is split approximately equally into B^* many samples), we get that,

$$\tilde{\Delta} \;\leqslant\; \max_{1 \leqslant j \leqslant B_{\alpha,0}} \left| \mathbb{P}(\widehat{\theta}_j - \theta_0 \leqslant 0) - \frac{1}{2} \right| \;\leqslant\; \mathfrak{C}_X \sqrt{\frac{B_{\alpha,0}}{n}}. \tag{12}$$

Note that this conclusion requires a weaker bound than the one in (11) because we only care about $t = 0$ in (11). For example, in case of the sample mean, if the observations are symmetric around the population mean, then $\tilde{\Delta} = 0$ irrespective of any moment assumptions, but a general Berry–Esseen bound (11) need not hold true without additional moment assumptions. If we take $\mathcal{M}$ to be the set of all strictly increasing functions and $\mathcal{S}$ is the class of all continuous distributions F with $F(0) = 1/2$, then $\tilde{\Delta}$ can also be bounded as

$$\tilde{\Delta} \leqslant \max_{1 \leqslant j \leqslant B_{\alpha,0}} \inf_{h \in \mathcal{M}} \inf_{F \in \mathcal{S}} \left| \mathbb{P}\left(h(\widehat{\theta}_j) - h(\theta_0) \leqslant 0 \right) - F(0) \right|. \tag{13}$$

This follows from the fact that $\{\widehat{\theta}_j \leqslant \theta_0\} = \{h(\widehat{\theta}_j) \leqslant h(\theta_0)\}$ for all strictly increasing functions h. Allowing for arbitrary increasing transformations may result in better Normal approximations in many cases. Classical examples include the Fisher's z-transformation for the correlation coefficient and Anscombe's arcsine transformation for Binomial random variable; see Borges (1970); Gebhardt (1969); Borges (1971); Efron (1982) for some examples. Because symmetric distributions belong to $\mathcal{S}$ and the standard Normal distribution belongs to it, the right hand side of (13) is always better (i.e., smaller) than the bound attained by (11). Moreover, if $\widehat{\theta}_j$ has zero median, then the right hand side of (13) is zero but (11) can result in a constant order upper bound.

If inequality (12) holds true, then the requirement (9) holds true if

$$\mathfrak{C}_X^2 \frac{B_{\alpha,0}^3}{2n} \leqslant \frac{\alpha}{P(B_{\alpha,0};0)} - 1 \quad \Leftrightarrow \quad \mathfrak{C}'_X \frac{(\log(2/\alpha))^3}{n} \leqslant \frac{\alpha}{P(B_{\alpha,0};0)} - 1, \tag{14}$$

where $\mathfrak{C}'_X$ is a slightly adjusted version of the constant $\mathfrak{C}_X^2$. This equivalence follows from inequality (7) for $B_{\alpha,0}$. From Figure 3, it is clear that $\alpha/P(B_{\alpha,0};0) - 1$ in (9) can be as large as 0.4 even for small values of α.

By making use of an Edgeworth expansion, estimators with smaller median bias can be constructed via median bias reduction (Pfanzagl, 1973b; Kenne Pagui et al., 2017). Pfanzagl (1973b, Section 6) provides a general recipe for constructing estimators with a median bias of $o(n^{-(s-2)/2})$ for any $s \geqslant 3$. Kenne Pagui et al. (2017, Eq. (3)) yields estimates $\widehat{\theta}_j$ that satisfy $|\mathbb{P}(\widehat{\theta}_j \leqslant \theta_0) - 1/2| = O((B_{\alpha,0}/n)^{3/2})$. In this case, $\widetilde{\Delta}^2 \leqslant \mathfrak{D}_X^2 B_{\alpha,0}^3/n^3$ for some constant $\mathfrak{D}_X$ and hence, requirement (14) can be relaxed to

$$\mathfrak{D}_X^2 \frac{B_{\alpha,0}^5}{n^3} \leqslant \frac{\alpha}{P(B_{\alpha,0};0)} - 1 \quad \Leftrightarrow \quad \mathfrak{D}'_X \frac{(\log(2/\alpha))^5}{n^3} \leqslant \frac{\alpha}{P(B_{\alpha,0};0)} - 1,$$

for a slightly adjusted constant $\mathfrak{D}'_X$. The reduction in median bias, hence, leads to a smaller threshold sample size N_α after which our intervals are finite-sample valid.

The above argument for asymptotically median unbiased estimators implies that the smallest interval containing $B_{\alpha,0}$ many independent estimators of θ_0 has a finite sample coverage of at least $1-\alpha$ after a sample size of N_α. This, however, does not imply coverage validity for the confidence interval returned by Algorithm 1. This happens because with non-zero probability Algorithm 1 uses $B_{\alpha,0} - 1 < B_{\alpha,0}$ estimators. The following result proves upper and lower bounds on the miscoverage of the confidence interval returned by Algorithm 1 with Δ whenever the estimation procedure $\mathcal{A}(\cdot)$ has a median bias of $\widetilde{\Delta}$ "converging" to Δ. For simplicity, the result is stated only for asymptotically median unbiased estimators, i.e., $\Delta = 0$. See Remark 2.2 and the proof of Theorem 2 for upper and lower bounds on true coverage when Algorithm 1 is applied with Δ when the estimators has a median bias of at most $\widetilde{\Delta}$.

Recall that $\widehat{\mathrm{CI}}_{\alpha,\Delta}$ is the confidence interval returned by Algorithm 1 when it is applied with Δ as the median bias parameter. Theorem 2 below is proved in Section S.5.

Theorem 2. *Suppose $X_1, \ldots, X_n$ are independent random variables. If the estimation procedure $\mathcal{A}(\cdot)$ returns estimators that have a median bias of at most $\widetilde{\Delta} \geqslant 0$, then*

$$\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,0}) \leqslant \alpha\left(1 + \frac{B_{\alpha,0}(B_{\alpha,0}-1)}{2}\widetilde{\Delta}^2\right) \quad \textit{for every} \quad \alpha \in (0,1). \tag{15}$$

Furthermore, if $\mathbb{P}(\widehat{\theta}_j = \theta_0) = 0$ and the estimators $\widehat{\theta}_j, j \geqslant 1$ all have the same median bias $\widetilde{\Delta}$, then

$$\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,0}) \geqslant \alpha, \quad \textit{for every} \quad \alpha \in (0,1). \tag{16}$$

Remark 2.2 In Section S.5, we also consider the case when Δ is not necessarily 0. If the finite-sample median bias $\widetilde{\Delta}$ of the estimator procedure $\mathcal{A}(\cdot)$ is close to Δ (rather than zero), then

$$\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta}) \leqslant \alpha\left(1 + 2|\widetilde{\Delta} - \Delta|\right)^{B_{\alpha,\Delta}}.$$

Further, if $\mathbb{P}(\widehat{\theta}_j = \theta_0) = 0$ and the estimators $\widehat{\theta}_j, j \geqslant 1$ all have the same median bias $\widetilde{\Delta}$, then

$$\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta}) \geqslant \alpha\left(1 + 2|\widetilde{\Delta} - \Delta|\right)^{-B_{\alpha,\Delta}}.$$

These two inequalities imply that if $B_{\alpha,\Delta}|\widetilde{\Delta}-\Delta| = o(1)$, then $\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta})$ converges to α. $\diamond$

Remark 2.3 The main conclusion of Theorem 2 is that Algorithm 1 can be used with $\Delta = 0$ and it retains asymptotic validity for large sample sizes if the estimation procedure $\mathcal{A}(\cdot)$ produces asymptotically median unbiased estimators. $\diamond$

Theorem 2 is a finite sample result characterizing explicitly the effect of misspecifying Δ in Algorithm 1. The misspecification of Δ is measured by how far the median bias $\widetilde{\Delta}$ of the estimators $\widehat{\theta}_j, j \geqslant 1$ is from Δ, the asymptotic median bias. To illustrate Theorem 2, consider the setting under which (12) holds true. Theorem 2 along with (12) implies that,

$$\alpha \;\leqslant\; \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,0}) \;\leqslant\; \alpha\left(1 + \mathfrak{C}'_X \frac{B^3_{\alpha,0}}{n}\right), \quad \text{for every} \quad \alpha \in (0,1). \tag{17}$$

In case an estimation procedure $\mathcal{A}(\cdot)$ with reduced median bias is employed in Algorithm 1, then we get $\widetilde{\Delta} \leqslant \mathfrak{D}_X (B_{\alpha,0}/n)^{3/2}$ for some constant $\mathfrak{D}_X$ and hence, Theorem 2 yields

$$\alpha \;\leqslant\; \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,0}) \;\leqslant\; \alpha\left(1 + \mathfrak{D}'_X \frac{B^5_{\alpha,0}}{n^3}\right), \quad \text{for every} \quad \alpha \in (0,1). \tag{18}$$

Theorem 2 (and the conclusions (17), (18)) can be compared to the guarantees offered by classical confidence intervals constructed based on the assumption of asymptotic Normality. Under a bound like (11), such confidence intervals only satisfy

$$\left|\mathbb{P}\left(\theta_0 \notin \widehat{\mathrm{CI}}^{\texttt{Wald}}_{\alpha}\right) - \alpha\right| \leqslant \frac{\mathfrak{C}_X}{\sqrt{n}}, \quad \text{for every} \quad \alpha \in (0,1). \tag{19}$$

In other words, the coverage of $\widehat{\mathrm{CI}}^{\texttt{Wald}}_{\alpha}$ differs from $(1-\alpha)$ by a quantity of order $1/\sqrt{n}$ and can significantly miscover if $\alpha \ll 1/\sqrt{n}$. The same comment also applies to the bootstrap and subsampling confidence intervals. Confidence intervals obtained by various methods are often compared in terms of the rate of convergence in (19). In parametric models or, more generally, cases where θ_0 is estimable at an $n^{-1/2}$ rate, confidence intervals which attain a rate of $n^{-1/2}$ in (19) are called first-order accurate, those that attain a rate of n^{-1} are called second-order accurate and so on. Asymptotic Normality based Wald confidence intervals $\widehat{\mathrm{CI}}^{\texttt{Wald}}_{\alpha}$ are usually first-order accurate. Bootstrap confidence intervals can be constructed to be second-order accurate (Hall, 1986, 1988; Mammen, 1992). Subsampling intervals can also be constructed to satisfy second-order accuracy (Bertail and Politis, 2001). In contrast, the HulC readily obtains second-order accuracy and is valid even if α converges to zero. Further, if we use an estimator with reduced median bias, the HulC is sixth-order accurate (18). Another important difference is that the HulC attains relative accuracy (i.e., $|\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,0})/\alpha - 1|$ is small) instead of absolute accuracy as in (19). In the problem of mean estimation, some results for relative accuracy of Wald confidence intervals are available using self-normalized large deviation techniques (Shao, 1997; Jing et al., 2003). To our knowledge, such refined results are unavailable for a large class of M-estimators.

2.3 Comparison with Wald confidence intervals

In this section we show that our intervals have lengths close to those of the Wald intervals. In order to facilitate this comparison, we assume in this section that $\alpha \to 0$ slowly as a function of n. Let $\widehat{\theta} = \mathcal{A}(\{X_1, \ldots, X_n\})$ and $\widehat{\theta}_j = \mathcal{A}(\{X_i : i \in S_j)$. Suppose that

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2) \quad \text{and} \quad \sqrt{|S_j|}(\widehat{\theta}_j - \theta_0) \xrightarrow{d} N(0, \sigma^2), \tag{20}$$

as $n \to \infty$ and $|S_j| \to \infty$ for all $1 \leqslant j \leqslant B^*$. Under this assumption, if $\widehat{\sigma}^2$ is a consistent estimator σ^2, then the Wald confidence interval is given by $\widehat{\mathrm{CI}}_\alpha^{\texttt{Wald}} := [\widehat{\theta} - \widehat{\sigma} z_{\alpha/2}/\sqrt{n}, \widehat{\theta} + \widehat{\sigma} z_{\alpha/2}/\sqrt{n}]$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$-th quantile of the standard Gaussian distribution. The width of this confidence interval is given by $\sqrt{n}\mathrm{Width}(\widehat{\mathrm{CI}}_\alpha^{\texttt{Wald}}) = 2z_{\alpha/2}\widehat{\sigma}$. This converges in probability to $2z_{\alpha/2}\sigma$. From the properties of the Normal distribution, it follows that $z_{\alpha/2} = \sqrt{2\log(2/\alpha) - \log(\log(2/\alpha)) - \log(2\pi)} + o(1)$ as $\alpha \to 0$. See, for example, Proposition 4.1 of Boucheron and Thomas (2012). Hence, the width of the Wald confidence interval is asymptotically equal to $2\sigma\sqrt{2\log(2/\alpha)}/\sqrt{n}$, as $\alpha \to 0$ and $n \to \infty$.

To compare this width to the width of the HulC, for simplicity, we treat B^* as a fixed value and assume that n is a multiple of B^* so that each split has n/B^* many observations. Assumption (20) implies that $\sqrt{n/B^*}(\widehat{\theta}_j - \theta_0) \xrightarrow{d} N(0, \sigma^2)$ for $1 \leqslant j \leqslant B^*$. Because the estimators are independent and $B^* \leqslant B_{\alpha,\Delta} < \infty$, we get that the convergence is joint for all the estimators $\widehat{\theta}_j, 1 \leqslant j \leqslant B^*$. Recall that our confidence interval is the smallest rectangle containing these estimators and hence

$$\sqrt{\frac{n}{B^*}}\mathrm{Width}(\widehat{\mathrm{CI}}_{\alpha,\Delta}) = \sqrt{\frac{n}{B^*}}\left[\max_{1\leqslant j\leqslant B^*}\widehat{\theta}_j - \min_{1\leqslant j\leqslant B^*}\widehat{\theta}_j\right] = \max_{1\leqslant j<k\leqslant B^*}\sqrt{\frac{n}{B^*}}(\widehat{\theta}_j - \widehat{\theta}_k). \tag{21}$$

Joint asymptotic convergence of the estimators implies that

$$\sqrt{\frac{n}{B^*}}\mathrm{Width}(\widehat{\mathrm{CI}}_{\alpha,\Delta}) \xrightarrow{d} \max_{1\leqslant j<k\leqslant B^*}(G_j - G_k) = \max_{1\leqslant j\leqslant B^*} G_j - \min_{1\leqslant j\leqslant B^*} G_j, \tag{22}$$

where $(G_1, \ldots, G_{B^*})$ is a Gaussian random vector with mean zero and a diagonal covariance matrix with all diagonal entries equal to σ^2. This shows the first difference in widths. Unlike the classical Wald confidence intervals, the width of our confidence interval does not degenerate after scaling by $\sqrt{n}$; the width after proper scaling converges weakly to a non-degenerate distribution. Using (22), we can control of the width of our confidence region in terms of the width of the convex hull of B^* many independent mean zero Gaussian random variables. Because G_j's are symmetric around zero,

$$\mathbb{E}\left[\max_{1\leqslant j\leqslant B^*} G_j - \min_{1\leqslant j\leqslant B^*} G_j\right] = 2\mathbb{E}\left[\max_{1\leqslant j\leqslant B^*} G_j\right] = 2\sqrt{2\log(B^*)}\left[1 - \frac{\log\log B^*}{4\log B^*} + O\left(\frac{1}{\log B^*}\right)\right].$$

The last equality here holds as $\alpha \to 0$ and follows from Theorem 1.2 of Kabluchko and Zaporozhets (2019) (and the discussion before that theorem). Therefore, the width of our confidence interval is asymptotically $2\sigma\sqrt{2B^*\log(B^*)/n}$. From inequalities (7), we know that $B_{\alpha,\Delta}$ and B^* are of order $\log_2(2/\alpha)$; note that under asymptotic Normality, we can take $\Delta = 0$. Hence, the width of our confidence interval is asymptotically

$$2\sigma\sqrt{\frac{2\log(2/\alpha)}{n}}\sqrt{\log_2(\log_2(2/\alpha))}. \tag{23}$$

This implies that the ratio of the expected width of our confidence interval to that of the Wald interval is approximately equal to $\sqrt{\log_2(\log_2(2/\alpha))}$. This is always larger than 1, and grows very slowly as $\alpha \to 0$. For $\alpha \in [0.01, 0.2]$, this ratio ranges between 1.71 and 1.32. In a way, this is the price to pay for the generality of the confidence interval. While the Wald confidence interval makes complete use of asymptotic Normality, our confidence interval only makes use of the fact that its median is zero; we do not even make use of symmetry.

2.3.1 Numerical Comparisons

Figure 4 shows the coverage and width of the 95% HulC interval and the Wald interval from ordinary least squares linear regression. The simulation setting is as follows: for $n \in \{20, 50, 100, 1000\}$, independent

observations $(X_i, Y_i), 1 \leqslant i \leqslant n$ are generated from

$$X_i \sim \text{Uniform}[0, 10],\ \xi_i \sim N(0, 1), \quad \text{and} \quad Y_i = 1 + 2X_i + \gamma X_i^{1.7} + \exp(\gamma X_i)\xi_i. \tag{24}$$

For $\gamma = 0$, observations (X_i, Y_i) follow the standard linear model and for $\gamma > 0$, observations do not follow a linear model with non-linear mean function and a heteroscedastic error variable. The misspecification increases with γ. We define the estimator and target as $\widehat{\beta}$ and β^*_γ, where

$$(\widehat{\alpha}, \widehat{\beta}) := \arg\min_{\alpha,\beta} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2, \quad \text{and} \quad (\alpha^*_\gamma, \beta^*_\gamma) := \arg\min_{\alpha,\beta} \mathbb{E}_\gamma[(Y - \alpha - \beta X)^2].$$

Here $\mathbb{E}_\gamma[\cdot]$ represents the expectation when (X, Y) are generated from (24). Note that β^*_γ need not be equal to 2 for $\gamma > 0$. By Monte-Carlo approximation of $\mathbb{E}_\gamma[\cdot]$ with 10^8 samples, we have $\beta^*_{0.25} = 3.2791, \beta^*_{0.5} = 4.5567, \beta^*_{0.75} = 5.8239$, and $\beta^*_1 = 6.8093$. The Wald interval in this case are obtained using the sandwich variance estimator as in Buja et al. (2019).

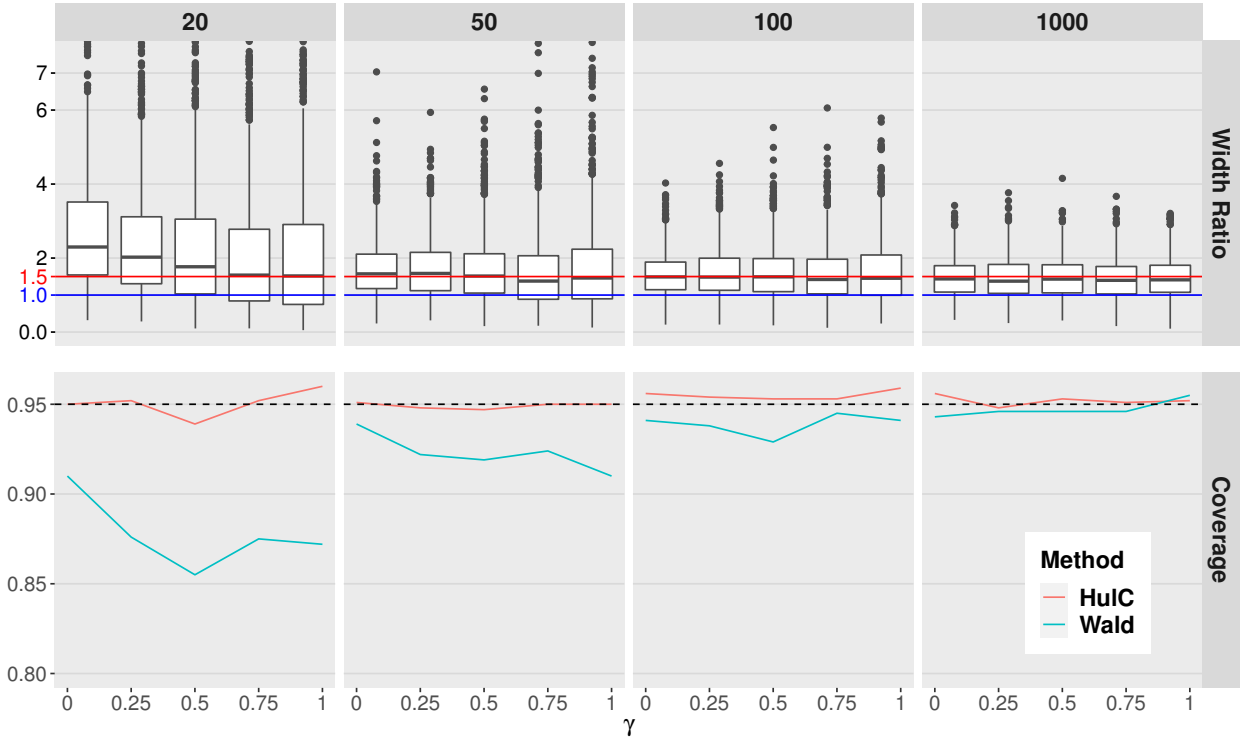


Figure 4: Comparison of width and coverage attained by our confidence interval and the Wald confidence interval in a potentially misspecified linear model. The coverage in bottom panel is based on 1000 replications. Our method is shown as "HulC" and Wald's is shown as "Wald." The four column plots correspond to four different sample sizes $n = 20, 50, 100, 1000$. The top panel shows the ratio of the widths of our confidence interval to that of the Wald confidence interval; the red line is horizontal at 1.5 representing the case where the HulC yields a 50% larger interval than Wald's. The width ratio plot is truncated at $y = 7$.

Figure 5 provides an illustration when the estimator is obtained from multiple linear regression. The setting for Figure 5 is as follows: for $20 \leqslant n \leqslant 500$, independent observations $(X_i, Y_i) \in \mathbb{R}^6 \times \mathbb{R}, 1 \leqslant i \leqslant n$ are generated from $Y_i = |\theta_0^\top X_i| + \xi_i$, where $\xi_i \sim N(0, 1)$ and $X_i \in \mathbb{R}^6$ is generated according

to the following law: $(X_{i,1}, X_{i,2}) \sim \text{Uniform}[-1,1]^2$, $X_{i,3} := 0.2X_{i,1} + 0.2(X_{i,2}+2)^2 + 0.2Z_{i,1}$, $X_{i,4} := 0.1 + 0.1(X_{i,1}+X_{i,2}) + 0.3(X_{i,1}+1.5)^2 + 0.2Z_{i,2}$, $X_{i,5} \sim \text{Ber}(\exp(X_{i,1})/\{1+\exp(X_{i,1})\})$, and $X_{i,6} \sim \text{Ber}(\exp(X_{i,2})/\{1+\exp(X_{i,2})\})$. Here $(Z_{i,1}, Z_{i,2}) \sim \text{Uniform}[-1,1]^2$ are independent of $(X_{i,1}, X_{i,2})$ and $\theta_0 = (1.3, -1.3, 1, -0.5, -0.5, -0.5)/\sqrt{5.13}$. This is also a misspecified linear regression model and is taken from Kuchibhotla et al. (2021). Our estimator and target are defined as

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) := \underset{\alpha,\beta,\gamma}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha - \beta X_{i,1} - \gamma^\top X_{i,-1})^2, \quad \text{and} \quad (\alpha^*, \beta^*, \gamma^*) := \underset{\alpha,\beta,\gamma}{\arg\min}\, \mathbb{E}[(Y - \alpha - \beta X_1 - \gamma^\top X_{-1})^2],$$

where $X_{i,-1}$ and X_{-1} represent the last 5 coordinates of X_i and X respectively. With Monte Carlo approximation of $\mathbb{E}[\cdot]$, we found that $\beta^* = -0.137323$. For level $\alpha = 0.05$, the HulC requires splitting the data into approximately 5 parts. This implies that for a sample of size 20, each part only has 4 observations and one cannot fit uniquely a linear regression estimator because the model has 6 covariates. Interestingly, when we just use the output from `R` function `lm()`, the HulC still covers the true β^* with required confidence because in this case `lm()` simply ignores the last 2 covariates.

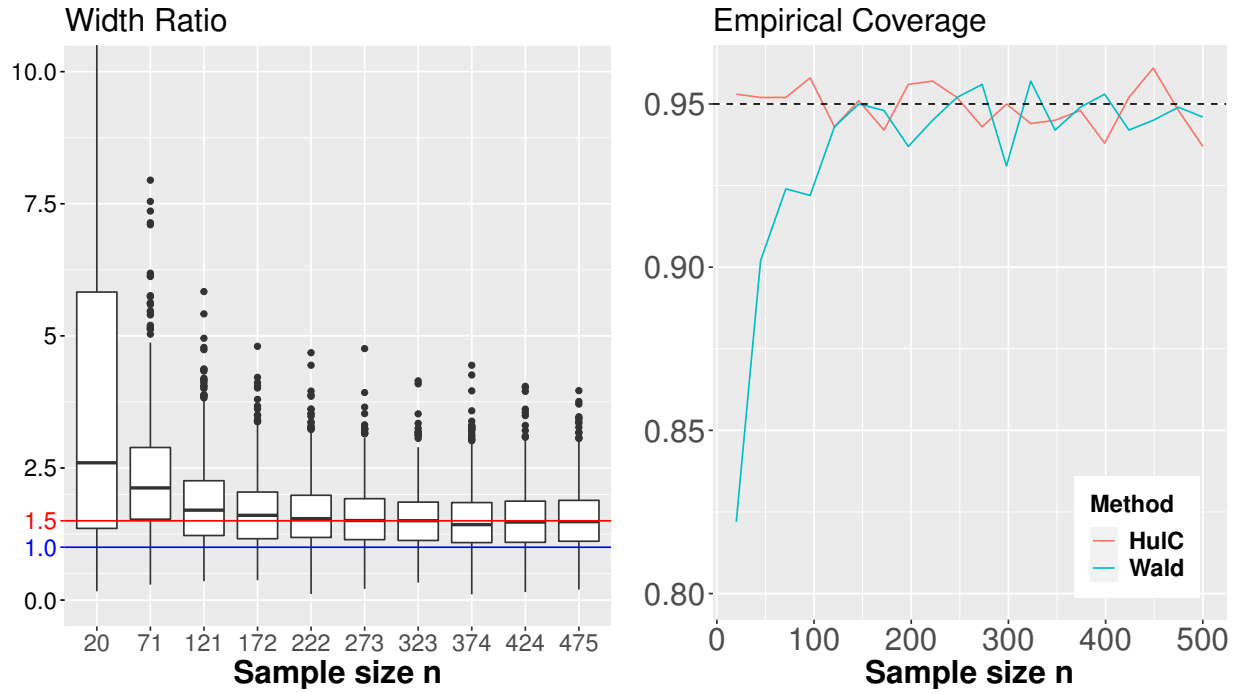


Figure 5: Comparison of width and coverage between our confidence interval with Wald's with a multiple linear regression estimator as the sample size changes from 20 to 500. Our method is shown as "HulC" and Wald's is shown as "Wald." The red horizontal line in width ratio plot (left) is at 1.5, which means our confidence interval is 50% larger than Wald's. The empirical coverage in the right plot is computed based on 1000 replications for each sample size. The width ratio plot is truncated at $y = 10$.

Unlike the Wald confidence interval, the HulC does not explicitly or implicitly estimate the variance of the estimator but its width as given in (23) adapts to the unknown standard deviation σ. The calculation shown above uses asymptotic arguments, but some simple bounds can be obtained using no more than two

moments for $\widehat{\theta}_j$. Observe from (21) that

$$\begin{aligned}\mathbb{E}\left[\text{Width}(\widehat{\text{CI}}_{\alpha,\Delta})\right] &= \mathbb{E}\left[\max_{1\leqslant j\leqslant B^*}\widehat{\theta}_j - \min_{1\leqslant j\leqslant B^*}\widehat{\theta}_j\right] \\ &\leqslant 2\mathbb{E}\left[\max_{1\leqslant j\leqslant B^*}|\widehat{\theta}_j-\theta^*|\right] \leqslant 2\left(\mathbb{E}\left[\max_{1\leqslant j\leqslant B^*}|\widehat{\theta}_j-\theta^*|^2\right]\right)^{1/2} \\ &\leqslant 2\left(\mathbb{E}\left[\sum_{j=1}^{B^*}|\widehat{\theta}_j-\theta^*|^2\right]\right)^{1/2} \leqslant 2\sqrt{B^*}\max_{1\leqslant j\leqslant B^*}\left(\mathbb{E}[|\widehat{\theta}_j-\theta^*|^2]\right)^{1/2}.\end{aligned} \tag{25}$$

Assuming convergence in mean square of $\sqrt{n/B^*}(\widehat{\theta}_j-\theta^*)$ to a distribution with mean zero and variance σ^2, we get that the expected width is asymptotically bounded by $2\sigma B^*/\sqrt{n}$. This calculation does not require convergence to Gaussianity and shows that the width of our confidence interval, in general, adapts to the standard deviation of the estimators. The calculation (25) can be significantly improved if the estimators are known to have higher moments. In the first inequality of (25) we only use second moment Jensen's inequality. Replacing the second moments by q-th moment here will yield $(B^*)^{1/q}$ instead of $\sqrt{B^*}$ in the last line of (25).

Transformed Parameters. In contrast to Wald intervals, the HulC interval is equivariant to monotone transformations. It is worth noting that the validity of our confidence interval does not require any smoothness conditions on the transformation $g(\cdot)$. In comparison, the delta method requires continuous differentiability of $g(\cdot)$.

2.4 HulC for multivariate parameters

Suppose now that $\theta_0 \in \mathbb{R}^d$. A slight modification of the HulC still works if we replace the interval with either the convex hull or the rectangular hull of the estimators.

As before, let $\widehat{\theta}_j, 1\leqslant j\leqslant B$ be independent estimators of $\theta_0\in\mathbb{R}^d$. The convex hull of a set of points in $\mathbb{R}^d$ is the smallest convex set containing these points. The smallest rectangle containing the estimators $\widehat{\theta}_j, 1\leqslant j\leqslant B$, which we call the *rectangular hull*, is

$$\text{RectHull}(\{\widehat{\theta}_j : 1\leqslant j\leqslant B\}) \;:=\; \bigotimes_{k=1}^{d}\left[\min_{1\leqslant j\leqslant B} e_k^\top\widehat{\theta}_j,\; \max_{1\leqslant j\leqslant B} e_k^\top\widehat{\theta}_j\right],$$

where $e_k, 1\leqslant k\leqslant d$ represent the canonical basis vectors in $\mathbb{R}^d$; and $\bigotimes$ denotes the Cartesian product.

To compactly state our next result, we define the following coordinate-wise maximum median bias,

$$\Delta = \max_{1\leqslant k\leqslant d}\max_{1\leqslant j\leqslant B}\text{Med-bias}_{e_k^\top\theta_0}(e_k^\top\widehat{\theta}_j). \tag{26}$$

Similar to Lemma 1, we have the following result (proved in Section S.6) on the coverage of the convex hull and the smallest rectangle.

Lemma 2. *Suppose $\widehat{\theta}_j, 1\leqslant j\leqslant B$ are independent estimators of $\theta_0\in\mathbb{R}^d$.*

1. *If $\mathbb{P}(c^\top(\widehat{\theta}_j-\theta_0)\leqslant 0) = 1/2$ for all $c\in\mathbb{R}^d\backslash\{0\}$, then for $B\geqslant d+1$,*

$$\mathbb{P}\left(\theta_0\notin\text{ConvHull}(\{\widehat{\theta}_j : 1\leqslant j\leqslant B\})\right) \;=\; \frac{1}{2^{B-1}}\sum_{i=0}^{B-d-1}\binom{B-1}{i}. \tag{27}$$

2. Recall the definition of Δ *in* (26)*. For all* $B \geqslant 1$*,*

$$\mathbb{P}\left(\theta_0 \notin \text{RectHull}(\{\widehat{\theta}_j : 1 \leqslant j \leqslant B\})\right) \;\leqslant\; d\left\{\left(\frac{1}{2}+\Delta\right)^B + \left(\frac{1}{2}-\Delta\right)^B\right\}. \tag{28}$$

The proof of (27) follows from the works of Wendel (1962) and Wagner and Welzl (2001). The requirement of more than $d+1$ estimators can be restrictive in practice. This is especially so in near high dimensional problems where the dimension d can grow with the sample size n. The proof of (28) follows by using the union bound on the univariate confidence region in Lemma 1. Furthermore, to obtain (28) we only assume that the coordinate-wise median bias of the estimators is bounded and this condition is much weaker than the corresponding condition used to obtain (27).

Inequality (28) is written with a single parameter Δ as a bound on the median bias for all coordinates. It is, however, easy to obtain similar bounds when the median bias is different for different coordinates; see the proof of Lemma 2 for details. Similarly, we also note that, one need not compute B random vector estimators. One might construct a different number of estimators for $e_k^\top\theta_0$ and construct univariate confidence intervals along each coordinate to obtain a multivariate confidence rectangle for θ_0. Formally, if $\widehat{\text{CI}}^{(k)}_{\alpha/d,\Delta}$ is a confidence interval of level $1-\alpha/d$ for $e_k^\top\theta_0$ constructed using Algorithm 1 with (a known) Δ, then

$$\mathbb{P}\left(\theta_0 \in \bigotimes_{k=1}^{d} \widehat{\text{CI}}^{(k)}_{\alpha/d,\Delta}\right) \geqslant 1-\alpha.$$

Note that construction of $\widehat{\text{CI}}^{(k)}_{\alpha/d,\Delta}$ requires $B_{\alpha/d,\Delta}$ estimators of $e_k^\top\theta_0$. Following inequalities (7), we conclude that

$$\max\left\{\left\lceil\frac{\log(d/\alpha)}{\log(2/(1+2\Delta))}\right\rceil, \left\lceil\frac{\log(2d/\alpha)}{\log(2)}\right\rceil\right\} \leqslant B_{\alpha/d,\Delta} \leqslant \left\lceil\frac{\log(2d/\alpha)}{\log(2/(1+2\Delta))}\right\rceil \quad\Rightarrow\quad B_{\alpha/d,\Delta} \asymp \log(2d/\alpha).$$

This implies that one only needs to split the original data $X_1,\ldots,X_n$ into (about) $\log(2d/\alpha)$ many batches. In Lemma 2, (28) requires $B \geqslant C\log(2d/\alpha)$ for a constant C for a coverage of $1-\alpha$. This can be compared with the requirement $B \geqslant d+1$ for the validity of (27). Hence, for moderate to high dimensional problems, the smallest rectangle is an economical choice.

Similar to the univariate case, one need not know median bias of $e_k^\top\widehat{\theta}_j$ exactly. It suffices to know them approximately as dictated by Proposition 1. The conclusions from Proposition 1 continue to hold true even with a growing dimension. For instance, for asymptotically median unbiased estimators, if

$$\mathfrak{C}'_X \frac{\log^3(2d/\alpha)}{n} \;\leqslant\; \frac{\alpha}{dP(B_{\alpha/d,0};0)} - 1, \tag{29}$$

for some constant $\mathfrak{C}'_X$, then irrespective of the dimension $d \geqslant 1$, we obtain

$$\mathbb{P}\left(\theta_0 \in \text{RectHull}(\{\widehat{\theta}_j : 1 \leqslant j \leqslant B_{\alpha/d,0}\})\right) \geqslant 1-\alpha.$$

It follows from Figures 2 & 3 that the right hand side of (29) can be as large as 0.4 for certain choices of α, even for $d \gg n$. Similarly, Theorem 2 (in particular its implication (17)) yields

$$\mathbb{P}\left(\theta_0 \notin \bigotimes_{k=1}^{d} \widehat{\text{CI}}^{(k)}_{\alpha/d,0}\right) \leqslant \sum_{k=1}^{d} \mathbb{P}\left(\theta_0 \notin \widehat{\text{CI}}^{(k)}_{\alpha/d,0}\right) \leqslant \sum_{k=1}^{d} \frac{\alpha}{d}\left(1+\mathfrak{C}'_X\frac{B^3_{\alpha/d,0}}{n}\right) = \alpha\left(1+\mathfrak{C}'_X\frac{B^3_{\alpha/d,0}}{n}\right).$$

Here the union bound is valid irrespective of what the dimension d is relative to the sample size n. If the estimators are median bias reduced, then the same argument as above yields

$$\mathbb{P}\left(\theta_0 \notin \bigotimes_{k=1}^{d} \widehat{\mathrm{CI}}^{(k)}_{\alpha/d,0}\right) \leqslant \alpha\left(1 + \mathfrak{C}'_X \frac{B^5_{\alpha/d,0}}{n^3}\right).$$

Note, once again, that $B_{\alpha/d,0} \asymp \log(2d/\alpha)$ and hence the miscoverage probabilities are bounded by $\alpha(1 + \mathfrak{C}'_X \log^3(d/\alpha)/n)$ (under (12)) and $\alpha(1 + \mathfrak{C}'_X \log^5(d/\alpha)/n^3)$ (under median bias reduction).

We see that we only require $\log(d/\alpha) = o(n^{1/3})$ (or $\log(d/\alpha) = o(n^{3/5})$ in the bias reduced case) and we do not require joint/multivariate distributional convergence whatsoever. This can be contrasted with the results from the literature on high-dimensional central limit theorems (Belloni et al., 2018; Koike, 2020; Fang and Koike, 2020; Deng, 2020; Chernozhukov et al., 2020). These results concern Gaussian approximation for high-dimensional averages. Under certain moment assumptions on the *joint* distribution, these results imply a "joint" Gaussian approximation with a minimum requirement of $\log(d) = o(n^{1/3})$ (Fang and Koike, 2020, Proposition 1.1). Belloni et al. (2018, Theorem 2.3) uses the union bound based on moderate deviations but still requires joint moment conditions and the condition that $\log(d) = o(n^{1/3})$. With usual estimators, the HulC also has the same dimensionality requirement while only making use of marginal median bias. With median bias reduced estimators, the HulC only requires $\log(d) = o(n^{3/5})$ and this is even weaker than $\log(d) = o(n^{1/2})$, which is the best possible dimension restriction for a Gaussian approximation (Das and Lahiri, 2020, Theorem 3). It is worth mentioning that by slightly enlarging the bootstrap confidence regions, the dimensionality requirement can be reduced to $\log(d) = o(n)$ in the case of mean estimation (Deng, 2020).

Although we have used union bound above to obtain a coverage of $1 - \alpha$ for a multivariate parameter, we only required *asymptotic* median unbiasedness of $e_k^\top \widehat{\theta}_j$ marginally. There is no requirement whatsoever on the asymptotic joint convergence or symmetry of $\widehat{\theta}_j \in \mathbb{R}^d$. Interestingly, such a result is not possible with the usual confidence intervals. This point is discussed further in Section S.1 of the supplementary file.

3 Applications to standard problems

In this section, we present some simple applications including mean estimation, median estimation, and parametric exponential models. In parametric and semi-parametric models, regularity conditions and efficiency theory implies the existence of estimators which when centered at the target have an asymptotic mean zero Gaussian distribution. In these cases, often one can modify the estimators to ensure reduced median bias. For some examples of (approximately) median-unbiased estimators, see Birnbaum (1964); John (1974); Pfanzagl (1970a,b, 1972, 1979); Hirji et al. (1989); Andrews and Phillips (1987); Kenne Pagui et al. (2017).

3.1 Mean estimation

Suppose $X_1, \ldots, X_n$ are independent real-valued random variables with a common mean $\mu \in \mathbb{R}$. Consider the problem of constructing a confidence interval for μ. Note that the random variables need not be identically distributed. If the random variables have a finite second moment and satisfy the Lindeberg condition, then the sample mean $\bar{X}_j = |S_j|^{-1} \sum_{i \in S_j} X_i$ satisfies $\sqrt{|S_j|}(\bar{X}_j - \mu) \xrightarrow{d} N(0, \sigma_j^2)$, where $\sigma_j^2 = \sum_{i \in S_j} \mathrm{Var}(X_i)/|S_j|$. This implies that the estimator $\bar{X}_j$ is asymptotically median unbiased and Algorithm 1 with $\Delta = 0$ yields an asymptotically valid confidence interval for μ. In this case, Wald intervals also have asymptotic coverage.

The setting becomes more interesting when we consider random variables with less than two finite moments. In this case, the limiting distribution of $\bar{X}_j$ is known to be a stable law and its rate of convergence also changes depending on the tail decay of the random variables. If the random variables satisfy

$$\lim_{x\to\infty} x^{\alpha}\mathbb{P}(X_i > x) = \lim_{x\to\infty} x^{\alpha} P(X_i < -x) \quad \text{for some} \quad \alpha \in [1, 2), \tag{30}$$

then the limiting stable law of $\bar{X}_j$ is symmetric around zero (see, for instance, Theorem 9.34 in Breiman (1968)). In this special case, Algorithm 1 continues to provide asymptotically valid confidence intervals, while Wald intervals and the bootstrap are known to fail for $\alpha < 2$; see, for example, Athreya (1987) and Knight (1989). In particular, if the underlying distributions are all symmetric around the mean μ, then without any moment assumptions the confidence interval returned by Algorithm 1 is finite sample valid. It is worth noting that subsampling (Romano and Wolf, 1999) is still applicable in the case of infinite variance.

If the assumption (30) does not hold true, then the limiting stable law is not symmetric and the asymmetry depends on the gap between left and right hand side quantities in (30). In this case, the median bias of the limiting distribution is not readily available and the methods presented in previous sections are not applicable. This can be resolved using the Adaptive HulC which we describe in Section 4.

3.2 Median estimation

Suppose $X_1, \ldots, X_n$ are independent real-valued random variables with common median $m \in \mathbb{R}$. Consider the problem of constructing a confidence interval for m. The usual estimator for the population median is the sample median. Set $\widehat{\theta}_j = \text{median}(X_i : i \in S_j)$. If the average distribution function $\bar{F}_j(t) = |S_j|^{-1}\sum_{i\in S_j}\mathbb{P}(X_i \leqslant t)$ has a derivative bounded away from zero at m, then it is known (Sen, 1968) that $\sqrt{|S_j|}(\widehat{\theta}_j - m) \xrightarrow{d} N(0, \sigma_j^2)$, where $\sigma_j^2 = (4\bar{f}_j^2(m))^{-1}$. Here $\bar{f}_j(m)$ is the derivative of $\bar{F}_j(t)$ at $t = m$. There are several classical methods for constructing confidence intervals for m including Wald's, quantile or rank based intervals. Wald confidence intervals in this case require estimating of the density $\bar{f}_j(\cdot)$ at m and the quantile based intervals require choosing the appropriate quantiles for end points. Unlike the Wald interval, the quantile based intervals are finite sample valid (Lanke, 1974). Because the limiting distribution is mean zero Gaussian, the HulC applies and yields an asymptotically valid confidence interval.

Once again the setting becomes interesting when the underlying distributions do not satisfy the conditions for Normality. For example, if the density $\bar{f}_j(\cdot)$ is not bounded away from zero at the common median m, then the limiting distribution of $\widehat{\theta}_j$ is not Gaussian and hence Wald as well as bootstrap intervals break down. The limiting distribution in this case is explicitly described in Knight (1998, Section 2), Knight and Bassett (2002, Section 5), and Geyer (1996, Example 2). In this case, the rate of convergence of the median depends on how fast the density decays to zero as t approaches m. When the population median m is unique, the sample median computed based on odd number of observations is known to be median unbiased (Desu and Rodine, 1969). This observation implies that Algorithm 1 with $\Delta = 0$ yields a finite sample valid confidence interval for m if each S_j has an odd number of observations (which can be trivially ensured). In fact, with any given number of observations (even or odd), an estimator that randomly (equally likely) chooses between the r-th order statistic and $(|S_j| - r + 1)$-th order statistic is median unbiased for m as shown in Section 4 of Desu and Rodine (1969).

3.3 Binomial distribution

Consider $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ for some $p \in (0,1)$. The problem of constructing confidence intervals for p is a well-studied problem with focus on coverage as p changes with the sample size n (Brown et al., 2002). It is well-known that the limiting distribution of $\text{Binomial}(n,p)$ as $n \to \infty$ changes from a Gaussian to a Poisson distribution depending on whether $np \to \infty$ or $np \to \lambda \in (0,\infty)$. Because of this change, the Wald confidence intervals can undercover p when p is small relative to the sample size n (Brown et al., 2001). We will now consider the coverage properties of the HulC when using the sample proportion as an estimator for p. For any set $S \subseteq \{1,2,\ldots,n\}$, $\sum_{i \in S} X_i \sim \text{Binomial}(|S|, p)$. Theorem 10 of Doerr (2018) shows that whenever $p \in [\log(4/3)/|S|, 1 - \log(4/3)/|S|]$, the estimator $\sum_{i \in S} X_i/|S|$ has a median bias of at most $1/4$. Theorem 1 of Greenberg and Mohri (2014) yields this result for $p \in [1/|S|, 1 - 1/|S|]$. For a more precise result, see Lemma 8 of Doerr (2018). Hence, Algorithm 1 with $\Delta = 1/4$ yields finite sample coverage for all $p \in [\log(4/3)/m, 1 - \log(4/3)/m]$; here m represents the minimum number of observations in each split of the data. Because $m \asymp n/\log(2/\alpha)$, we get finite sample coverage validity even for $p = \Theta(1/n)$. Note that the Binomial distribution is not approximately Normal in this case.

Allowing for some modifications of either the estimator or the final confidence set, we can obtain finite sample coverage for all $p \in [0,1]$. Firstly, note that Algorithm 1 with $\Delta = 0$ will always cover the true median of the estimators. With the proportion estimator $\sum_{i \in S_j} X_i/|S_j|$, Algorithm 1 with $\Delta = 0$ with a probability of at least $1-\alpha$ will contain the median of $\text{Binomial}(m,p)/m$ where $m = |S_j|$ for all $1 \leqslant j \leqslant B^*$. Hamza (1995) proves that

$$\left|\text{mean}\left(\frac{\text{Binom}(m,p)}{m}\right) - \text{median}\left(\frac{\text{Binom}(m,p)}{m}\right)\right| = \left|p - \text{median}\left(\frac{\text{Binom}(m,p)}{m}\right)\right| \leqslant \frac{\log(2)}{m}. \tag{31}$$

Therefore, if $\widehat{\text{CI}}_{\alpha,0} = [\widehat{L}, \widehat{U}]$ represents the confidence interval from Algorithm 1 with $\Delta = 0$, we get that for all $p \in [0,1]$,

$$\mathbb{P}\left(p \notin \left[\widehat{L} - \frac{\log(2)}{m}, \widehat{U} + \frac{\log(2)}{m}\right] \cap [0,1]\right) \leqslant \alpha.$$

This is a modification of confidence interval returned by Algorithm 1 but uses the classical Binomial proportion estimator. If we modify the estimator, then no changes are required in Algorithm 1 with $\Delta = 0$ to obtain a finite sample coverage. Because the Binomial distribution has a monotone likelihood ratio, the results of Pfanzagl (1970a, 1972) can be applied to obtain a median unbiased estimator of p. It might be worth noting here that Binomial distribution being discrete, any median unbiased estimator has to be randomized; see page 74 of Pfanzagl (2011) for a discussion. The exact median unbiased estimator of Pfanzagl (1970a, 1972) is computationally intensive. A simpler estimator for p with reduced median bias can be obtained from Hirji et al. (1989), and Kenne Pagui et al. (2017). These works discuss binary regression and estimating a Binomial proportion is the special case when there are no regressors except for an intercept. Hamza (1995) also proves that (31) holds true for $\text{Binom}(m,p)$ replaced by $\text{Poisson}(m\lambda)$. This implies that the confidence interval $\widehat{\text{CI}}_{\alpha,0}$ from Algorithm 1 inflated by $\log(2)/m$ also has a finite sample coverage for every $\lambda \geqslant 0$.

3.4 Exponential families

Pfanzagl (1979) provides an algorithm to construct an exactly median unbiased estimator for every sample size in a full rank exponential family, even in the presence of nuisance parameters. This paper considers a more general parametric model than exponential families. A related result for exponential families is also

obtained in Brown et al. (1976). For brevity, we will not describe this algorithm here and refer to the papers mentioned above; also, see Cabrera and Watson (1997) for some computational methods. With such an estimator, the HulC can be applied with $\Delta = 0$ to obtain a finite sample valid confidence interval.

3.5 Squared mean estimation

Suppose $X_1, \ldots, X_n$ are independent random variables with common mean μ and common variance $\sigma^2 < \infty$. Consider the estimation of $\theta_0 = \mu^2$. A natural estimator of θ_0 is $\widetilde{\theta} = \bar{X}_n^2$, the square of the sample mean. The asymptotic distribution of $\widetilde{\theta}$ depends on the true mean and the population variance:

$$n^{1/2}(\widetilde{\theta} - \theta_0) \xrightarrow{d} N(0, 4\mu^2\sigma^2), \quad \text{if } \theta_0 = \mu^2 \neq 0, \quad \text{and} \quad n(\widetilde{\theta} - \theta_0) \xrightarrow{d} \sigma^2\chi_1^2, \quad \text{if } \theta_0 = \mu^2 = 0.$$

There are two aspects to consider here. Firstly, the rate of convergence changes from $n^{-1/2}$ to n^{-1} as μ changes from non-zero to zero. Secondly, the limiting distribution of $\widetilde{\theta}$ becomes one-sided for $\mu = 0$ and this implies that the estimator has an asymptotic median bias of $1/2$ for $\mu = 0$. The first aspect is not an issue for Algorithm 1, but the second aspect renders Algorithm 1 useless for μ close to zero because it would require nearly infinite many splits of the data. Alternatively, for each subset S_j of the data, consider the U-statistic estimator

$$\widehat{\theta}_j = \frac{1}{|S_j|(|S_j|-1)} \sum_{i \neq k \in S_j} X_i X_k. \tag{32}$$

It readily follows that $\mathbb{E}[\widehat{\theta}_j] = \mu^2$ for all $\mu \in \mathbb{R}$, unlike $\bar{X}_n^2$ which is biased for μ close to zero. Once again $\widehat{\theta}_j$ has different limiting distributions depending the magnitude of μ.

$$\begin{aligned} \sqrt{|S_j|}(\widehat{\theta}_j - \theta_0) &\xrightarrow{d} N(0, 4\mu^2\sigma^2), \quad \text{if } \theta_0 = \mu^2 \neq 0, \\ |S_j|(\widehat{\theta}_j - \theta_0) &\xrightarrow{d} \sigma^2(\chi_1^2 - 1), \quad \text{if } \theta_0 = \mu^2 = 0. \end{aligned}$$

The rate of convergence changes between $\mu \neq 0$ and $\mu = 0$, but now the limiting distribution has median bias bounded away from zero. It may be worth pointing out that the limiting distribution in general would be a mixture of Normal and Chi-square as μ becomes close to zero. We can prove the following result on the median bias of $\widehat{\theta}_j$ that is uniform over all $\mu \in \mathbb{R}$. The proof in Section S.7 can be easily extended to accommodate non-identically distributed observations expect for common mean and variance.

Proposition 2. *Suppose $\xi_i = (X_i - \mu)/\sigma, 1 \leqslant i \leqslant n$ are independent and identically distributed. Then for any μ and σ, the median bias of $\widehat{\theta}_j$ is bounded by*

$$\begin{aligned} &\sup_{\theta \in \mathbb{R}} \left| \frac{1}{2} - \left\{ \Phi\left(\frac{-\theta + \sqrt{\theta^2 + |S_j|/(|S_j|-1)^2}}{\sqrt{|S_j|}/(|S_j|-1)} \right) - \Phi\left(\frac{-\theta - \sqrt{\theta^2 + |S_j|/(|S_j|-1)^2}}{\sqrt{|S_j|}/(|S_j|-1)} \right) \right\} \right| \\ &+ \sqrt{\frac{4\mathbb{E}[\xi_1^4]\log(|S_j|)}{|S_j|\pi}} + \frac{\mathbb{E}[|\xi_1|^3] + \mathbb{E}[|\xi_1|^6]/(\mathbb{E}[\xi_1^4])^{3/2}}{\sqrt{|S_j|}} + \frac{2}{|S_j|}. \end{aligned} \tag{33}$$

Note that the first term on the right hand side of (33) can be computed given $|S_j|$ without the knowledge of μ and σ. Further, the last three terms of (33) all disappear as $|S_j| \to \infty$ whenever certain moments of ξ_1 are bounded away from 0 and ∞. Exact computation for some sample sizes ($|S_j|$) shows that the supremum in the first term in (33) is attained at $\theta = 0$ and equals $|\mathbb{P}(\chi_1^2 \leqslant 1) - 1/2| \approx 0.183$. Hence, Algorithm 1 can be applied with $\Delta = |\mathbb{P}(\chi_1^2 \leqslant 1) - 1/2|$ and the estimator $\widehat{\theta}_j$ to attain an uniformly valid asymptotic confidence

interval for $\theta_0 = \mu^2$. It is worth pointing out that the resulting confidence interval is adaptive in its width as μ approaches zero, i.e. the expected width of the HULC interval scales as n^{-1} when μ is close to 0, and as $n^{-1/2}$ when μ is large. This follows from the fact that $\widehat{\theta}_j$ has an adaptive rate of convergence.

3.6 Uniform model

Suppose $X_1, \ldots, X_n$ are independent real-valued random variables from $U[0, \theta_0]$, the uniform distribution on $[0, \theta_0]$. The maximum likelihood estimator of θ_0 is given by $\widetilde{\theta}_j = \max\{X_i : i \in S_j\}$ which is both mean and median biased. The median bias is 1/2 and hence Algorithm 1 would be inapplicable because it requires infinitely many splits of the data. Note that in this model, $\widetilde{\theta}_j$ converges to θ_0 at an n^{-1} rate.

Interestingly, there are estimators of θ_0 that are median unbiased in this case. For instance, with $\widehat{b}_{S_j}$ and $\widehat{a}_{S_j}$ representing the largest and the second largest values in $\{X_i : i \in S_j\}$, it can be shown that the estimator $\widehat{\theta}_j = 2\widehat{b}_{S_j} - \widehat{a}_{S_j}$ is finite sample median unbiased for θ_0; see Section 3 of Robson and Whitlock (1964) for a proof. Hence, Algorithm 1 can be applied with $\widehat{\theta}_j$ and $\Delta = 0$ to obtain a finite-sample valid confidence set for θ_0. Note that $\widehat{\theta}_j$ also has an n^{-1} rate of convergence. In this case, it is known that the classical bootstrap is invalid, but subsampling works; see e.g., Politis and Romano (1994) and Loh (1984).

The estimator $\widehat{\theta}_j$ described above is approximately median unbiased for a large class of distributions of the form $F(x)/F(\theta_0)$ for $x \in [0, \theta_0]$; see Robson and Whitlock (1964, Section 3). Also, see Hall (1982) for other estimators of θ_0, in a large class of non-parametric distributions, that have a limiting distribution that is symmetric around θ_0.

3.7 Constrained Estimation

Suppose $X_1, \ldots, X_n$ are independent real-valued random variables with mean μ. Consider the estimation of $\theta_0 = \mu \mathbb{1}\{\mu \geqslant 0\}$. We have seen in Section 3.1 how to apply the HULC for μ. Although $\mu \mapsto \mu \mathbb{1}\{\mu \geqslant 0\}$ is a simple transformation, it changes the behavior of many commonly used estimators of θ_0. This is because θ_0 is a non-regular functional and hence, there does not exist any regular estimator for θ_0; this follows from Hirano and Porter (2012, Theorem 2). The implication is that classical Wald confidence intervals based on the estimator $\bar{X}_n \mathbb{1}\{\bar{X}_n \geqslant 0\}$ can fail to cover θ_0 for μ "close" to zero (Robins, 2004, Appendix 1.1). Further, bootstrap and subsampling are also similarly inconsistent; see Fang and Santos (2019, Section 3.2) and Andrews (2000, Section 3) for bootstrap, and Andrews and Guggenberger (2010, Eq. (1)–(2)) for subsampling. It is, however, easy to show that the estimator $\bar{X}_n \mathbb{1}\{\bar{X}_n \geqslant 0\}$ is asymptotically median unbiased for $\theta_0 = \mu \mathbb{1}\{\mu \geqslant 0\}$ because $\bar{X}_n$ is asymptotically median unbiased for μ. This follows simply from the fact that $\kappa(t) = t \mathbb{1}\{t \geqslant 0\}$ is monotonic in μ and hence,

$$\mathbb{1}\{\kappa(\bar{X}_n) \geqslant \kappa(\mu)\} \geqslant \mathbb{1}\{\bar{X}_n \geqslant \mu\} \quad \text{and} \quad \mathbb{1}\{\kappa(\bar{X}_n) \leqslant \kappa(\mu)\} \geqslant \mathbb{1}\{\bar{X}_n \leqslant \mu\}. \tag{34}$$

Note that $\kappa(\cdot)$ is not strictly increasing. This implies that

$$\text{Med-bias}_{\kappa(\mu)}(\kappa(\bar{X}_n)) \;\leqslant\; \text{Med-bias}_{\mu}(\bar{X}_n). \tag{35}$$

Hence, the HULC with the estimator $\kappa(\bar{X}_n)$ and $\Delta = 0$ yields a second-order accurate confidence interval for $\kappa(\mu) = \mu \mathbb{1}\{\mu \geqslant 0\}$.

Inequalities (34) and (35) do not require the specific form of the function $\kappa(\cdot)$. They hold for any monotone function $\kappa(\cdot)$ and, in particular, for any piecewise constant function. Theorem 3.2 of Fang and

Santos (2019) implies that bootstrap is inconsistent unless $\kappa(\cdot)$ is differentiable. This shows the wide range of applicability of our confidence interval. Finally, we note that projection to *any* set on the real line is a monotone function and hence our confidence interval from the HulC is asymptotically valid with the natural estimator that projects the MLE (or any other estimator) to the constraint set. A simple example where this is useful is in the squared mean estimation example of Section 3.5. The estimator $\widehat{\theta}_j$ in (32) is not necessarily non-negative, but its target μ^2 is always non-negative. Using the facts discussed here, we can safely use $\widehat{\theta}_j \mathbb{1}\{\widehat{\theta}_j \geqslant 0\}$ instead of $\widehat{\theta}_j$ in the squared mean estimation example.

3.8 Matching Estimators

In causal inference, matching estimators for the average treatment effect (ATE) are popular, partly because they are intuitive. Under certain regularity conditions, matching estimators are known to be asymptotically Normal centered at the true ATE. Hence, the HulC with $\Delta = 0$ yields a second order accurate confidence interval for ATE. Abadie and Imbens (2008) proved that the bootstrap is inconsistent for matching estimators. They also commented that subsampling can still be used, but given the computational cost of matching, subsampling becomes computationally intensive with larger samples.

3.9 Semiparametric Estimation

In all the examples above, we have cases where the bootstrap and subsampling are either not easily applicable or fail to provide an asymptotically valid confidence interval. There are many cases where all the usual methods apply but the HulC is much simpler and computationally cheaper.

In non- and semi-parametric problems, when a functional of interest can be estimated by an estimator that is asymptotically Normal, two possibilities arise. In the simpler case, the estimator is regular and asymptotically linear with a known (or easily estimable) influence function, while in general the estimator may not have a simple asymptotic expansion. In the first case, we may estimate the asymptotic variance consistently via the sample variance of the (estimated) influence function but obtaining finite-sample guarantees (say via a Berry-Esseen bound) typically requires a case-by-case analysis. More generally however, the variance often involves more nuisance (non-parametric) components than the functional and hence, variance estimation often requires more assumptions or regularity conditions than estimation of the functional. On the other hand, the HulC requires no more nuisance estimation than required for the estimation of the functional. We give three simple examples to illustrate this.

1. **Integral functionals of density.** Consider the estimation of $\theta_0 = \int \phi(f(x), f'(x), \ldots, f^{(k)}(x), x)dx$, when $X_1, \ldots, X_n$ are independent and identically distributed observations from f supported on a compact set. Theorem 2 of Laurent (1997) provides an asymptotically efficient estimator $\widehat{\theta}_n$ for θ_0 that is asymptotically Normal under certain smoothness assumptions on f. The asymptotic variance of $\widehat{\theta}_n$, however, involves higher order ($\geqslant k$) derivatives of f when $k \geqslant 1$. For a more concrete example, consider the Fisher information $\theta_0 = \int_{-\pi}^{\pi} (f'(x))^2/f(x)dx$. The asymptotic variance of the efficient estimator is given by $\int_{-\pi}^{\pi} (2f^{(2)}(x)/f(x) - (f'(x))^2/f(x))^2 f(x)dx - (\int_{-\pi}^{\pi} (f'(x))^2/f(x))dx$, which involves the second derivative of f.

2. **Single Index Model.** Suppose $(X_i, Y_i), 1 \leqslant i \leqslant n$ are independent observations satisfying $\mathbb{E}[Y_i|X_i] = m_0(\theta_0^\top X_i)$ when $m_0(\cdot)$ is an unknown convex function. Consider the least squares estimator $(\widehat{m}, \widehat{\theta})$ which is obtained as a minimizer of $\sum_{i=1}^{n} (Y_i - m(\theta^\top X_i))^2$ over m that is convex, Lipschitz, and θ in

$\{\eta : \|\eta\|_2 = 1\}$. Kuchibhotla et al. (2021) prove that $\widehat{\theta}$ is asymptotically Normal with an asymptotic variance depending on nuisance components such as the conditional mean of X on $\theta_0^\top X$, the derivative of m_0, and the conditional variance of Y given X.

3. **Functionals of Normal Models.** Suppose $X_1, \ldots, X_n$ are independent observations from $N(\mu, \Sigma)$ in the space E (either a Hilbert or a Banach space). Consider the estimation of $\theta_0 = f(\mu)$, if E is Banach or $\theta_0 = f(\mu, \Sigma)$, if E is Hilbert. Koltchinskii and Zhilova (2019, 2021) provide asymptotically efficient estimators of θ_0 which have a Normal limiting distribution. Also, see Koltchinskii (2020). The asymptotic variance is $\langle \Sigma f'(\mu), f'(\mu) \rangle$ if $\theta_0 = f(\mu)$ and is $\|\Sigma^{1/2} f'_\mu(\mu, \Sigma)\|^2 + 2\|\Sigma^{1/2} f'_\Sigma(\mu, \Sigma)\Sigma^{1/2}\|^2_{op}$ if $\theta_0 = f(\mu, \Sigma)$. Estimating the asymptotic variance hence requires estimating more complicated functionals of μ, Σ. Such variance estimation is not discussed in these works.

In all of these cases our approach using the HulC yields a conceptually simpler confidence interval without any additional nuisance estimation.

4 Adaptive HulC

In this section, we provide a method, the Adaptive HulC, to estimate Δ based on subsampling (Politis and Romano, 1994) and consequently, provide a simple method for constructing a valid confidence interval. One might wonder at this point "why not just use subsampling to construct the confidence interval directly?" The answer is that the Adaptive HulC does not require the knowledge of the rate of convergence of the estimator. As an example, in the mean estimation case with fewer than 2 moments, we do not know the rate of convergence a priori. Further, we do not estimate the rate of convergence as suggested in Bertail et al. (1999) for subsampling.

We return now to the univariate parameter setting. Suppose $r_{|S_j|}(\widehat{\theta}_j - \theta_0)$ converges in distribution to W a continuous random variables as $|S_j| \to \infty$. Then it follows that

$$\text{Med-bias}_{\theta_0}(\widehat{\theta}_j) \;\to\; \Delta := \left| \mathbb{P}(W \leqslant 0) - \frac{1}{2} \right|, \quad \text{as} \quad |S_j| \to \infty.$$

Hence, Δ is the asymptotic median bias and can be estimated using subsampling. Let $S_1^{(b)}, \ldots, S_{K_n}^{(b)}$ denote K random subsamples of size $b = b(n)$ and let $\widehat{\theta}_j^{(b)}, 1 \leqslant j \leqslant K_n$ be the estimates based on the subsamples. Let $\widehat{\theta}$ be the estimate based on the full data (of size n). Then Δ can be estimated by

$$\widehat{\Delta}_n := |L_n(0) - 1/2|, \quad \text{where} \quad L_n(0) := \frac{1}{K_n} \sum_{j=1}^{K_n} \mathbb{1}\{\widehat{\theta}_j^{(b)} - \widehat{\theta} \leqslant 0\}.$$

Given this estimator $\widehat{\Delta}_n$, we can estimate the miscoverage probability $P(B; \Delta)$ in (3) of the convex hull of B estimators by $P(B; \widehat{\Delta}_n)$. The results of Politis and Romano (1994) imply that $\widehat{\Delta}_n$ is (asymptotically) consistent for Δ (see also, our Lemma 3, which develops finite-sample bounds) and hence, $B_{\alpha,\Delta} = B_{\alpha,\widehat{\Delta}_n}$ for large enough n; see Proposition 1. Therefore, the convex hull based on $B_{\alpha,\widehat{\Delta}_n}$ estimators has an asymptotic miscoverage probability of at most α. To avoid conservativeness, one can randomize the number of estimators between $B_{\alpha,\widehat{\Delta}_n}$ and $B_{\alpha,\widehat{\Delta}_n} - 1$ to attain asymptotically exact coverage as shown in Algorithm 2.

We now prove bounds on the miscoverage probabilities of the confidence intervals of the Adaptive HulC. The first result provides a bound without using the fact that $\widehat{\Delta}_n$ is obtained from subsampling and

Algorithm 2: Adaptive Confidence Interval with Unknown Median Bias (Adaptive HulC)

Input: data $X_1, \ldots, X_n$, coverage probability $1-\alpha$, and an estimation procedure $\mathcal{A}(\cdot)$ that takes as input observations and returns an estimator, subsample size b, number of subsamples K_n.

Output: A confidence interval $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$ such that $\mathbb{P}(\theta_0 \in \widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}) \geqslant 1-\alpha$ (asymptotically).

1 Draw K_n many subsamples of size b from $X_1, \ldots, X_n$. Apply $\mathcal{A}(\cdot)$ for each subsample and obtain estimators $\widehat{\theta}_b^{(j)}, 1 \leqslant j \leqslant K_n$.

2 Compute the estimator of the (asymptotic) median bias of $\mathcal{A}(\cdot)$ as $\widehat{\Delta}_n := |L_n(0) - 1/2|$.

3 Use Algorithm 1 with input data $X_1, \ldots, X_n$, coverage probability $1-\alpha$, the value $\widehat{\Delta}_n$ and the estimation procedure $\mathcal{A}(\cdot)$

4 **return** the confidence interval obtained as output from Algorithm 1 as $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$.

then using distributional convergence assumptions, we obtain the final miscoverage bound for $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$. Define $\Delta_{n,\alpha}$ as the median bias of $\mathcal{A}(\{X_i : i \in S\})$ with $n/(2B_{\alpha,\Delta}) \leqslant |S| \leqslant 2n/B_{\alpha,\Delta}$, i.e.,

$$\Delta_{n,\alpha} := \max_{1/2 \leqslant B_{\alpha,\Delta}|S|/n \leqslant 2} \text{Med-bias}_{\theta_0}(\mathcal{A}(\{X_i : i \in S\})).$$

Consider the following assumption:

(A1) There exists a random variable W and a decreasing sequence $\{\delta_m\}_{m \geqslant 1}$ converging to zero such that the estimator $\widehat{\theta}^{(m)}$ obtained by applying $\mathcal{A}(\cdot)$ on m observations satisfies

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}(r_m(\widehat{\theta}^{(m)} - \theta_0) \leqslant t) - \mathbb{P}(W \leqslant t) \right| \leqslant \delta_m.$$

Further, $0 < \mathbb{P}(W \leqslant 0) < 1$.

Define the asymptotic median bias of the estimation procedure $\mathcal{A}(\cdot)$ as $\Delta := (1/2 - \max\{\mathbb{P}(W \leqslant 0), \mathbb{P}(W \geqslant 0)\})_+$. For any $\alpha \in (0,1)$ and $\Delta \in (0, 1/2)$, define

$$C_{\alpha,\Delta} := \frac{1}{2}\left[\min\left\{\left(\frac{\alpha}{P(B_{\alpha,\Delta};\Delta)}\right)^{1/B_{\alpha,\Delta}}, \left(\frac{P(B_{\alpha,\Delta}-1;\Delta)}{\alpha}\right)^{1/B_{\alpha,\Delta}}\right\} - 1\right],$$

and for $\Delta = 0$,

$$C_{\alpha,0} := \frac{2}{B_{\alpha,0}(B_{\alpha,0}-1)}\left[\frac{\alpha}{P(B_{\alpha,0};0)} - 1\right].$$

These quantities are taken from Proposition 1 which implies that for any $\gamma \in (0, 1/2)$, if $|\gamma - \Delta| \leqslant C_{\alpha,\Delta}$, then $B_{\alpha,\gamma} = B_{\alpha,\Delta}$. Finally, recall that $\widehat{\mathrm{CI}}_{\alpha,\Delta}$ represents the confidence interval returned by the HulC when the median bias parameter is chosen to be Δ (irrespective of what the true finite sample median bias is).

To succinctly state our next result we define some additional quantities. Given an estimate $\widehat{\Delta}_n$ we compute the number of splits, $B_{\alpha,\widehat{\Delta}_n}$. We then hypothesize splitting the data twice into $B_{\alpha,\widehat{\Delta}_n}$ and $B_{\alpha,\widehat{\Delta}_n} - 1$ parts with approximately equal number of observations in each split. We then define,

$$\widehat{\mathrm{CI}}_\alpha^{(0)} := \left[\min_{1 \leqslant j \leqslant B_{\alpha,\widehat{\Delta}_n}-1} \widehat{\theta}_j, \max_{1 \leqslant j \leqslant B_{\alpha,\widehat{\Delta}_n}-1} \widehat{\theta}_j\right], \quad \text{and} \quad \widehat{\mathrm{CI}}_\alpha^{(1)} := \left[\min_{1 \leqslant j \leqslant B_{\alpha,\widehat{\Delta}_n}} \widehat{\theta}_j, \max_{1 \leqslant j \leqslant B_{\alpha,\widehat{\Delta}_n}} \widehat{\theta}_j\right].$$

Here $\widehat{\theta}_j$ are estimators computed based on $\mathcal{A}(\cdot)$. We have the following result:

Theorem 3. *Suppose the random variables $X_1, \ldots, X_n$ are independent and assumption* **(A1)** *holds true. Then for any $\alpha \in (0,1)$, the* Adaptive HulC *confidence intervals satisfy*

$$\max\left\{\frac{\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{(0)})}{2}, \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{(1)})\right\} \leqslant \mathbb{P}(B_{\alpha,\hat{\Delta}_n} \neq B_{\alpha,\Delta}) + \alpha \times \begin{cases} (1 + B_{\alpha,0}(B_{\alpha,0}-1)\Delta_{n,\alpha}^2/2), & \text{if } \Delta = 0, \\ (1+2|\Delta_{n,\alpha}-\Delta|)^{B_{\alpha,\Delta}}, & \text{if } \Delta \neq 0 \end{cases}, \tag{36}$$

and for any $0 \leqslant \eta \leqslant C_{\alpha,\Delta}$,

$$\left|\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}) - \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta})\right| \leqslant 2\mathbb{P}(|\hat{\Delta}_n - \Delta| \geqslant \eta) + 3\alpha \times \begin{cases} \eta^2 B_{\alpha,0}^2(1 + B_{\alpha,0}^2\Delta_{n,\alpha}^2/2)/2, & \text{if } \Delta = 0, \\ 2\sqrt{e}\eta B_{\alpha,\Delta}(1+2|\Delta_{n,\alpha}-\Delta|)^{B_{\alpha,\Delta}}/(1/2-\Delta), & \text{if } \Delta \neq 0. \end{cases} \tag{37}$$

Theorem 3 (proved in Section S.8) provides a bound on miscoverage of the confidence interval $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$ from Algorithm 2 but does not assume that $\hat{\Delta}_n$ is obtained from subsampling. The miscoverage probabilities of confidence intervals obtained from non-random choices of number of splits $\widehat{\mathrm{CI}}_\alpha^{(0)}$ and $\widehat{\mathrm{CI}}_\alpha^{(1)}$ only requires controlling the probability of $B_{\alpha,\hat{\Delta}_n} \neq B_{\alpha,\Delta}$. From Proposition 1, it follows that we do not need $\hat{\Delta}_n$ to be consistent for Δ. Note that the second term in (36) only depends on how close $\Delta_{n,\alpha}$ to Δ is.

For the miscoverage probability of $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$ that randomizes the number of splits to avoid overcoverage, we require consistency of $\hat{\Delta}_n$ to Δ. If $\hat{\Delta}_n$ is obtained from an independent sample, then we would not require such consistency and can apply Theorem 2 to prove miscoverage. Regarding inequality (37), we recall from Theorem 2 (and Remark 2.2) that $\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta})$ can be upper and lower bounded by quantities close to α. Such lower bounds do not hold true for $\widehat{\mathrm{CI}}_\alpha^{(0)}$ and $\widehat{\mathrm{CI}}_\alpha^{(1)}$. Finally, because $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$ is a random selection of one of $\widehat{\mathrm{CI}}_\alpha^{(0)}$ and $\widehat{\mathrm{CI}}_\alpha^{(1)}$, we get

$$\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}) \leqslant \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{(0)}) + \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{(1)}),$$

and inequality (36) can be used to imply that $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$ has an approximate miscoverage probability of 3α when $B_{\alpha,\hat{\Delta}_n} = B_{\alpha,\Delta}$ holds with high probability.

In the multivariate case, one can apply union bound directly on (36) with α replaced by α/d to obtain a bound on miscoverage. But using the proof, one can refine this by replacing $\mathbb{P}(B_{\alpha,\hat{\Delta}_n} \neq B_{\alpha,\Delta})$ by $\mathbb{P}(B_{\alpha,\hat{\Delta}_n^{(k)}} \neq B_{\alpha,\Delta^{(k)}}$ for any $1 \leqslant k \leqslant d$). Here $\Delta^{(k)}$ is the limiting median bias of the estimator of k-th coordinate of θ_0 and $\hat{\Delta}_n^{(k)}$ is its estimator. Because the second term on the right hand side of (36) is multiplicative in α, a union bound can be safely applied to obtain a non-trivial guarantee, as in Section 2.4. Similarly, one can replace the first term on the right hand side of (37) with $\mathbb{P}(|\hat{\Delta}_n^{(k)} - \Delta^{(k)}| \geqslant \eta$ for any $1 \leqslant k \leqslant d$). The second term being multiplicative in α does not effect the applicability of a union bound to obtain a non-trivial bound.

Inequality (37) holds true for all $\eta \in [0, C_{\alpha,\Delta}]$. With $\hat{\Delta}_n$ a consistent estimator for Δ, one can take η converging to zero with sample size. In the following, we will prove a bound on $\mathbb{P}(|\hat{\Delta}_n - \Delta| \geqslant \eta)$ when $\hat{\Delta}_n$ is obtained using subsampling (as in Algorithm 2). It is worth emphasizing that any method of estimating Δ can be used in Theorem 3.

(A2) There exists $r^* > 0$ and $\mathfrak{C} < \infty$ such that the distribution function of W satisfies

$$0 \leqslant \frac{\mathbb{P}(W \leqslant t) - \mathbb{P}(W \leqslant -t)}{t} \leqslant \mathfrak{C}, \quad \text{for all} \quad 0 \leqslant t \leqslant r^*.$$

(A3) The subsample size b satisfies $b/n \to 0$ and $r_b/r_n \to 0$ as $n \to \infty$. Further, the number of subsamples diverges: $K_n \to \infty$.

These assumptions are similar to those used in the analysis of subsampling. In contrast to the classical analysis of subsampling by Politis and Romano (1994) we provide a finite sample analysis.

Lemma 3. *Fix any $t > 0$ such that $r_b t/r_n \leqslant r^*$, then under assumptions* **(A1)**, **(A2)**, *and* **(A3)** *with probability at least $1 - 2\delta_n - (b+1)/n - \mathbb{P}(|W| > t)$,*

$$|\widehat{\Delta}_n - \Delta| \;\leqslant\; \sqrt{\frac{\log(2n)}{2K_n}} + \sqrt{\frac{\log(2n/b)}{2[n/b]}} + 2\delta_b + 2\mathfrak{C}\frac{r_b t}{r_n}.$$

The proof follows a similar structure to that of Politis and Romano (1994) and appears in Section S.9.

Choosing $t \to \infty$ in Lemma 3 such that $r_b t/r_n \to 0$ as $n \to \infty$, we conclude that $|\widehat{\Delta}_n - \Delta| = o_p(1)$; for example, one can take $t = \sqrt{r_n/r_b}$. This combined with Theorem 3 implies that Adaptive HulC yields an asymptotically valid confidence interval for θ_0 under assumptions **(A1)**, **(A2)**, and **(A3)**.

5 Applications to non-standard problems

Many commonly used estimators are derived from classical parametric and semi-parametric efficiency theory and have an asymptotic Normal distribution with zero mean. This implies that these estimators have an asymptotic median bias of zero, making them "standard" problems and allowing for the application of the HulC with $\Delta = 0$. There do exist estimators that have a non-standard rate of convergence and a non-standard limiting distribution. In this section, we discuss three "non-standard" examples where either the rate of convergence or the limiting distribution or both are unknown in practice. With the rate of convergence unknown, subsampling does not readily apply to yield a confidence interval; one needs to estimate the rate of convergence as in Bertail et al. (1999).

5.1 Heavy-tailed mean estimation

In Section 3.1, we discussed the application of the HulC in the context of mean estimation when the limiting distribution is symmetric around zero. When the random variables $X_1, \ldots, X_n$ do not have a finite second moment, then the limiting distribution of the sample mean $\bar{X}_n$ can be asymmetric around the population mean μ with the amount of asymmetry depending on the tail decay on either side of μ. In this case, the rate of convergence also depends on tail decay and is unknown a priori, which makes subsampling inapplicable. See Romano and Wolf (1999) for an application of subsampling using the studentized statistic, which does not require estimating the rate of convergence. Without knowing the rate of convergence, we can apply Algorithm 2 to obtain an estimate of the median bias and create a confidence interval for the population mean. In this case, provided that the median bias is not too close to 1/2 the Adaptive HulC will yield non-trivial confidence intervals.

5.2 Shape constrained regression

Constructing confidence intervals in the context of general non-parametric regression is a difficult task. In order to obtain optimal estimation rates we aim to explicitly balance (squared) bias and variance. On the other hand, the exact bias is often intractable and difficult to account for in confidence interval construction. As a consequence, often under-smoothing is used to ensure that the squared bias is negligible compared to the variance asymptotically. In practice, however, under-smoothing can be sensitive to the precise choice of tuning parameters.

If the conditional mean function is assumed to satisfy a shape constraint such as monotonicity or convexity, then the least squares estimator of the conditional mean has negligible bias uncomplicating the inference problem. However, the rate of convergence and the limiting distribution depends on the local smoothness of the conditional mean. To be concrete, consider the setting of univariate monotone regression with equi-spaced design, i.e., $Y_i = f_0(i/n) + \varepsilon_i$ where ε_is are independent and identically distributed with mean zero and finite variance $\sigma^2 > 0$, and f_0 is our shape constrained target. Consider the least squares estimator (LSE) of f_0 as

$$\widehat{f}_n \;:=\; \operatorname*{arg\,min}_{f:\,\text{increasing}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(i/n))^2.$$

Note that $\widehat{f}_n(\cdot)$ is defined uniquely only at $i/n, 1 \leqslant i \leqslant n$, and is, conventionally, defined to as a piecewise constant increasing function on $[0,1]$. In this setting, for any $t \in (0,1)$ such that $f_0(\cdot)$ has a positive continuous derivative on some neighborhood of t, the LSE satisfies $n^{1/3}(\widehat{f}_n(t) - f_0(t)) \xrightarrow{d} [4\sigma^2 f_0'(t)]^{1/3}\mathbb{C}$, where $\mathbb{C} = \arg\min_{h\in\mathbb{R}}\{\mathbb{W}(h) + h^2\}$ has Chernoff's distribution (here $\mathbb{W}(\cdot)$ is a two-sided Brownian motion starting from 0). It is important here that $f_0'(t) \neq 0$. If $f_0^{(j)}(t) = 0$ for $1 \leqslant j \leqslant p-1$ and $f_0^{(p)}(t) \neq 0$ (for $p \geqslant 1$), where $f_0^{(j)}$ denotes the j-th derivative, then $n^{p/(2p+1)}(\widehat{f}_n(t) - f_0(t))$ converges to a non-degenerate distribution depending on $f_0^{(p)}(t)$ and σ^2. Finally, if $f_0(\cdot)$ is flat at t, then the rate of convergence becomes $n^{1/2}$. These results are known in both the fixed and random design settings (see for instance, Wright (1981); Durot (2008); Guntuboyina and Sen (2018)). These rates of convergences imply that the LSE admits an adaptive behavior and for arbitrary monotone functions and consequently it is unclear how to perform inference. The situation becomes more complicated in the multi-dimensional case where the limiting distribution depends on the anisotropic smoothness of f_0. Recently Deng et al. (2020b) proved that the rates of convergence along with the nuisance parameters in the limiting distributions can be estimated consistently using $\widehat{f}_n$. This theory requires substantial new techniques and still requires estimation of σ^2. Alternatively, we can use the Adaptive HulC in all of these cases to obtain asymptotically valid confidence intervals. It is worth mentioning that in most of these settings, the median bias of the limiting distribution is also unknown because it depends on the unknown local smoothness. The same discussion also holds true for other shape constrained models such as convex regression and current status regression; see Guntuboyina and Sen (2018, Section 4) and Deng et al. (2020a) for details.

We note that for shape constrained regression, it is possible to obtain confidence bands from confidence intervals at several points on the domain. For example if we know $\ell(t_1) \leqslant f_0(t_1) \leqslant u(t_1)$ and $\ell(t_2) \leqslant f_0(t_2) \leqslant u(t_2)$ for two points $t_1, t_2 \in [0,1]$ in the domain, then using the information $f_0(\cdot)$ is non-decreasing we can

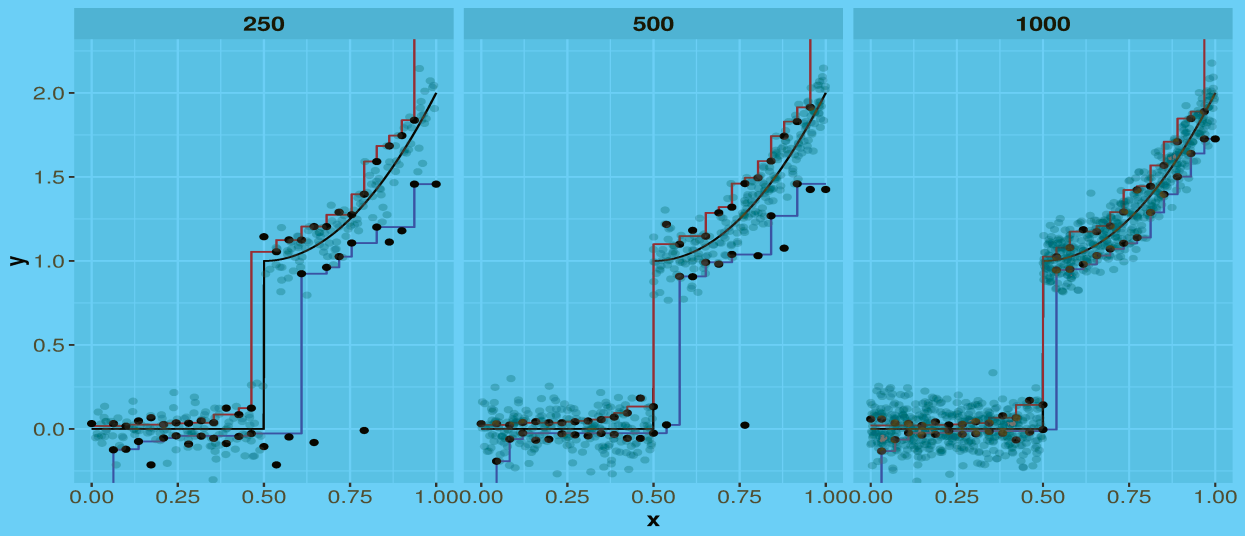


Figure 6: Confidence bands for a monotone conditional mean function as sample size changes from 250 to 1000. The black line shows the true function which is a constant on $[0, 0.5]$ and is a (strictly increasing) quadratic on $[0.5, 1]$. The LSE attains an $n^{1/2}$ rate on $[0, 0.5]$ and an $n^{1/3}$ rate on $[0.5, 1]$. The Adaptive HulC simultaneous confidence intervals at 25 equi-spaced points on $[n^{-1/2}, 1 - n^{-1/2}]$ are shown as dark black points. The confidence band obtained via (38) is shown as red and blue lines. The obtained sample for each sample size is plotted in gray.

conclude that $\bar{\ell}(t) \leqslant f_0(t) \leqslant \bar{u}(t)$ for all $t \in [0, 1]$, where

$$\bar{\ell}(t) = \begin{cases} -\infty, & \text{for } t < t_1, \\ \ell(t_1), & \text{for } t_1 \leqslant t < t_2, \\ \ell(t_2), & \text{for } t_2 \leqslant t \leqslant 1, \end{cases} \quad \text{and} \quad \bar{u}(t) = \begin{cases} u(t_1), & \text{for } t \leqslant t_1, \\ u(t_2), & \text{for } t_1 < t \leqslant t_2, \\ \infty, & \text{for } t_2 < t \leqslant 1. \end{cases} \tag{38}$$

Of course, the more points at which confidence intervals are available, the better the confidence band is. Figure 6 shows the simultaneous confidence intervals obtained from the Adaptive HulC (with $b = n^{2/3}$ and $K_n = 1000$) for a monotone conditional mean from observations $Y_i = f_0(X_i) + \varepsilon_i$ where $X_i \sim \text{Unif}[0, 1], \varepsilon_i \sim N(0, 0.1^2)$ and $f_0(x) = 1 - \mathbb{1}\{x \leqslant 0.5\} + ((x - 0.5)/0.5)^2 \mathbb{1}\{x > 0.5\}$. This figure only shows one replication of the experiment and suggests that the width of the band seems to adapt to the local smoothness of f_0.

5.3 Nonparametric Regression and Forests

The HulC also yields confidence intervals for nonparametric regression even in the presence of unknown asymptotic bias. We briefly sketch the main ideas, deferring most of the details to future work. We focus on constructing a confidence interval for the non-parametric regression function f_0 at a fixed point $x_0 \in \mathbb{R}$. For example, let $\widehat{f}_n(x_0)$ be a kernel regression estimator with bandwidth h. If $h = h_n$ is chosen to balance bias and variance then $\sqrt{nh_n}(\widehat{f}_n(x_0) - f_0(x_0))$ converges to a Gaussian law with mean $Q = \lim_{n\to\infty} \sqrt{nh_n}\mathbb{E}[\widehat{f}_n(x_0) - f_0(x_0)]$ (which is the asymptotic bias), and finite, non-zero variance. As we discussed earlier, classical methods often rely on undersmoothing to ensure that $Q = 0$. However, the Adaptive HulC works as long as Q is finite, since in this case the asymptotic median bias is bounded away from $1/2$. We also emphasize that, in contrast to undersmoothing for which there are relatively few guidelines on practical implementation,

it is more conventional in non-parametric regression to balance (squared) bias and variance, and in many cases cross-validation methods yield tuning parameters which achieve this balancing under various conditions (see for instance, Theorem 2.2 in Li and Racine (2004)).

The argument above is not specific to kernel regression estimators. More generally, let $\widehat{f}_n(x_0)$ be a complicated nonparametric estimator such as a random forest. The Adaptive HulC yields a valid interval for $f_0(x_0)$ provided that we are able to balance (squared) bias and variance, i.e. so long as we can ensure that for some possibly unknown r_n, we have that $r_n(\widehat{f}_n(x_0) - f_0(x_0))$ converges to a Gaussian law with possibly non-zero (but finite) mean, and non-zero, finite variance.

6 Confidence Regions under Unimodality

In previous sections, we have considered the construction of confidence intervals based on the median bias of the estimation procedure. In some cases, the estimation procedure has a large median bias close to 1/2. For example in the mean square estimation problem, $\bar{X}_n^2$ has a median bias of 1/2 when $\mu = 0$ and in the uniform model, the MLE has a median bias of 1/2. In these cases, the HulC and Adaptive HulC are not useful because they would require infinite splits of the data. Interestingly, in these examples, the limiting distribution of the estimation procedure is unimodal at the true parameter. In the univariate setting, unimodality at θ_0 means that the distribution function is convex for $t \leqslant \theta_0$ and concave for $t \geqslant \theta_0$. It is important to note that unimodality of a random variable is a global property of the distribution function unlike median bias, which is a local property.

Using the results of Lanke (1974), we can construct a confidence interval based on unimodality of the estimation procedure. The resulting confidence interval is very similar to the one from the HulC. The Unimodal HulC method is presented in Algorithm 3.

The Unimodal HulC can be seen as a generalization of the HulC where we also use the unimodality of estimators, if available. Taking $t = 0$ in the Unimodal HulC gives exactly the HulC. The confidence interval with $t = 0$ need not have coverage validity if the asymptotic median bias is 1/2 and by taking $t > 0$, we get asymptotic coverage when the limiting distribution is unimodal even if the asymptotic median bias is 1/2. In the Unimodal HulC, we assume that the limiting median bias Δ is known, but one can always substitute $\Delta = 1/2$ if median bias is unknown; recall that $P(B; 1/2) = 1$ for all B. Alternatively, one can use the subsampling approach from Section 4 to replace Δ with the subsampling estimator. We leave it to future work to derive a final miscoverage bound for this subsampling-based procedure.

The following theorem (proved in Section S.10) shows that the confidence interval returned by the Unimodal HulC has a miscoverage probability bounded asymptotically by α.

Theorem 4. *Suppose the estimators $\widehat{\theta}_j$ in the* Unimodal HulC *are independent and are constructed based on approximately equal sized samples. Further, suppose the estimators are continuously distributed and satisfy*

$$\sup_{u \in \mathbb{R}} |\mathbb{P}(r_{n,\alpha}(\widehat{\theta}_j - \theta_0) \leqslant u) - \mathbb{P}(W \leqslant u)| \;\leqslant\; \delta_{n,\alpha},$$

for some sequence $\{r_{n,\alpha}\}_{n \geqslant 1}$ and a continuous random variable W that is unimodal at 0 and has a median bias of Δ (i.e., $\Delta = |1/2 - \mathbb{P}(W \leqslant 0)|$). Then for all $t \geqslant 0$, $\Delta \in [0, 1/2]$, and $\alpha \in [0, 1]$,

$$\mathbb{P}\left(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{\mathrm{mode}}\right) \;\leqslant\; \alpha\,(1 - 10B_{\alpha,t,\Delta}(1+t)\delta_{n,\alpha})_+^{-1}. \tag{39}$$

Algorithm 3: Confidence Interval based on Unimodality and Median Bias (UNIMODAL HULC)

Input: data $X_1, \ldots, X_n$, coverage probability $1-\alpha$, a parameter $t > 0$ and an estimation procedure $\mathcal{A}(\cdot)$ that yields estimators asymptotically unimodal at θ_0 and have an asymptotic median bias of $\Delta \in [0, 1/2]$.

Output: A confidence interval $\widehat{\mathrm{CI}}_\alpha^{\mathrm{mode}}$ such that $\mathbb{P}(\theta_0 \in \widehat{\mathrm{CI}}_\alpha^{\mathrm{mode}}) \geqslant 1-\alpha$ asymptotically.

1 Set $Q(B; t, \Delta) := P(B; \Delta)(1+t)^{-B+1}$, and find the smallest integer $B = B_{\alpha,t,\Delta} \geqslant 1$ such that $Q(B; t, \Delta) \leqslant \alpha$. Recall $P(B; \Delta)$ from (3).

2 Generate a random variable U from Uniform distribution on $[0, 1]$ and set

$$\eta_{\alpha,t} := \frac{Q(B_{\alpha,t,\Delta} - 1; t, \Delta) - \alpha}{Q(B_{\alpha,t,\Delta} - 1; t, \Delta) - Q(B_{\alpha,t,\Delta}; t, \Delta)} \quad \text{and} \quad B^* := \begin{cases} B_{\alpha,t,\Delta}, & \text{if } U \leqslant \eta_{\alpha,t}, \\ B_{\alpha,t,\Delta} - 1, & \text{if } U > \eta_{\alpha,t}. \end{cases}$$

3 Randomly split the data $X_1, \ldots, X_n$ into B^* many disjoint sets $\{\{X_i : i \in S_j\} : 1 \leqslant j \leqslant B^*\}$ of approximately equal sizes.

4 Compute estimators $\widehat{\theta}_j := \mathcal{A}(\{X_i : i \in S_j\})$, for $1 \leqslant j \leqslant B^*$ and set

$$\widehat{\theta}_{\max} := \max_{1 \leqslant j \leqslant B^*} \widehat{\theta}_j, \quad \text{and} \quad \widehat{\theta}_{\min} := \min_{1 \leqslant j \leqslant B^*} \widehat{\theta}_j.$$

5 **return** the confidence interval $\widehat{\mathrm{CI}}_\alpha^{\mathrm{mode}} := [\widehat{\theta}_{\min} - t(\widehat{\theta}_{\max} - \widehat{\theta}_{\min}),\ \widehat{\theta}_{\max} + t(\widehat{\theta}_{\max} - \widehat{\theta}_{\min})]$.

Similar to Theorem 2, Theorem 4 shows that the confidence interval from the UNIMODAL HULC attains the required miscoverage probability up to a multiplicative error. Once again, this is unlike the coverage guarantee for Wald's interval. Because of the multiplicative error, the guarantee from Theorem 4 is also suitable for an application of the union bound to obtain a valid multivariate confidence region, as discussed previously in Section 2.4.

Note that the right hand side of (39) is finite if and only if $10B_{\alpha,t,\Delta}(1+t)\delta_{n,\alpha} < 1$. It is easy to prove that $B_{\alpha,t,\Delta} = O(\log(1/\alpha))$ when either $t > 0$ or $\Delta < 1/2$ and in many cases, $\delta_{n,\alpha} = O(\sqrt{\log(1/\alpha)/n})$. Hence, the condition for finiteness would hold true as long as $n \gg \log^3(1/\alpha)$; this is similar to the requirement in the HULC. The importance of the UNIMODAL HULC stems from its ability to tackle problems where the median bias of the estimator is large (near 1/2).

The width of the confidence interval $\widehat{\mathrm{CI}}_\alpha^{\mathrm{mode}}$ is given by $(1+2t)(\widehat{\theta}_{\max} - \widehat{\theta}_{\min})$. This is $1+2t$ times larger than the width of confidence interval from the HULC. The confidence interval has a coverage for any parameter $t \geqslant 0$ and as t increases, the number of splits B in the UNIMODAL HULC decreases leading to a smaller value of $\widehat{\theta}_{\max} - \widehat{\theta}_{\min}$. Similar to the map $\Delta \mapsto B_{\alpha,\Delta}$, the map $(t, \Delta) \mapsto B_{\alpha,t,\Delta}$ is a piecewise constant function.

6.1 Application 1: standard problems

The UNIMODAL HULC can make use of both asymptotic unimodality and asymptotic median unbiasedness which holds true for most of the standard problems where the limiting distribution is Gaussian (a symmetric unimodal distribution). In many of the examples discussed in Section 3, one can use the UNIMODAL HULC to (potentially) obtain a tighter confidence interval. Note that $B_{\alpha,t,\Delta}$ in the UNIMODAL HULC is always smaller

than $B_{\alpha,\Delta} = B_{\alpha,0,\Delta}$ in the HulC. Once again the advantage is that we do not need to estimate the limiting variance of the estimators being used and need not know the rate of convergence.

6.2 Application 2: shape constrained regression (revisited)

In Section 5.2, we used subsampling to estimate the median bias of the LSE in shape constrained regression. Experimentally, we found that the distribution of the LSE is unimodal at the true value. Consider the regression problem $Y_i = f_0(i/n) + \varepsilon_i$ where $\varepsilon_i \sim N(0,1)$ and $f_0(x) \equiv 0$. The histograms of the LSE error $\widehat{f}_n(x_0) - f_0(x_0)$ at $x_0 = 0.25, 0.75$ are shown in Figure 7 when the estimator is computed based on 10^6 samples and 1000 replications. We are not aware of a result proving unimodality of the limiting distribution of the

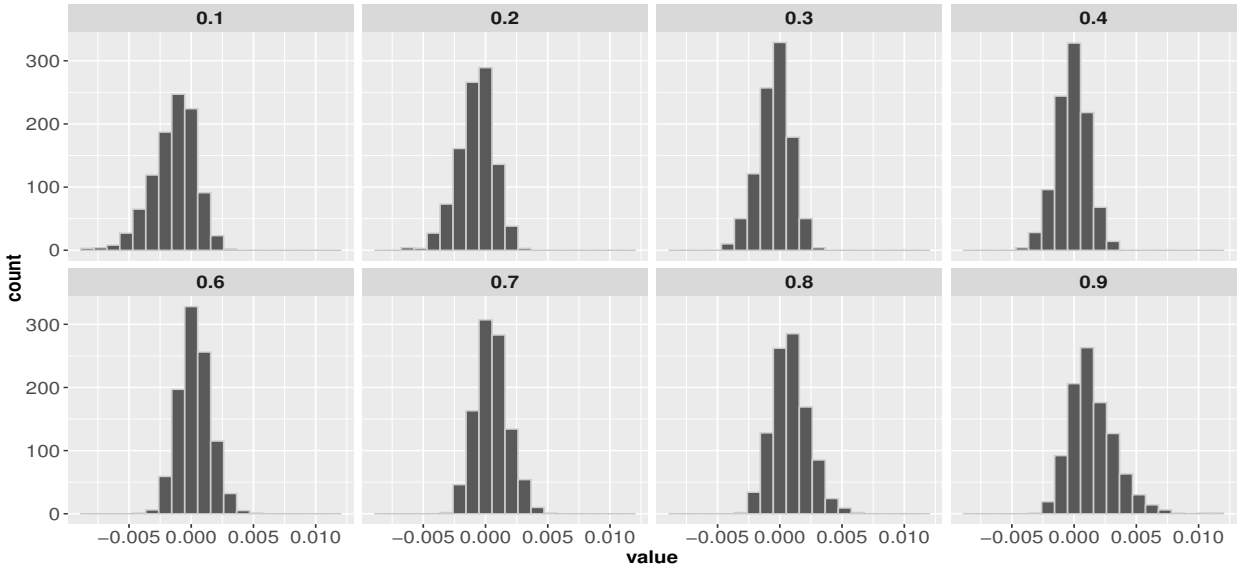


Figure 7: Histogram of monotone LSE at $x_0 = 0.25, 0.75$ when the true function is identically zero. Clearly, the distributions are asymmetric but the mode is at the right place (0). The distribution at 0.25 is left skewed and the one at 0.75 is right skewed. The farther we move from the center of the support $[0,1]$, the more asymmetric the distribution becomes.

LSE in general (when higher derivatives of f_0 may vanish at x_0). But motivated by our experimental results, we apply the Unimodal HulC to construct a confidence band, and leave a more rigorous investigation of its validity to future work.

The performance of the Unimodal HulC for monotone regression is shown in Figure 8. It shows adaptation and shows higher uncertainty around the change point. The confidence band here is noticeably larger than the one from the Adaptive HulC.

7 Conclusions and Future Directions

In this paper, we developed and analyzed a simple and broadly applicable method, the HulC for constructing confidence sets, using the convex hull of estimates constructed on independent subsamples of the data. The HulC bypasses the difficult problem of estimating nuisance components in the limiting distribution, requires fewer regularity conditions than the bootstrap and unlike subsampling does not require knowledge of the

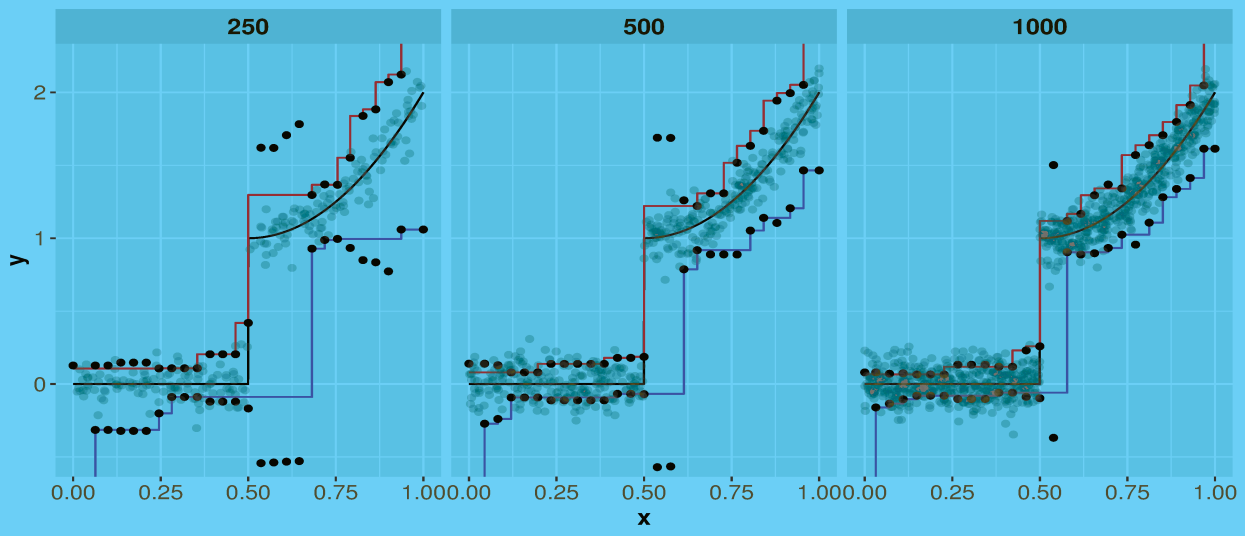


Figure 8: Performance of the UNIMODAL HULC (with $t = 1/2$ and $\Delta = 1/2$) as sample size increases. The black line is the true monotone function which is a constant 0 on $[0, 0.5]$ and is smooth on $[0.5, 1]$. The confidence intervals using the UNIMODAL HULC along with union bound at 25 equi-spaced points on $[n^{-1/2}, 1 - n^{-1/2}]$ are shown in black points. The confidence bands from these simultaneous confidence intervals using (38) are shown in red and blue.

rate of convergence of the underlying estimates on which it is based. These advantages, in many cases, come at a surprisingly small price in the width of the interval. The width of the intervals are determined in general by the accuracy of the underlying estimators, as well as their median bias. We also present two variants, the ADAPTIVE HULC which estimates the median bias using subsampling, and the UNIMODAL HULC which can be useful even in cases when the median bias is large so long as the limiting distribution is unimodal. Beyond these methodological contributions, we also studied several challenging confidence set construction problems and showed how our methods can often provide simple solutions to these problems.

From a computational standpoint, the HULC only requires computing the estimator B times where B is typically around 5 or 10, and so is less computationally intensive than the bootstrap. In cases where the underlying estimator has computational complexity which is super-linear in the number of samples, computing B estimates on n/B samples can in fact be cheaper than computing a single estimate on the whole dataset.

Intuitively, the HULC is also quite robust in the sense that we only use qualitative properties, such as an upper bound on the median bias, of the limiting distribution rather than its exact form. In finite samples, the distribution of a statistic might be close to symmetric even if it does not resemble a Gaussian distribution.

Our analysis has not discussed the important problem of obtaining confidence intervals with uniform coverage. This is especially important in irregular problems where the rate of convergence and limiting distribution can vary across the parameter space. Developing this understanding is an important open problem that we plan to address in future work.

References

Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557.

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, pages 399–405.

Andrews, D. W. and Guggenberger, P. (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory*, pages 426–468.

Andrews, D. W. and Phillips, P. C. (1987). Best median-unbiased estimation in linear regression with bounded asymmetric loss functions. *Journal of the American Statistical Association*, 82(399):886–893.

Athreya, K. (1987). Bootstrap of the mean in the infinite variance case. *The Annals of Statistics*, pages 724–731.

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018). High-dimensional econometrics and regularized gmm. *arXiv preprint arXiv:1806.01888*.

Bentkus, V. (2005). A lyapunov-type bound in r^d. *Theory of Probability & Its Applications*, 49(2):311–323.

Bentkus, V., Bloznelis, M., and Götze, F. (1997). A berry–esséen bound for m-estimators. *Scandinavian journal of statistics*, 24(4):485–502.

Bertail, P. and Politis, D. N. (2001). Extrapolation of subsampling distribution estimators: The iid and strong mixing cases. *Canadian Journal of Statistics*, 29(4):667–680.

Bertail, P., Politis, D. N., and Romano, J. P. (1999). On subsampling estimators with unknown rate of convergence. *Journal of the American Statistical Association*, 94(446):569–579.

Bickel, P., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Bickel, P. J. (1982). On Adaptive Estimation. *The Annals of Statistics*, 10(3):647 – 671.

Birnbaum, A. (1964). Median-unbiased estimators. *Bulletin of Mathematical Statistics*, 11(1):25–34.

Borges, R. (1970). Eine approximation der binomialverteilung durch die normalverteilung der ordnung 1/n. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 14(3):189–199.

Borges, R. (1971). Derivation of normalizing transformations with an error of order 1/n. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 441–460.

Boucheron, S. and Thomas, M. (2012). Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17.

Breiman, L. (1968). *Probability*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical science*, pages 101–117.

Brown, L. D., Cai, T. T., and Dasgupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*, 30(1):160–201.

Brown, L. D., Cohen, A., and Strawderman, W. E. (1976). A complete class theorem for strict monotone likelihood ratio with applications. *The Annals of Statistics*, 4(4):712–722.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019). Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544.

Cabrera, J. and Watson, G. (1997). Simulation methods for mean and median bias reduction in parametric estimation. *Journal of statistical planning and inference*, 57(1):143–152.

Chakravarti, P., Balakrishnan, S., and Wasserman, L. (2019). Gaussian mixture clustering using relative tests of fit.

Chernozhukov, V., Chetverikov, D., and Koike, Y. (2020). Nearly optimal central limit theorem and bootstrap approximations in high dimensions. *arXiv preprint arXiv:2012.09513*.

Das, D. and Lahiri, S. (2020). Central limit theorem in high dimensions: The optimal bound on dimension growth rate. *arXiv preprint arXiv:2008.04389*.

Deng, H. (2020). Slightly conservative bootstrap for maxima of sums. *arXiv preprint arXiv:2007.15877*.

Deng, H., Han, Q., and Sen, B. (2020a). Inference for local parameters in convexity constrained models. *arXiv preprint arXiv:2006.10264*.

Deng, H., Han, Q., and Zhang, C.-H. (2020b). Confidence intervals for multiple isotonic regression and other monotone models. *arXiv preprint arXiv:2001.07064*.

Desu, M. M. and Rodine, R. H. (1969). Estimation of the population median. *Scandinavian Actuarial Journal*, 1969(1-2):67–70.

Doerr, B. (2018). An elementary analysis of the probability that a binomial random variable exceeds its expectation. *Statistics & Probability Letters*, 139:67–74.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95(1):125–140.

Durot, C. (2008). Monotone nonparametric regression with random design. *Mathematical Methods of Statistics*, 17(4):327–341.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, pages 1–26.

Efron, B. (1982). Transformation theory: How normal is a family of distributions? *The Annals of Statistics*, pages 323–339.

Fang, X. and Koike, Y. (2020). High-dimensional central limit theorems by stein's method. *arXiv preprint arXiv:2001.10917*.

Fang, Z. and Santos, A. (2019). Inference on directionally differentiable functions. *The Review of Economic Studies*, 86(1):377–412.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Gebhardt, F. (1969). Some numerical comparisons of several approximations to the binomial distribution. *Journal of the American Statistical Association*, 64(328):1638–1646.

Geyer, C. J. (1996). On the asymptotics of convex stochastic optimization. *Unpublished manuscript*.

Greenberg, S. and Mohri, M. (2014). Tight lower bound on the probability of a binomial exceeding its expectation. *Statistics & Probability Letters*, 86:91–98.

Guntuboyina, A. and Sen, B. (2018). Nonparametric shape-restricted regression. *Statistical Science*, 33(4):568–594.

Hall, P. (1982). On estimating the endpoint of a distribution. *The Annals of Statistics*, pages 556–568.

Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics*, pages 1431–1452.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, pages 927–953.

Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Hamza, K. (1995). The smallest uniform upper bound on the distance between the mean and the median of the binomial and poisson distributions. *Statistics & Probability Letters*, 23(1):21–25.

Han, Q. and Kato, K. (2019). Berry-esseen bounds for chernoff-type non-standard asymptotics in isotonic regression. *arXiv preprint arXiv:1910.09662*.

Hartigan, J. (1970). Exact confidence intervals in regression problems with independent symmetric errors. *The Annals of Mathematical Statistics*, pages 1992–1998.

Hartigan, J. A. (1969). Using subsample values as typical values. *Journal of the American Statistical Association*, 64(328):1303–1317.

Hayakawa, S., Lyons, T., and Oberhauser, H. (2021). Estimating the probability that a given vector is in the convex hull of a random sample. *arXiv preprint arXiv:2101.04250*.

Hirano, K. and Porter, J. R. (2012). Impossibility results for nondifferentiable functionals. *Econometrica*, 80(4):1769–1790.

Hirji, K. F., Tsiatis, A. A., and Mehta, C. R. (1989). Median unbiased estimation for binary data. *The American Statistician*, 43(1):7–11.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30.

Jing, B.-Y., Shao, Q.-M., and Wang, Q. (2003). Self-normalized cramér-type large deviations for independent random variables. *The Annals of probability*, 31(4):2167–2215.

John, S. (1974). Median-unbiased most acceptable estimates of poisson, binomial and negative-binomial distributions. *Communications in Statistics-Theory and Methods*, 3(12):1155–1159.

Kabluchko, Z. and Zaporozhets, D. (2019). Expected volumes of gaussian polytopes, external angles, and multiple order statistics. *Transactions of the American Mathematical Society*, 372(3):1709–1733.

Kenne Pagui, E. C., Salvan, A., and Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika*, 104(4):923–938.

Kim, K. I. (2016). Higher order bias correcting moment equation for m-estimation and its higher order efficiency. *Econometrics*, 4(4):48.

Knight, K. (1989). On the bootstrap of the sample mean in the infinite variance case. *The Annals of Statistics*, pages 1168–1175.

Knight, K. (1998). Limiting distributions for l1 regression estimators under general conditions. *Annals of statistics*, pages 755–770.

Knight, K. and Bassett, G. W. (2002). Second order improvements of sample quantiles using subsamples. *Unpublished Manuscript*.

Koike, Y. (2020). Notes on the dimension dependence in high-dimensional central limit theorems for hyper-rectangles. *Japanese Journal of Statistics and Data Science*, pages 1–41.

Koltchinskii, V. (2020). Estimation of smooth functionals in high-dimensional models: bootstrap chains and gaussian approximation. *arXiv preprint arXiv:2011.03789*.

Koltchinskii, V. and Zhilova, M. (2019). Estimation of smooth functionals in normal models: bias reduction and asymptotic efficiency. *arXiv preprint arXiv:1912.08877*.

Koltchinskii, V. and Zhilova, M. (2021). Efficient estimation of smooth functionals in gaussian shift models. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 57, pages 351–386. Institut Henri Poincaré.

Kosmidis, I. and Firth, D. (2009). Bias reduction in exponential family nonlinear models. *Biometrika*, 96(4):793–804.

Kosmidis, I., Pagui, E. C. K., and Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing*, 30(1):43–59.

Kuchibhotla, A. K., Patra, R. K., and Sen, B. (2021). Semiparametric efficiency in convexity constrained single index model. *Journal of American Statistical Association (Forthcoming)*.

Lanke, J. (1974). Interval estimation of a median. *Scandinavian Journal of Statistics*, pages 28–32.

Laurent, B. (1997). Estimation of integral functionals of a density and its derivatives. *Bernoulli*, 3(2):181–211.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Li, Q. and Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*.

Loh, W.-Y. (1984). Estimating an endpoint of a distribution with resampling methods. *The Annals of Statistics*, 12(4):1543–1550.

Mammen, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields*, 93(4):439–455.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283.

Pfanzagl, J. (1970a). Median unbiased estimates for mlr-families. *Metrika*, 15(1):30–39.

Pfanzagl, J. (1970b). On the asymptotic efficiency of median unbiased estimates. *Annals of Mathematical Statistics*, 41(5):1500–1509.

Pfanzagl, J. (1971). The berry-esseen bound for minimum contrast estimates. *Metrika*, 17(1):82–91.

Pfanzagl, J. (1972). On median unbiased estimates. *Metrika*, 18(1):154–173.

Pfanzagl, J. (1973a). The accuracy of the normal approximation for estimates of vector parameters. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 25(3):171–198.

Pfanzagl, J. (1973b). Asymptotic expansions related to minimum contrast estimators. *The Annals of Statistics*, pages 993–1026.

Pfanzagl, J. (1979). On optimal median unbiased estimators in the presence of nuisance parameters. *Annals of Statistics*, 7(1):187–193.

Pfanzagl, J. (2011). *Parametric statistical theory*. Walter de Gruyter.

Pfanzagl, J. (2017). Optimality of unbiased estimators: Nonasymptotic theory. In *Mathematical Statistics*, pages 83–106. Springer.

Pinelis, I. (2017). Optimal-order uniform and nonuniform bounds on the rate of convergence to normality for maximum likelihood estimators. *Electronic Journal of Statistics*, 11(1):1160–1179.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics*, 47(6):3438 – 3469.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer.

Robson, D. and Whitlock, J. (1964). Estimation of a truncation point. *Biometrika*, 51(1/2):33–39.

Romano, J. P. and Wolf, M. (1999). Subsampling inference for the mean in the heavy-tailed case. *Metrika*, 50(1):55–69.

Sen, P. K. (1968). Asymptotic normality of sample quantiles for m-dependent processes. *The annals of mathematical statistics*, 39(5):1724–1730.

Shao, J. and Tu, D. (2012). *The jackknife and bootstrap*. Springer Science & Business Media.

Shao, Q.-M. (1997). Self-normalized large deviations. *The Annals of Probability*, pages 285–328.

Sherman, M. and Carlstein, E. (1997). Omnibus confidence intervals. *Texas A&M University, Dept. of Statistics, Technical Report*, 278.

Stigler, S. M. (2007). The epic story of maximum likelihood. *Statistical Science*, 22(4):598–620.

van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Wagner, U. and Welzl, E. (2001). A continuous analogue of the upper bound theorem. *Discrete & Computational Geometry*, 26(2):205–219.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Wendel, J. G. (1962). A problem in geometric probability. *Math. Scand*, 11(109-111):123.

Wright, F. T. (1981). The Asymptotic Behavior of Monotone Regression Estimates. *The Annals of Statistics*, 9(2):443 – 448.

Zhang, J.-J. and Liang, H.-Y. (2011). Berry–esseen type bounds in heteroscedastic semi-parametric model. *Journal of statistical planning and inference*, 141(11):3447–3462.

Supplement to
"The HulC: Confidence Regions from Convex Hulls"

Abstract

This supplement contains the proofs of all the main results in the paper.

S.1 Union bound with Wald intervals

If, for each $1 \leqslant k \leqslant d$, the estimators $e_k^\top \widehat{\theta}_j$ are asymptotically Normal, then asymptotic Normality implies that for all $\gamma \in (0,1)$,

$$\left|\mathbb{P}\left(e_k^\top \theta_0 \in \widehat{\mathrm{CI}}_\gamma^{\texttt{Wald},k}\right) - (1-\gamma)\right| \leqslant \delta_n,$$

for some δ_n converging to zero as $n \to \infty$; an example is (11). First order accurate confidence intervals (such as Wald's) satisfy $\delta_n = O(n^{-1/2})$, second order accurate ones satisfy $\delta_n = O(n^{-1})$ and so on. Taking $\gamma = \alpha/d$ and applying the union bound, we only obtain

$$\mathbb{P}\left(\bigcup_{k=1}^d \left\{e_k^\top \theta_0 \notin \widehat{\mathrm{CI}}_{\alpha/d}^{\texttt{Wald},k}\right\}\right) \leqslant \sum_{k=1}^d \mathbb{P}\left(e_k^\top \theta_0 \notin \widehat{\mathrm{CI}}_{\alpha/d}^{\texttt{Wald},k}\right) \leqslant \alpha + d\delta_n.$$

In order for the right hand side to be α asymptotically, one needs $d\delta_n = o(1)$. This is a very stringent requirement, especially when the dimension d grows faster than the sample size n.

There is a simple way to resolve this issue following our proposed methodology. The idea is to construct 1/2 Wald confidence regions for each coordinate $e_k^\top \theta_0$ from each of $e_k^\top \widehat{\theta}_j, 1 \leqslant j \leqslant B$ and then take the union of these regions. Formally, set $\bar{B}_{d,\alpha} = \lceil \log(d/\alpha)/\log(2) \rceil$. For $1 \leqslant j \leqslant \bar{B}_{d,\alpha}$, suppose $\widehat{\mathrm{CI}}_j^{\texttt{Wald},k}$ is the Wald confidence region of coverage 1/2 based on $e_k^\top \widehat{\theta}_j$. This means that

$$\left|\mathbb{P}\left(e_k^\top \theta_0 \in \widehat{\mathrm{CI}}_j^{\texttt{Wald},k}\right) - \frac{1}{2}\right| \leqslant \delta_{n,d}, \quad \text{for all} \quad 1 \leqslant j \leqslant \bar{B}_{d,\alpha}, 1 \leqslant k \leqslant d.$$

The right hand side $\delta_{n,d}$ here depends on the dimension d because $\widehat{\theta}_j$ is computed based on $n/\bar{B}_{d/\alpha}$ many observations. Under these conditions, the following result provides a valid $1-\alpha$ confidence regions using Wald's confidence intervals and a union bound. The interesting aspect (similar to (10)) is that the coverage implied by Proposition 3 is eventually finite sample (i.e., holds after some sample size) even though the coverage of $\widehat{\mathrm{CI}}_j^{\texttt{Wald},k}$ is asymptotic.

Proposition 3. *If*

$$\bar{B}_{d,\alpha}\delta_{n,d} \;\leqslant\; \bar{B}_{d,\alpha}\left[\left(\frac{\alpha}{d}\right)^{1/\bar{B}_{d,\alpha}} - \frac{1}{2}\right], \tag{E.1}$$

then

$$\mathbb{P}\left(\theta_0 \notin \bigotimes_{k=1}^d \bigcup_{j=1}^{\bar{B}_{d,\alpha}} \widehat{\mathrm{CI}}_j^{\texttt{Wald},k}\right) \;\leqslant\; \alpha.$$

Proof. Following the proof of Lemma 1, we obtain

$$\mathbb{P}\left(e_k^\top \theta_0 \notin \bigcup_{j=1}^{\bar{B}_{d,\alpha}} \widehat{\mathrm{CI}}_j^{\texttt{Wald},k}\right) \;\leqslant\; \left(\frac{1}{2} + \delta_{n,d}\right)^{\bar{B}_{d,\alpha}} \leqslant \frac{\alpha}{d}.$$

The last inequality here follows from (E.1). Hence the union bound applies and proves the result. □

Similar to the right hand side of (8), the right hand side of (E.1) is be bounded away from zero unless d/α is a power of 2. It is worth noting that, unlike (10), the construction of the confidence region $\otimes_{k=1}^{d} \cup_{j=1}^{\bar{B}_{d,\alpha}} \widehat{\mathrm{CI}}_j^{\mathtt{Wald},k}$ requires estimation of variance of the estimators and its validity requires (marginal) distributional convergence of the estimators $e_k^\top \widehat{\theta}_j$.

S.2 Proof of Lemma 1

It is clear that

$$\begin{aligned}
\mathbb{P}\left(\theta_0 \notin \left[\min_{1\leqslant j\leqslant B} \widehat{\theta}_j, \max_{1\leqslant j\leqslant B} \widehat{\theta}_j\right]\right) &= \mathbb{P}\left(\theta_0 < \min_{1\leqslant j\leqslant B} \widehat{\theta}_j(k)\right) + \mathbb{P}\left(\max_{1\leqslant j\leqslant B} \widehat{\theta}_j < \theta_0\right) \\
&= \prod_{j=1}^{B} \mathbb{P}(\widehat{\theta}_j > \theta_0) + \prod_{j=1}^{B} \mathbb{P}(\widehat{\theta}_j < \theta_0) \\
&= \prod_{j=1}^{B} \left\{1 - \mathbb{P}(\widehat{\theta}_j \leqslant \theta_0)\right\} + \prod_{j=1}^{B} \left\{1 - \mathbb{P}(\widehat{\theta}_j \geqslant \theta_0)\right\}.
\end{aligned}$$

The result now follows from the definition (2) of Δ. Note that the inequality in result stems from the fact that Δ in (2) is the maximum value over all estimators. Furthermore, we use the fact that $\mathbb{P}(\widehat{\theta}_j < \theta_0) + \mathbb{P}(\widehat{\theta}_j > \theta_0) \leqslant 1$. If this inequality is strict, then the inequality in the coverage is strict.

S.3 Proof of Theorem 1

The confidence interval from the HulC is either based on $B_{\alpha,\Delta}$ estimators or based on $B_{\alpha,\Delta} - 1$ estimators. The number of estimators used depends on the realization of the uniform random variable U in step 2 of the HulC. With probability $\tau_{\alpha,\Delta}$, the confidence interval will be based on $B_{\alpha,\Delta} - 1$ estimators and with probability $1 - \tau_{\alpha,\Delta}$, the confidence interval will be based on $B_{\alpha,\Delta}$ estimators. Hence from Lemma 1, we obtain

$$\begin{aligned}
\mathbb{P}\left(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta}\right) &= \tau_{\alpha,\Delta} \mathbb{P}\left(\theta_0 \notin \left[\min_{1\leqslant j\leqslant B_{\alpha,\Delta}-1} \widehat{\theta}_j, \max_{1\leqslant j\leqslant B_{\alpha,\Delta}-1} \widehat{\theta}_j\right]\right) \\
&\quad + (1 - \tau_{\alpha,\Delta}) \mathbb{P}\left(\theta_0 \notin \left[\min_{1\leqslant j\leqslant B_{\alpha,\Delta}} \widehat{\theta}_j, \max_{1\leqslant j\leqslant B_{\alpha,\Delta}} \widehat{\theta}_j\right]\right) \\
&\leqslant \tau_{\alpha,\Delta} P(B_{\alpha,\Delta} - 1; \Delta) + (1 - \tau_{\alpha,\Delta}) P(B_{\alpha,\Delta}; \Delta) \;=\; \alpha.
\end{aligned} \tag{E.2}$$

This proves (5). Under the additional assumptions for (6), the only inequality in (E.2) becomes equality. Firstly, $\mathbb{P}(\widehat{\theta}_j = \theta_0) = 0$ for all $j \geqslant 1$ implies that

$$\mathbb{P}\left(\theta_0 \notin \left[\min_{1\leqslant j\leqslant B} \widehat{\theta}_j, \max_{1\leqslant j\leqslant B} \widehat{\theta}_j\right]\right) = \prod_{j=1}^{B} \mathbb{P}(\widehat{\theta}_j > \theta_0) + \prod_{j=1}^{B} (1 - \mathbb{P}(\widehat{\theta}_j > \theta_0)).$$

Secondly, the assumption that all the estimators have a median bias of Δ exactly, implies that this probability is exactly $(1/2 - \Delta)^B + (1/2 + \Delta)^B$. This proves (6).

S.4 Proof of Proposition 1

If $\widetilde{\Delta} = \Delta$, then the conclusion is obvious. Consider the case, $\widetilde{\Delta} \neq \Delta$. Note that $P(B;\Delta)$ is an increasing function of Δ. Hence, for any $0 \leqslant \Delta \leqslant \widetilde{\Delta} < 1/2$, we have

$$\begin{aligned} 1 \leqslant \frac{P(B;\widetilde{\Delta})}{P(B;\Delta)} &= \frac{(1+2\widetilde{\Delta})^B + (1-2\widetilde{\Delta})^B}{(1+2\Delta)^B + (1-2\Delta)^B} \\ &\leqslant \max\left\{ \frac{(1+2\widetilde{\Delta})^B}{(1+2\Delta)^B}, \frac{(1-2\widetilde{\Delta})^B}{(1-2\Delta)^B} \right\} = \left(\frac{1+2\widetilde{\Delta}}{1+2\Delta}\right)^B = \left(1 + \frac{2(\widetilde{\Delta}-\Delta)}{(1+2\Delta)}\right)^B \\ &\leqslant \Big(1 + 2(\widetilde{\Delta}-\Delta)\Big)^B. \end{aligned} \tag{E.3}$$

Reversing the roles of $\widetilde{\Delta}$ and Δ, we conclude that if $0 \leqslant \widetilde{\Delta} \leqslant \Delta < 1/2$, then

$$\Big(1 + 2(\Delta - \widetilde{\Delta})\Big)^{-B} \leqslant \frac{P(B;\widetilde{\Delta})}{P(B;\Delta)} \leqslant 1.$$

Hence, for all $\widetilde{\Delta}, \Delta \in [0, 1/2)$, and all $B \geqslant 1$, we have

$$P(B;\Delta)\left(1 + 2|\Delta - \widetilde{\Delta}|\right)^{-B} \;\leqslant\; P(B;\widetilde{\Delta}) \;\leqslant\; P(B;\Delta)\left(1 + 2|\Delta - \widetilde{\Delta}|\right)^{B}. \tag{E.4}$$

Because $P(B_{\alpha,\Delta};\Delta) < \alpha$, using (E.4) with $B = B_{\alpha,\Delta}$, we get $P(B_{\alpha,\Delta};\widetilde{\Delta}) \leqslant \alpha$ if

$$\Big(1 + 2|\Delta - \widetilde{\Delta}|\Big)^{B_{\alpha,\Delta}} \leqslant \frac{\alpha}{P(B_{\alpha,\Delta};\Delta)}. \tag{E.5}$$

Similarly, $P(B_{\alpha,\Delta} - 1;\Delta) > \alpha$, using (E.4) with $B = B_{\alpha,\Delta} - 1$, we get $P(B_{\alpha,\Delta} - 1;\widetilde{\Delta}) > \alpha$ if

$$\Big(1 + 2|\Delta - \widetilde{\Delta}|\Big)^{B_{\alpha,\Delta}-1} < \frac{P(B_{\alpha,\Delta} - 1;\Delta)}{\alpha}. \tag{E.6}$$

Note that both the ratios on the right hand side of (E.5) and (E.6) are at least 1. Furthermore, inequality (E.6) will be satisfied if $(1+2|\Delta - \widetilde{\Delta}|)^{B_{\alpha,\Delta}} \leqslant P(B_{\alpha,\Delta} - 1;\Delta)/\alpha$. Combining inequalities (E.5) and (E.6), we get that if (8) holds true, then $P(B_{\alpha,\Delta};\widetilde{\Delta}) \leqslant \alpha < P(B_{\alpha,\Delta} - 1;\widetilde{\Delta})$ and hence $B_{\alpha,\Delta} = B_{\alpha,\widetilde{\Delta}}$.

Note that in inequality (E.3), we used an inequality for the ratio $P(B;\widetilde{\Delta})/P(B;\Delta)$. Observe that if $\Delta = 0$, then $P(B;\widetilde{\Delta})/P(B;\Delta)$ has zero derivative at $\widetilde{\Delta} = 0$. Observe that, with $\Delta = 0$,

$$1 \leqslant \frac{P(B;\widetilde{\Delta})}{P(B;\Delta)} = \frac{(1+2\widetilde{\Delta})^B + (1-2\widetilde{\Delta})^B}{2} \;=\; 1 + 0\widetilde{\Delta} + \frac{B(B-1)}{2}\widetilde{\Delta}^2 P(B-2;\widetilde{\Delta}^*),$$

for some $\widetilde{\Delta}^* \in [0, \widetilde{\Delta}]$. Because $P(B;\delta) \leqslant 1$ for all $B \geqslant 1, \delta \in [0, 1/2]$, we get

$$1 \;\leqslant\; \frac{P(B;\widetilde{\Delta})}{P(B;0)} \;\leqslant\; 1 + \frac{B(B-1)}{2}\widetilde{\Delta}^2. \tag{E.7}$$

Because $P(B_{\alpha,0} - 1;0) > \alpha$, taking $B = B_{\alpha,0} - 1$ in (E.7), we obtain

$$P(B_{\alpha,0} - 1;\widetilde{\Delta}) \geqslant P(B_{\alpha,0} - 1;0) > \alpha \quad \Rightarrow \quad P(B_{\alpha,0} - 1;\widetilde{\Delta}) > \alpha. \tag{E.8}$$

Because $P(B_{\alpha,0};0) \leqslant \alpha$, taking $B = B_{\alpha,0}$ in (E.7), we obtain

$$P(B_{\alpha,0};\widetilde{\Delta}) \leqslant \alpha, \quad \text{if} \quad \frac{B_{\alpha,0}(B_{\alpha,0}-1)}{2}\widetilde{\Delta}^2 \leqslant \frac{\alpha}{P(B_{\alpha,0};0)} - 1. \tag{E.9}$$

Combining (E.8) and (E.9), we obtain (9).

S.5 Proof of Theorem 2

Because the median bias of estimators from $\mathcal{A}(\cdot)$ is $\widetilde{\Delta}$, we set

$$\widetilde{\Delta} \;\geqslant\; \max_{1\leqslant j\leqslant B^*}\left(\frac{1}{2}-\max\left\{\mathbb{P}(\widehat{\theta}_j\geqslant\theta_0),\,\mathbb{P}(\widehat{\theta}_j\leqslant\theta_0)\right\}\right)_+.$$

Lemma 1 implies that

$$\mathbb{P}(\theta_0\notin\widehat{\mathrm{CI}}_{\alpha,0})\leqslant\mathbb{E}\left[\left(\frac{1}{2}+\widetilde{\Delta}\right)^{B^*}+\left(\frac{1}{2}-\widetilde{\Delta}\right)^{B^*}\right].$$

The right hand side involves an expectation because B^* is a random variable satisfying

$$\mathbb{P}(B^*=B_{\alpha,0})=1-\tau_{\alpha,0},\quad\text{and}\quad\mathbb{P}(B^*=B_{\alpha,0}-1)=\tau_{\alpha,0}.$$

This follows from (4). Therefore,

$$\mathbb{P}(\theta_0\notin\widehat{\mathrm{CI}}_{\alpha,0})\leqslant\tau_{\alpha,0}P(B_{\alpha,0}-1;\widetilde{\Delta})+(1-\tau_{\alpha,0})P(B_{\alpha,0};\widetilde{\Delta}).\tag{E.10}$$

From the proof of Proposition 1 (in particular (E.7)), it follows that

$$\max\left\{\frac{P(B_{\alpha,0};\widetilde{\Delta})}{P(B_{\alpha,0};0)},\,\frac{P(B_{\alpha,0}-1;\widetilde{\Delta})}{P(B_{\alpha,0}-1;0)}\right\}\leqslant 1+\frac{B_{\alpha,0}(B_{\alpha,0}-1)}{2}\widetilde{\Delta}^2.\tag{E.11}$$

Substituting this inequality in (E.10) yields

$$\begin{aligned}\mathbb{P}(\theta_0\notin\widehat{\mathrm{CI}}_{\alpha,0})&\leqslant\tau_{\alpha,0}P(B_{\alpha,0}-1;0)\left(1+\frac{B_{\alpha,0}(B_{\alpha,0}-1)}{2}\widetilde{\Delta}^2\right)\\&\quad+(1-\tau_{\alpha,0})P(B_{\alpha,0};0)\left(1+\frac{B_{\alpha,0}(B_{\alpha,0}-1)}{2}\widetilde{\Delta}^2\right)\\&=\left[\tau_{\alpha,0}P(B_{\alpha,0}-1;0)+(1-\tau_{\alpha,0})P(B_{\alpha,0};0)\right]\left(1+\frac{B_{\alpha,0}(B_{\alpha,0}-1)}{2}\widetilde{\Delta}^2\right)\\&=\alpha\left(1+\frac{B_{\alpha,0}(B_{\alpha,0}-1)}{2}\widetilde{\Delta}^2\right).\end{aligned}$$

The last equality follows from the definition (4) of $\tau_{\alpha,0}$. This completes the proof of upper bound in (15).

From the proof of Lemma 1, it follows, under the assumption of $\mathbb{P}(\widehat{\theta}_j=\theta_0)=0$ and the exact median bias of $\widetilde{\Delta}$, that

$$\mathbb{P}(\theta_0\notin\widehat{\mathrm{CI}}_{\alpha,\Delta})=\mathbb{E}\left[\left(\frac{1}{2}-\widetilde{\Delta}\right)^{B^*}+\left(\frac{1}{2}-\widetilde{\Delta}\right)^{B^*}\right]\geqslant\mathbb{E}\left[\frac{2}{2^{B^*-1}}\right]=\alpha,$$

the last equality follows again from the definition of $\tau_{\alpha,0}$ in (4). This completes the proof of (16).

For the case where the estimators have an asymptotic median bias of Δ ($\neq 0$), we use the HulC to obtain $\widehat{\mathrm{CI}}_{\alpha,\Delta}$, while the true finite sample median bias is bounded by $\widetilde{\Delta}$. In this case, to prove the upper bound, we use the inequality

$$\max\left\{\frac{P(B_{\alpha,\Delta};\widetilde{\Delta})}{P(B_{\alpha,\Delta};\Delta)},\,\frac{P(B_{\alpha,\Delta}-1;\widetilde{\Delta})}{P(B_{\alpha,\Delta}-1;\Delta)}\right\}\leqslant\left(1+2|\widetilde{\Delta}-\Delta|\right)^{B_{\alpha,\Delta}},$$

in place of (E.11). Using this inequality in (E.10) (with $\tau_{\alpha,0}$ replaced by $\tau_{\alpha,\Delta}$), we obtain

$$\begin{aligned}\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta}) &\leqslant [\tau_{\alpha,\Delta} P(B_{\alpha,\Delta}-1;\Delta) + (1-\tau_{\alpha,\Delta})P(B_{\alpha,\Delta};\Delta)]\left(1+2|\widetilde{\Delta}-\Delta|\right)^{B_{\alpha,\Delta}} \\ &= \alpha\left(1+2|\widetilde{\Delta}-\Delta|\right)^{B_{\alpha,\Delta}}.\end{aligned}$$

From the proof of Lemma 1, it follows, under the assumption of $\mathbb{P}(\widehat{\theta}_j = \theta_0) = 0$ and the exact median bias of $\widetilde{\Delta}$, that

$$\begin{aligned}\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta}) &= P(B_{\alpha,\Delta};\widetilde{\Delta})(1-\tau_{\alpha,\Delta}) + P(B_{\alpha,\Delta}-1;\widetilde{\Delta})\tau_{\alpha,\Delta} \\ &\geqslant [P(B_{\alpha,\Delta};\Delta)(1-\tau_{\alpha,\Delta}) + P(B_{\alpha,\Delta}-1;\Delta)\tau_{\alpha,\Delta}]\left(1+2|\widetilde{\Delta}-\Delta|\right)^{-B_{\alpha,\Delta}}, \quad \text{using (E.4)}, \\ &= \alpha\left(1+2|\widetilde{\Delta}-\Delta|\right)^{-B_{\alpha,\Delta}}.\end{aligned}$$

This proves the upper and lower bounds for a non-zero asymptotic median bias of Δ.

S.6 Proof of Lemma 2

Equality (27) follows from Theorem 3 of Hayakawa et al. (2021). This result was originally proved in Wendel (1962) under symmetry of $\widehat{\theta}_j - \theta_0$. Wagner and Welzl (2001) proved that the miscoverage probability is lower bounded by the quantity on the right hand side whenever $\widehat{\theta}_j - \theta_0$ has an absolutely continuous distribution; this does not require the assumption of $\mathbb{P}(c^\top(\widehat{\theta}_j - \theta_0) \leqslant 0) = 1/2$ for all $c \in \mathbb{R}^d \backslash \{0\}$.

Equality (28) follows readily from Lemma 1 and union bound. Formally,

$$\begin{aligned}\mathbb{P}\left(\theta_0 \notin \mathrm{RectHull}(\{\widehat{\theta}_j : 1 \leqslant j \leqslant B\})\right) &= \mathbb{P}\left(\bigcup_{k=1}^{d}\left\{e_k^\top\theta_0 \notin \left[\min_{1\leqslant j\leqslant B} e_k^\top\widehat{\theta}_j,\ \max_{1\leqslant j\leqslant B} e_k^\top\widehat{\theta}_j\right]\right\}\right) \\ &\leqslant \sum_{k=1}^{d}\mathbb{P}\left(e_k^\top\theta_0 \notin \left[\min_{1\leqslant j\leqslant B} e_k^\top\widehat{\theta}_j,\ \max_{1\leqslant j\leqslant B} e_k^\top\widehat{\theta}_j\right]\right) \\ &\leqslant \sum_{k=1}^{d}\left\{\left(\frac{1}{2}-\Delta_k\right)^B + \left(\frac{1}{2}+\Delta_k\right)^B\right\},\end{aligned}$$

where Δ_k is an upper bound on the median bias of $e_k^\top\widehat{\theta}_j$ for $1 \leqslant j \leqslant B$. Hence inequality (28) follows.

S.7 Proof of Proposition 2

For notational convenience and without loss of generality, we will prove the result when $\widehat{\theta}_j$ is computed based on n observations. Set

$$\widehat{T} = \frac{1}{\binom{n}{2}}\sum_{i<j} X_i X_j.$$

This can be rewritten as

$$\widehat{T} = \frac{n}{(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 - \frac{1}{(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right).$$

In terms of ξ_i and $Z = n^{-1/2}\sum_{i=1}^n \xi_i$, this becomes

$$\widehat{T} - \mu^2 = \frac{2\mu\sigma Z}{\sqrt{n}} + \frac{\sigma^2}{n-1}\left[Z^2 - 1\right] + \frac{\sigma^2}{n-1}\left[1 - \frac{1}{n}\sum_{i=1}^n \xi_i^2\right].$$

It is clear that

$$\frac{\widehat{T}}{\sigma^2} - \frac{\mu^2}{\sigma^2} \;=\; 2\frac{\mu}{\sigma}\frac{Z}{\sqrt{n}} + \frac{Z^2-1}{n-1} + \frac{1}{\sqrt{n}(n-1)}\left[\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - 1\right)\right].$$

Set $W = \sqrt{n}(n^{-1}\sum_{i=1}^n \xi_i^2 - 1) = O_p(1)$. By the univariate Berry–Esseen bound, we get

$$\left|\mathbb{P}\left(W \leqslant t\right) - \mathbb{P}\left(N(0, \mathbb{E}[\xi^4]) \leqslant t\right)\right| \leqslant \frac{\mathbb{E}[\xi^6]}{(\mathbb{E}[\xi^4])^{3/2}\sqrt{n}}. \tag{E.12}$$

Similarly,

$$|\mathbb{P}(Z \leqslant t) - \mathbb{P}(N(0,1) \leqslant t)| \leqslant \frac{\mathbb{E}[|\xi|^3]}{(\mathbb{E}[\xi^2])^{3/2}\sqrt{n}}. \tag{E.13}$$

Inequality (E.12) implies that

$$\mathbb{P}\left(|W| > (3\mathbb{E}[\xi^4]\log n)^{1/2}\right) \leqslant \frac{1}{n} + \frac{\mathbb{E}[\xi^6]}{(\mathbb{E}[\xi^4])^{3/2}\sqrt{n}}.$$

Clearly,

$$\begin{aligned}
\mathbb{P}\left(\widehat{T} \leqslant \mu^2\right) &= \mathbb{P}\left(2\frac{\mu}{\sigma}Z + \frac{\sqrt{n}}{n-1}(Z^2-1) + \frac{W}{n-1} \leqslant 0\right) \\
&= \mathbb{P}\left(2\frac{\mu}{\sigma}Z + \frac{\sqrt{n}}{n-1}(Z^2-1) + \frac{W}{n-1} \leqslant 0,\ |W| \leqslant (3\mathbb{E}[\xi^4]\log n)^{1/2}\right) \\
&\quad + \mathbb{P}\left(2\frac{\mu}{\sigma}Z + \frac{\sqrt{n}}{n-1}(Z^2-1) + \frac{W}{n-1} \leqslant 0,\ |W| > (3\mathbb{E}[\xi^4]\log n)^{1/2}\right) \\
&\leqslant \mathbb{P}\left(2\frac{\mu}{\sigma}Z + \frac{\sqrt{n}}{n-1}(Z^2-1) \leqslant \frac{(2\mathbb{E}[\xi^4]\log n)^{1/2}}{n-1}\right) \\
&\quad + \mathbb{P}(|W| > (3\mathbb{E}[\xi^4]\log n)^{1/2}).
\end{aligned} \tag{E.14}$$

Similarly,

$$\begin{aligned}
\mathbb{P}(\widehat{T} \leqslant \mu^2) \geqslant \mathbb{P}&\left(2\frac{\mu}{\sigma}Z + \frac{\sqrt{n}}{n-1}(Z^2-1) \leqslant -\frac{(3\mathbb{E}[\xi^4]\log n)^{1/2}}{n-1}\right) \\
&- \mathbb{P}(|W| > (3\mathbb{E}[\xi^4]\log n)^{1/2}).
\end{aligned} \tag{E.15}$$

Note that

$$\begin{aligned}
\{aZ + b(Z^2-1) \leqslant c\} &\equiv \left\{\frac{-a - \sqrt{a^2 + 4b(b+c)}}{2b} \leqslant Z \leqslant \frac{-a + \sqrt{a^2 + 4b(b+c)}}{2b}\right\} \\
&\equiv \left\{\left|Z + \frac{a}{2b}\right| \leqslant \frac{\sqrt{a^2 + 4b(b+c)}}{2b}\right\}.
\end{aligned}$$

Using inequality (E.13), we obtain

$$\begin{aligned}\mathbb{P}\left(aZ+b(Z^2-1)\leqslant c\right) &= \mathbb{P}\left(\left|Z+\frac{a}{2b}\right|\leqslant\frac{\sqrt{a^2+4b(b+c)}}{2b}\right)\\ &= \mathbb{P}\left(\left|N(0,1)+\frac{a}{2b}\right|\leqslant\frac{\sqrt{a^2+4b(b+c)}}{2b}\right)\pm\frac{\mathbb{E}[\xi^3]}{(\mathbb{E}[\xi^2])^{3/2}\sqrt{n}}\\ &= \Phi\left(\frac{-a+\sqrt{a^2+4b(b+c)}}{2b}\right)-\Phi\left(\frac{-a-\sqrt{a^2+4b(b+c)}}{2b}\right)\\ &\quad\pm\frac{\mathbb{E}[|\xi|^3]}{(\mathbb{E}[\xi^2])^{3/2}\sqrt{n}}.\end{aligned}$$

Finally, note that

$$\begin{aligned}\left|\Phi\left(\frac{-a\pm\sqrt{a^2+4b(b\pm c)}}{2b}\right)-\Phi\left(\frac{-a\pm\sqrt{a^2+4b^2}}{2b}\right)\right| &\leqslant \frac{1}{\sqrt{2\pi}}\left|\sqrt{\frac{a^2+4b^2\pm4bc}{4b^2}}-\sqrt{\frac{a^2+4b^2}{4b^2}}\right|\\ &\leqslant \frac{1}{\sqrt{2\pi}}\times\frac{|c|}{b}.\end{aligned}$$

Combining the two inequalities above, we get

$$\begin{aligned}\mathbb{P}\left(aZ+b(Z^2-1)\leqslant c\right) &= \Phi\left(\frac{-a+\sqrt{a^2+4b^2}}{2b}\right)-\Phi\left(\frac{-a-\sqrt{a^2+4b^2}}{2b}\right)\\ &\quad\pm\frac{2}{\sqrt{2\pi}}\frac{|c|}{b}\pm\frac{\mathbb{E}[|\xi|^3]}{(\mathbb{E}[\xi^2])^{3/2}\sqrt{n}}\\ &= \mathbb{P}(aZ+b(Z^2-1)\leqslant 0)\pm\frac{2}{\sqrt{2\pi}}\frac{|c|}{b}\pm\frac{\mathbb{E}[|\xi|^3]}{(\mathbb{E}[\xi^2])^{3/2}\sqrt{n}}.\end{aligned}$$

Substituting these inequalities in (E.14) and (E.15), we conclude

$$\begin{aligned}&\left|\mathbb{P}(\widehat{T}\leqslant\mu^2)-\left\{\Phi\left(\frac{-a+\sqrt{a^2+4b^2}}{2b}\right)-\Phi\left(\frac{-a-\sqrt{a^2+4b^2}}{2b}\right)\right\}\right|\\ &\quad\leqslant\sqrt{\frac{2}{\pi}}\frac{|c|}{b}+\frac{\mathbb{E}[|\xi|^3]}{(\mathbb{E}[\xi^2])^{3/2}\sqrt{n}}+\frac{2}{n}+\frac{\mathbb{E}[\xi^6]}{(\mathbb{E}[\xi^4])^{3/2}\sqrt{n}}.\end{aligned}\tag{E.16}$$

Here

$$a=2\frac{\mu}{\sigma},\quad b=\frac{\sqrt{n}}{n-1},\quad\text{and}\quad c=\pm\frac{\sqrt{3\mathbb{E}[\xi^4]\log n}}{n-1}.$$

This implies

$$\frac{|c|}{b}=\sqrt{\frac{3\mathbb{E}[\xi^4]\log n}{n}}.$$

Note that the right hand side of (E.16) is of order $n^{-1/2}$ and does not depend on μ; it only depends on $\mathbb{E}[|\xi|^j], j=3,4,6$. Inequality (E.16) implies that the median bias of $\widehat{T}$ can be obtained by taking the maximum over all $\theta\in\mathbb{R}$ of

$$\begin{aligned}&\left|\frac{1}{2}-\left\{\Phi\left(\frac{-2\theta+\sqrt{4\theta^2+4n/(n-1)^2}}{2\sqrt{n}/(n-1)}\right)-\Phi\left(\frac{-2\theta-\sqrt{4\theta^2+4n/(n-1)^2}}{2\sqrt{n}/(n-1)}\right)\right\}\right|\\ &\quad=\left|\frac{1}{2}-\left\{\Phi\left(\frac{-\theta+\sqrt{\theta^2+n/(n-1)^2}}{\sqrt{n}/(n-1)}\right)-\Phi\left(\frac{-\theta-\sqrt{\theta^2+n/(n-1)^2}}{\sqrt{n}/(n-1)}\right)\right\}\right|.\end{aligned}$$

It seems the maximum is attained at $\theta=0$ for any n.

S.8 Proof of Theorem 3

Throughout the proof, we write $\widehat{\Delta}$ instead of Δ for convenience. Define the event

$$\mathcal{E} := \{B_{\alpha,\widehat{\Delta}_n} = B_{\alpha,\Delta}\}.$$

and set

$$\widetilde{\mathrm{CI}}_\alpha^{(0)} := \left[\min_{1\leqslant j\leqslant B_{\alpha,\Delta}-1} \widehat{\theta}_j, \min_{1\leqslant j\leqslant B_{\alpha,\Delta}-1} \widehat{\theta}_j\right], \quad \text{and} \quad \widetilde{\mathrm{CI}}_\alpha^{(1)} := \left[\min_{1\leqslant j\leqslant B_{\alpha,\Delta}} \widehat{\theta}_j, \min_{1\leqslant j\leqslant B_{\alpha,\Delta}} \widehat{\theta}_j\right].$$

Recall $\tau_{\alpha,\Delta}$ from (4). On the event $\mathcal{E}$, we get that

$$\tau_{\alpha,\widehat{\Delta}} = \frac{\alpha - P(B_{\alpha,\Delta};\widehat{\Delta})}{P(B_{\alpha,\Delta}-1;\widehat{\Delta}) - P(B_{\alpha,\Delta};\widehat{\Delta})}, \quad \text{and} \quad \widehat{\mathrm{CI}}_\alpha^{(0)} = \widetilde{\mathrm{CI}}_\alpha^{(0)}, \quad \widehat{\mathrm{CI}}_\alpha^{(1)} = \widetilde{\mathrm{CI}}_\alpha^{(1)}.$$

We first bound the miscoverage probabilities of $\widehat{\mathrm{CI}}_\alpha^{(0)}$ and $\widehat{\mathrm{CI}}_\alpha^{(1)}$ when event $\mathcal{E}$ occurs. From the definition of $\Delta_{n,\alpha}$ and Lemma 1,

$$\begin{aligned}
\mathbb{P}(\{\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{(0)}\} \cap \mathcal{E}) &\leqslant \mathbb{P}(\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}) \\
&\leqslant P(B_{\alpha,\Delta}-1;\Delta_{n,\alpha}) \\
&= \frac{P(B_{\alpha,\Delta}-1;\Delta_{n,\alpha})}{P(B_{\alpha,\Delta}-1;\Delta)} P(B_{\alpha,\Delta}-1;\Delta) \\
&\leqslant P(B_{\alpha,\Delta}-1;\Delta) \times \begin{cases} (1+(B_{\alpha,0}-1)(B_{\alpha,0}-2)\Delta_{n,\alpha}^2/2), & \text{if } \Delta = 0, \\ (1+2|\Delta_{n,\alpha}-\Delta|)^{B_{\alpha,\Delta}-1}, & \text{if } \Delta \neq 0 \end{cases} \\
&\leqslant 2\alpha \times \begin{cases} (1+(B_{\alpha,0}-1)(B_{\alpha,0}-2)\Delta_{n,\alpha}^2/2), & \text{if } \Delta = 0, \\ (1+2|\Delta_{n,\alpha}-\Delta|)^{B_{\alpha,\Delta}-1}, & \text{if } \Delta \neq 0 \end{cases}.
\end{aligned} \tag{E.17}$$

The first inequality follows from proof of Proposition 1 (in particular (E.4)) while the second inequality follows from the fact that $P(B_{\alpha,\Delta};\Delta) \leqslant \alpha$ and $P(B_{\alpha,\Delta}-1;\Delta)/P(B_{\alpha,\Delta};\Delta) \leqslant 2$ for any $\Delta \in [0,1/2]$. Similarly,

$$\begin{aligned}
\mathbb{P}(\{\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{(1)}\} \cap \mathcal{E}) &\leqslant P(B_{\alpha,\Delta};\Delta_{n,\alpha}) \\
&= P(B_{\alpha,\Delta};\Delta)\frac{P(B_{\alpha,\Delta};\Delta_{n,\alpha})}{P(B_{\alpha,\Delta};\Delta)} \leqslant \alpha \times \begin{cases} (1+B_{\alpha,0}(B_{\alpha,0}-1)\Delta_{n,\alpha}^2/2), & \text{if } \Delta = 0, \\ (1+2|\Delta_{n,\alpha}-\Delta|)^{B_{\alpha,\Delta}}, & \text{if } \Delta \neq 0 \end{cases}.
\end{aligned} \tag{E.18}$$

This completes the proof of (36). We will now prove a bound on the miscoverage probability of $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$. Writing $\mathbb{E}_U[\cdot]$ to represent the expectation with respect to U which is a uniform random variable independent of the data, we get

$$\mathbb{E}_U[\mathbb{1}\{\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}\}]\mathbb{1}\{\mathcal{E}\} \;=\; \tau_{\alpha,\widehat{\Delta}}\mathbb{1}\{\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{(0)}\}\mathbb{1}\{\mathcal{E}\} + (1-\tau_{\alpha,\widehat{\Delta}})\mathbb{1}\{\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{(1)}\}\mathbb{1}\{\mathcal{E}\}. \tag{E.19}$$

Note that the miscoverage probability of $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$ is $\mathbb{E}[\mathbb{E}_U[\mathbb{1}\{\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}\}]]$. Using inequalities (E.17) and (E.18), we can readily obtain

$$\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}) \leqslant \mathbb{P}(\mathcal{E}^c) + 3\alpha \times \begin{cases} (1+B_{\alpha,0}(B_{\alpha,0}-1)\Delta_{n,\alpha}^2/2), & \text{if } \Delta = 0, \\ (1+2|\Delta_{n,\alpha}-\Delta|)^{B_{\alpha,\Delta}}, & \text{if } \Delta \neq 0 \end{cases}.$$

If $\Delta = 0$, this only implies approximately 3α miscoverage probability. If $\widehat{\Delta} - \Delta$ converges to zero in probability, then $\tau_{\alpha,\widehat{\Delta}} - \tau_{\alpha,\Delta}$ converges to zero and we obtain asymptotically α miscoverage probability for $\widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}$. This is the aim in proving (37). For this, we cannot readily use inequality (36) because $\tau_{\alpha,\widehat{\Delta}}$ depend on the same data as $\widehat{\mathrm{CI}}_\alpha^{(0)}$, and $\widehat{\mathrm{CI}}_\alpha^{(1)}$. In order to overcome this, we show $\tau_{\alpha,\widehat{\Delta}}$ is close to $\tau_{\alpha,\Delta}$ in a relative error sense. For this, it is not sufficient to know $B_{\alpha,\widehat{\Delta}} = B_{\alpha,\Delta}$. We need $|\widehat{\Delta} - \Delta|$ to be small. Consider the event

$$\mathcal{E}_0 \;:=\; \{|\widehat{\Delta} - \Delta| \leqslant \eta\},$$

for some $\eta \in [0, C_{\alpha,\Delta}]$. Recall from Proposition 1 implies that if $|\widehat{\Delta} - \Delta| \leqslant C_{\alpha,\Delta}$, then $B_{\alpha,\widehat{\Delta}} = B_{\alpha,\Delta}$. Hence, if $\eta \leqslant C_{\alpha,\Delta}$, then $\mathcal{E}_0 \subseteq \mathcal{E}$ and $\mathbb{P}(\mathcal{E}^c) \leqslant \mathbb{P}(\mathcal{E}_0^c)$. To control the difference between $\tau_{\alpha,\widehat{\Delta}}$ and $\tau_{\alpha,\Delta}$, we write $\tau_{\alpha,\widehat{\Delta}}$ as

$$\tau_{\alpha,\widehat{\Delta}} = \frac{1-\widehat{a}}{\widehat{b}-\widehat{a}},$$

where $\widehat{a} = P(B_{\alpha,\Delta}; \widehat{\Delta})/\alpha$ and $\widehat{b} = P(B_{\alpha,\Delta} - 1; \widehat{\Delta})/\alpha$. Similarly, we can write $\tau_{\alpha,\Delta} = (1-a)/(b-a)$ for a, b defined similar to $\widehat{a}, \widehat{b}$ with Δ replacing $\widehat{\Delta}$. From inequalities (E.4) and (E.7), we get that

$$\begin{aligned}
(1 + 2|\widehat{\Delta} - \Delta|)^{-B_{\alpha,\Delta}} \leqslant \frac{\widehat{a}}{a} = \frac{P(B_{\alpha,\Delta}; \widehat{\Delta})}{P(B_{\alpha,\Delta}; \Delta)} \leqslant (1 + 2|\widehat{\Delta} - \Delta|)^{B_{\alpha,\Delta}}, \quad \text{if } \Delta \neq 0, \\
1 \leqslant \frac{\widehat{a}}{a} = \frac{P(B_{\alpha,\Delta}; \widehat{\Delta})}{P(B_{\alpha,\Delta}; \Delta)} \leqslant (1 + B_{\alpha,0}(B_{\alpha,0} - 1)\widehat{\Delta}^2/2), \quad \text{if } \Delta = 0.
\end{aligned}$$

The same inequalities also hold true for $\widehat{b}/b$. For notational convenience, let us write $\ell \leqslant \widehat{a}/a \leqslant u$ and the same for $\widehat{b}/b$. It is easy to verify that $\tau_{\alpha,\widehat{\Delta}}$ is a decreasing function of $\widehat{a}$ and $\widehat{b}$ and hence

$$\frac{1/u - a}{(b-a)} \leqslant \tau_{\alpha,\widehat{\Delta}} \leqslant \frac{1/\ell - a}{(b-a)} \quad \Rightarrow \quad \frac{1 - 1/u}{b-a} \leqslant \tau_{\alpha,\widehat{\Delta}} - \tau_{\alpha,\Delta} \leqslant \frac{1/\ell - 1}{b-a}.$$

Recall that $b = P(B_{\alpha,\Delta} - 1; \Delta)/\alpha$ and $a = P(B_{\alpha,\Delta}; \Delta)/\alpha$. It follows that

$$(1/2 - \Delta) \leqslant \frac{P(B_{\alpha,\Delta} - 1; \Delta)}{\alpha}\left(\frac{1}{2} - \Delta\right) \;\leqslant\; b - a \;\leqslant\; \frac{P(B_{\alpha,\Delta}; \Delta)}{\alpha} \leqslant 1.$$

This yields

$$(1 - 1/u) \leqslant \tau_{\alpha,\widehat{\Delta}} - \tau_{\alpha,\Delta} \leqslant \frac{(1/\ell - 1)}{(1/2 - \Delta)}.$$

Hence, on event $\mathcal{E}_0$,

$$|\tau_{\alpha,\widehat{\Delta}} - \tau_{\alpha,\Delta}| \;\leqslant\; D_\tau \;:=\; \begin{cases} B_{\alpha,0}(B_{\alpha,0} - 1)\eta^2/2, & \text{if } \Delta = 0, \\ (1/2 - \Delta)^{-1}|(1 + 2\eta)^{B_{\alpha,\Delta}} - 1|, & \text{if } \Delta \neq 0. \end{cases} \tag{E.20}$$

Note that D_τ is non-random and only depends on η in the event $\mathcal{E}_0$. Inequality (E.20) can be alternatively written as $|\tau_{\alpha,\widehat{\Delta}} - \tau_{\alpha,\Delta}|\mathbb{1}\{\mathcal{E}_0\} \leqslant D_\tau$.

From (E.19) and the fact that $\widehat{\mathrm{CI}}_\alpha^{(j)} = \widetilde{\mathrm{CI}}_\alpha^{(j)}$ for $j = 0, 1$ on the event $\mathcal{E}_0$ (when $\eta \leqslant C_{\alpha,\Delta}$), we get

$$\begin{aligned}
\mathbb{E}_U[\mathbb{1}\{\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}\}]\mathbb{1}\{\mathcal{E}_0\} &= \tau_{\alpha,\hat{\Delta}}\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}\}\mathbb{1}\{\mathcal{E}_0\} + (1-\tau_{\alpha,\hat{\Delta}})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)}\}\mathbb{1}\{\mathcal{E}_0\} \\
&= \tau_{\alpha,\Delta}\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}\}\mathbb{1}\{\mathcal{E}_0\} + (1-\tau_{\alpha,\Delta})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)}\}\mathbb{1}\{\mathcal{E}_0\} \\
&\quad + (\tau_{\alpha,\hat{\Delta}} - \tau_{\alpha,\Delta})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}\}\mathbb{1}\{\mathcal{E}_0\} - (\tau_{\alpha,\hat{\Delta}} - \tau_{\alpha,\Delta})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)}\}\mathbb{1}\{\mathcal{E}_0\} \\
&= \tau_{\alpha,\Delta}\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}\} + (1-\tau_{\alpha,\Delta})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)}\} \\
&\quad - \tau_{\alpha,\Delta}\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}\}\mathbb{1}\{\mathcal{E}_0^c\} - (1-\tau_{\alpha,\Delta})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)}\}\mathbb{1}\{\mathcal{E}_0^c\} \\
&\quad + (\tau_{\alpha,\hat{\Delta}} - \tau_{\alpha,\Delta})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}\}\mathbb{1}\{\mathcal{E}_0\} - (\tau_{\alpha,\hat{\Delta}} - \tau_{\alpha,\Delta})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)}\}\mathbb{1}\{\mathcal{E}_0\}.
\end{aligned} \tag{E.21}$$

Because

$$\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta}) = \mathbb{E}[\tau_{\alpha,\Delta}\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}\} + (1-\tau_{\alpha,\Delta})\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)}\}],$$

it follows from (E.21) and (E.20) that

$$\left|\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_\alpha^{\mathrm{sub}}) - \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_{\alpha,\Delta})\right| \leqslant 2\mathbb{P}(\mathcal{E}_0^c) + D_\tau \mathbb{E}[\mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)}\} + \mathbb{1}\{\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)}\}]. \tag{E.22}$$

The second reminder term in (E.22) is controlled using the definition of D_τ in (E.20) and the bounds (E.17), (E.18) for $\mathbb{P}(\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(0)})$ and $\mathbb{P}(\theta_0 \notin \widetilde{\mathrm{CI}}_\alpha^{(1)})$. Finally, we use the fact that $2\eta B_{\alpha,\Delta} \leqslant 2C_{\alpha,\Delta}B_{\alpha,\Delta} \leqslant 1/2$ (from $\eta \leqslant C_{\alpha,\Delta}$ and Figure 2) to bound D_τ for $\Delta \neq 0$ as

$$(1+2\eta)^{B_{\alpha,\Delta}} - 1 \leqslant e^{2\eta B_{\alpha,\Delta}} - 1 \leqslant 2\eta B_{\alpha,\Delta} e^{2\eta B_{\alpha,\Delta}} \leqslant 2\sqrt{e}\eta B_{\alpha,\Delta}.$$

This completes the proof of (37).

S.9 Proof of Lemma 3

Set $J(x) = \mathbb{P}(W \leqslant x)$ and $J_b(x) = \mathbb{P}(r_b(\widehat{\theta}_b - \theta_0) \leqslant x)$. Note that by the triangle inequality, $|\widehat{\Delta}_n - \Delta| \leqslant |L_n(0) - J(0)|$, so we obtain that,

$$\begin{aligned}
|\widehat{\Delta}_n - \Delta| &\leqslant \left|\frac{1}{K_n}\sum_{j=1}^{K_n}\mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\} - J(0)\right| \\
&\leqslant \left|\frac{1}{K_n}\sum_{j=1}^{K_n}\mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\} - \frac{1}{\binom{n}{b}}\sum_{s=1}^{\binom{n}{b}}\mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\}\right| \\
&\quad + \left|\frac{1}{\binom{n}{b}}\sum_{j=1}^{\binom{n}{b}}\mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\} - J(0)\right| \\
&\leqslant \left|\frac{1}{K_n}\sum_{j=1}^{K_n}\mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\} - \frac{1}{\binom{n}{b}}\sum_{j=1}^{\binom{n}{b}}\mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\}\right| \\
&\quad + \left|\frac{1}{\binom{n}{b}}\sum_{j=1}^{\binom{n}{b}}\mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant r_b(\widehat{\theta} - \theta_0)\} - J(0)\right|.
\end{aligned}$$

Fix $t > 0$ such that $r_b t / r_n \leqslant r^*$. Define the event

$$\mathcal{E} := \{r_b(\widehat{\theta}_n - \theta_0) \leqslant r_b t / r_n\}.$$

On the event $\mathcal{E}$, we have

$$\begin{aligned}
&\left| \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant r_b(\widehat{\theta} - \theta_0)\} - J(0) \right| \\
&\quad \leqslant \left| \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant r_b t / r_n\} - J(0) \right| \\
&\qquad + \left| \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant -r_b t / r_n\} - J(0) \right| \\
&\quad \leqslant \left| \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant r_b t / r_n\} - J_b(r_b t / r_n) \right| + |J_b(r_b t / r_n) - J(r_b t / r_n)| + |J(r_b t / r_n) - J(0)| \\
&\qquad + \left| \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant -r_b t / r_n\} - J_b(-r_b t / r_n) \right| + |J_b(-r_b t / r_n) - J(-r_b t / r_n)| + |J(-r_b t / r_n) - J(0)| .
\end{aligned}$$

Because $r_b t / r_n \leqslant r^*$, assumption **(A2)** implies that

$$\max\{|J(r_b t / r_n) - J(0)|, |J(-r_b t / r_n) - J(0)|\} \leqslant \mathfrak{C} r_b t / r_n.$$

From assumption **(A1)**, we conclude

$$\max\{|J_b(r_b t / r_n) - J(r_b t / r_n)|, |J_b(-r_b t / r_n) - J(-r_b t / r_n)|\} \leqslant \delta_b.$$

Therefore, on the event $\mathcal{E}$,

$$\begin{aligned}
|\widehat{\Delta}_n - \Delta| \leqslant &\left| \frac{1}{K_n} \sum_{j=1}^{K_n} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\} - \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\} \right| \\
&+ \left| \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant r_b t / r_n\} - J_b(r_b t / r_n) \right| \\
&+ \left| \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant -r_b t / r_n\} - J_b(-r_b t / r_n) \right| \\
&+ 2\delta_b + 2\mathfrak{C} r_b t / r_n.
\end{aligned}$$

Observe that $\widehat{\theta}_b^{(j)}, 1 \leqslant j \leqslant K_n$ are independent and identically distributed random variables conditional on the data drawn from the finite population $\widehat{\theta}_b^{(j)}, 1 \leqslant j \leqslant \binom{n}{b}$. Corollary 1 of Massart (1990) implies that

$$\mathbb{P}\left(\left| \frac{1}{K_n} \sum_{j=1}^{K_n} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\} - \frac{1}{\binom{n}{b}} \sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \widehat{\theta}_n) \leqslant 0\} \right| \geqslant \sqrt{\frac{\log(2n)}{2K_n}} \right) \leqslant \frac{1}{n}.$$

Furthermore, note that $\binom{n}{b}\sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant -r_b t/r_n\}$ is a non-degenerate U-statistics of order b and with a kernel bounded between 0 and 1. Hence, Hoeffding's inequality for U-statistics (Hoeffding, 1963, inequality (5.7)) implies that

$$\mathbb{P}\left(\left|\frac{1}{\binom{n}{b}}\sum_{j=1}^{\binom{n}{b}} \mathbb{1}\{r_b(\widehat{\theta}_b^{(j)} - \theta_0) \leqslant r_b t/r_n\} - J_b(r_b t/r_n)\right| \geqslant \sqrt{\frac{\log(2n/b)}{2[n/b]}}\right) \leqslant \frac{b}{n}.$$

Hence, with probability at least $1 - \mathbb{P}(\mathcal{E}^c) - (b+1)/n$,

$$|\widehat{\Delta}_n - \Delta| \leqslant \sqrt{\frac{\log(2n)}{2K_n}} + \sqrt{\frac{\log(2n/b)}{2[n/b]}} + 2\delta_b + 2\mathfrak{C} r_b t/r_n.$$

Now, note that

$$\mathbb{P}(\mathcal{E}^c) = \mathbb{P}(r_b|\widehat{\theta}_n - \theta_0| > r_b t/r_n) \leqslant 2\delta_n + \mathbb{P}(|W| > t).$$

Therefore, with probability at least $1 - 2\delta_n - (b+1)/n - \mathbb{P}(|W| > t)$,

$$|\widehat{\Delta}_n - \Delta| \leqslant \sqrt{\frac{\log(2n)}{2K_n}} + \sqrt{\frac{\log(2n/b)}{2[n/b]}} + 2\delta_b + 2\mathfrak{C}\frac{r_b t}{r_n}.$$

S.10 Proof of Theorem 4

Define

$$\widehat{\theta}_{\max}^{(B)} := \max_{1\leqslant j\leqslant B} \widehat{\theta}_j, \quad \text{and} \quad \widehat{\theta}_{\min}^{(B)} := \min_{1\leqslant j\leqslant B} \widehat{\theta}_j.$$

Similarly, define $\widehat{\theta}_{\max}^{(B-1)}$ and $\widehat{\theta}_{\min}^{(B-1)}$. Set

$$F_{n,j}(u) = \mathbb{P}(r_{n,\alpha}(\widehat{\theta}_j - \theta_0) \leqslant u), \quad \text{and} \quad F_W(u) = \mathbb{P}(W \leqslant u).$$

The assumed hypothesis implies that $|F_{n,j}(u) - F_W(u)| \leqslant \delta_{n,\alpha}$ for all u and all $1 \leqslant j \leqslant B$. Finally, define the miscoverage probability

$$M_B := \mathbb{P}\left(\theta_0 \notin \left[\widehat{\theta}_{\max}^{(B)} - t(\widehat{\theta}_{\max}^{(B)} - \widehat{\theta}_{\min}^{(B)}), \widehat{\theta}_{\max}^{(B)} + t(\widehat{\theta}_{\max}^{(B)} - \widehat{\theta}_{\min}^{(B)})\right]\right).$$

We will prove that for all $B \geqslant 1$, $t \geqslant 0$, and $\Delta \in [0, 1/2]$,

$$\frac{M_B}{Q(B; t, \Delta)} \leqslant \frac{1}{(1 - 10B(1+t)\delta_{n,\alpha})_+}. \tag{E.23}$$

The same bound holds true for $M_{B-1}/Q(B-1; t, \Delta)$. The definition of $\eta_{\alpha,t}$ implies that

$$\eta_{\alpha,t} Q(B_{\alpha,t,\Delta}; t, \Delta) + (1 - \eta_{\alpha,t}) Q(B_{\alpha,t,\Delta} - 1; t, \Delta) = \alpha.$$

Combining this with the inequalities for M_B and M_{B-1}, the result is proved. Note that

$$\begin{aligned}
M_B &= \mathbb{P}(\theta_0 < \widehat{\theta}_{\min}^{(B)} - t(\widehat{\theta}_{\max}^{(B)} - \widehat{\theta}_{\min}^{(B)})) + \mathbb{P}(\theta_0 > \widehat{\theta}_{\max}^{(B)} + t(\widehat{\theta}_{\max}^{(B)} - \widehat{\theta}_{\min}^{(B)})) \\
&= \mathbb{P}\left(r_{n,\alpha}(\widehat{\theta}_{\min}^{(B)} - \theta_0) > \frac{t}{1+t} r_{n,\alpha}(\widehat{\theta}_{\max}^{(B)} - \theta_0)\right) + \mathbb{P}\left(r_{n,\alpha}(\widehat{\theta}_{\max}^{(B)} - \theta_0) < \frac{t}{1+t} r_{n,\alpha}(\widehat{\theta}_{\min}^{(B)} - \theta_0)\right) \\
&= \mathbf{I}_B + \mathbf{II}_B.
\end{aligned}$$

This implies that M_B can be written in terms of the smallest and largest order statistic of $r_{n,\alpha}(\widehat{\theta}_j - \theta_0)$, $1 \leqslant j \leqslant B$. Bounding $\mathbf{I}_B$ will also provide a bound for $\mathbf{II}_B$ by taking negative random variables $r_{n,\alpha}(\theta_0 - \widehat{\theta}_j)$. Under the assumption of continuous distribution for $r_{n,\alpha}(\widehat{\theta}_j - \theta_0)$, we get following the proof of Lanke (1974, Theorem 1) that

$$\mathbf{I}_B = \sum_{j=1}^{B} \int_0^{\infty} \prod_{i \neq j} (F_{n,i}(x) - F_{n,i}(tx/(1+t))) dF_j(x). \tag{E.24}$$

Recall that $F_{n,i}(x)$ and $F_W(x)$ are close and satisfy

$$F_{n,i}(x) - F_{n,i}(tx/(1+t)) \leqslant F_W(x) - F_W(tx/(1+t)) + 2\delta_{n,\alpha},$$

and because the distribution of W is unimodal at 0, we get

$$F_{n,i}(x) - F_{n,i}(tx/(1+t)) \leqslant \frac{F_W(x) - F_W(0)}{1+t} + 2\delta_{n,\alpha}.$$

This follows from the fact that unimodality implies $F_W(\cdot)$ is convex below 0 and concave above 0 implying $F(\lambda x) \geqslant F(0) + \lambda(F(x) - F(0))$ for $\lambda \in [0,1]$ and $x \geqslant 0$. Finally, using the closeness of $F_{n,j}(\cdot)$ and $F_W(\cdot)$ once again, we conclude

$$0 \leqslant F_{n,i}(x) - F_{n,i}(tx/(1+t)) \leqslant \frac{F_{n,j}(x) - F_{n,j}(0)}{1+t} + 4\delta_{n,\alpha}.$$

Substituting this inequality in (E.24), we obtain

$$\begin{aligned}
\mathbf{I}_B &\leqslant \sum_{j=1}^{B} \int_0^{\infty} \left(\frac{F_{n,j}(x) - F_{n,j}(0)}{1+t} + 4\delta_{n,\alpha} \right)^{B-1} dF_{n,j}(x) \\
&= \frac{1}{(1+t)^{B-1}} \sum_{j=1}^{B} \int_0^{\infty} \left(F_{n,j}(x) - F_{n,j}(0) + 4(1+t)\delta_{n,\alpha} \right)^{B-1} dF_{n,j}(x) \\
&= \frac{1}{(1+t)^{B-1}} \sum_{j=1}^{B} \int_{F_{n,j}(0)}^{1} \left(u - F_{n,j}(0) + 4(1+t)\delta_{n,\alpha} \right)^{B-1} du \\
&\leqslant \frac{1}{B(1+t)^{B-1}} \sum_{j=1}^{B} \left[\left(1 - F_{n,j}(0) + 4(1+t)\delta_{n,\alpha} \right)^{B} \right].
\end{aligned}$$

Applying the same calculations with $r_{n,\alpha}(\theta_0 - \widehat{\theta}_j)$ which has the distribution function $G_n(t) = 1 - F_n(-t)$ would yield

$$\mathbf{II}_B \leqslant \frac{1}{B(1+t)^{B-1}} \sum_{j=1}^{B} \left[\left(F_{n,j}(0) + 4(1+t)\delta_{n,\alpha} \right)^{B} \right].$$

Therefore,

$$\begin{aligned}
M_B &\leqslant \frac{1}{B(1+t)^{B-1}} \sum_{j=1}^{B} \left[\left(1 - F_{n,j}(0) + 4(1+t)\delta_{n,\alpha} \right)^{B} + \left(F_{n,j}(0) + 4(1+t)\delta_{n,\alpha} \right)^{B} \right] \\
&\leqslant \frac{1}{(1+t)^{B-1}} \left[\left(1 - F_W(0) + 5(1+t)\delta_{n,\alpha} \right)^{B} + \left(F_W(0) + 5(1+t)\delta_{n,\alpha} \right)^{B} \right].
\end{aligned}$$

The second inequality here follows again from the closeness of $F_{n,j}(0)$ and $F_W(0)$. From the continuous distribution assumption, the asymptotic median bias is given by $\Delta = |1/2 - \mathbb{P}(W \leqslant 0)|$. Hence, it follows

that

$$M_B \leqslant (1+t)^{-B+1}\left[\left(\frac{1}{2}+5(1+t)\delta_{n,\alpha}-\Delta\right)^B+\left(\frac{1}{2}+5(1+t)\delta_{n,\alpha}+\Delta\right)^B\right].$$

Now consider $M_B/Q(B;t,\Delta)$.

$$\frac{M_B}{Q(B;t,\Delta)} \;\leqslant\; \frac{(1-2\Delta+10(1+t)\delta_{n,\alpha})^B+(1+2\Delta+10(1+t)\delta_{n,\alpha})^B}{(1-2\Delta)^B+(1+2\Delta)^B}.$$

To bound the right hand side, consider the function $g(x)=(x+1-2\Delta)^B+(x+1+2\Delta)^B$ for $x\geqslant 0$. It is clear that

$$0 \;\leqslant\; g(x)-g(0) \;\leqslant\; \int_0^x g'(t)dt \;\leqslant\; Bx\left[(x+1-2\Delta)^{B-1}+(x+1+2\Delta)^{B-1}\right].$$

Furthermore,

$$\frac{(x+1-2\Delta)^{B-1}+(x+1+2\Delta)^{B-1}}{(x+1-2\Delta)^B+(x+1+2\Delta)^B} \leqslant \frac{1}{x+1}.$$

Hence, we conclude that $g(x)\leqslant g(0)/(1-Bx/(x+1))_+$ and

$$\frac{M_B}{Q(B;t,\Delta)} \;\leqslant\; \frac{1}{(1-10B(1+t)\delta_{n,\alpha})_+}.$$

This completes the proof of (E.23) and implies (39).