Honest Inference for Stochastic Optimization

Kenta Takatsu and Arun Kumar Kuchibhotla

Department of Statistics and Data Science, Carnegie Mellon University

Abstract

This manuscript studies a general approach for constructing confidence sets for the solutions to stochastic optimization, rendering empirical risk minimization as a special case. Statistical inference for stochastic optimization poses significant challenges due to the non-standard limiting behaviors of the corresponding estimators, which arise in settings with increasing dimensions of parameters, non-smooth objectives, or constraints. We study simple and unified methods that guarantee validity in both regular and irregular cases. We provide a unified treatment of validity, conservativeness, and diameter bounds of the resulting confidence sets. We study applications to several high-dimensional and irregular statistical problems. Numerical results for all statistical applications are provided.

Keywords— Honest inference, Adaptive inference, Irregular M-estimation, Non-standard asymptotics, Extremum estimators.

Contents

1 Introduction

Many statistical parameters of interest are defined as solutions to population-level optimization problems. Classical examples include maximum likelihood estimation, generalized linear models, regression, and density estimation. More recent examples include machine learning algorithms and penalized estimation in high-dimensional settings. A unified study of inference for optimization-based parameters, therefore, has broad applicability across statistics and machine learning.

Deferring formal definitions to Section 1.1, let θ_0 denote a minimizer of a population objective $\mathbb{M}_P(\theta)$, where $\theta \in \Theta$ is the parameter and the subscript P indicates dependence on the unknown data-generating distribution. One example is $\mathbb{M}_P(\theta) = \mathbb{E}_P[\ell(Z;\theta)]$ for a loss function ℓ with observation $Z \sim P$. This subsumes many risk-minimizing algorithms through different choices of ℓ, Θ, or dependence structure. The goal of this manuscript is to study *honest confidence sets* (3) for θ_0 under mild assumptions on P, the geometry of Θ, and the structure of $\mathbb{M}_P$.

Despite the apparent simplicity of the problem formulation, inference for general stochastic optimization remains challenging. To illustrate, consider $\mathbb{M}_P(\theta) = \mathbb{E}_P[\ell(Z;\theta)]$ with a unique minimizer θ_0, estimated by its empirical counterpart:

$$\widehat{\theta}_N = \arg\min_{\theta\in\Theta} \frac{1}{N}\sum_{i=1}^{N} \ell(Z_i;\theta) \quad \text{given} \quad \{Z_1,\ldots,Z_N\}. \tag{1}$$

Traditional inferential tools include (i) Wald methods and (ii) resampling methods, such as the bootstrap and subsampling. Both require that a scaled quantity $r_N(\widehat{\theta}_N - \theta_0)$ converges in distribution for some diverging r_N. Under strong regularity conditions and in low-dimensional problems, $\sqrt{N}(\widehat{\theta}_N - \theta_0)$ is asymptotically normal (Huber, 1967). These conditions, however, fail in many contemporary problems: when Θ is high-dimensional or constrained, the rate r_N or the limiting distributions of $\widehat{\theta}_N$ can be intractable. In such settings, neither Wald nor resampling methods remain viable.

1.1 Problem setup

Let $\{Z_1,\ldots,Z_N\}$ be a possibly dependent stationary sequence with common marginal $P \in \mathcal{P}$. Given a criterion $\mathbb{M} : \Theta \times \mathcal{P} \mapsto \mathbb{R}$ on a metric space $(\Theta, \|\cdot\|)$, write $\mathbb{M}_P(\theta) = \mathbb{M}(\theta, P)$ at fixed $P \in \mathcal{P}$. The parameter of interest is

$$\theta(P) := \arg\min_{\theta\in\Theta} \mathbb{M}_P(\theta), \tag{2}$$

which may be set-valued when the minimizer is not unique. The goal is to construct a confidence set $\widehat{\mathrm{CI}}_{N,\alpha}$ such that

$$\limsup_{N\to\infty} \sup_{P\in\mathcal{P}} \mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}) \leqslant \alpha \quad \text{where} \quad \mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}) := \sup_{\theta\in\theta(P)} \mathbb{P}_P(\theta \notin \widehat{\mathrm{CI}}_{N,\alpha}). \tag{3}$$

This is the standard notion of (asymptotic) *honesty* (Li, 1989) where validity holds uniformly over $\mathcal{P}$. Without such uniformity, confidence sets can be anti-conservative along certain sequences of distributions (Pötscher, 2002). On the other hand, finite-sample honesty is generally unattainable without strong restrictions on $\mathcal{P}$ (Bahadur and Savage, 1956). Although we target an asymptotic guarantee, we provide finite-sample bounds on $\mathrm{MC}_P(\cdot)$ for all methods, regardless of the dimension or complexity of Θ. This property has recently been termed *dimension-agnostic* (Kim and Ramdas, 2024). See also Robins and van der Vaart (2006).

1.2 Motivation and prior works

The primary goal is to study confidence sets that satisfy (3) without strong distributional assumptions. This precludes traditional approaches that rely on asymptotic normality or any specific limiting distribution. The following settings are representative examples where traditional methods fail, yet fall within the scope of the framework studied here.

1. **High-dimensional linear regression**: Inference for OLS is surprisingly complicated when dimension d is large. Asymptotic methods fail due to a bias of order $d/N^{1/2}$ (Mammen, 1993). Bias-corrected alternatives exist, but typically require growth conditions on d (Cattaneo et al., 2019; Chang et al., 2023).

2. **Non-smooth objective and constrained optimization**: When $\mathbb{M}_P(\theta)$ is non-smooth, the limiting distribution of the empirical minimizer can be non-standard (Knight, 1998). A canonical example is quantile regression, whose limiting distribution depends on the unknown smoothness of the response CDF. Structural constraints such as sparsity or shape restrictions also complicate inference. This includes popular regularized minimizers, such as the Lasso, ridge regression, and the Dantzig selector, for which existing inference methods require strong assumptions (van de Geer et al., 2014).

3. **Cube-root problems**: Kim and Pollard (1990) identify a class of problems where the empirical minimizer exhibits *cube-root asymptotics*, including Manski's estimator, the Grenander estimator, and classification problems. Inference for these problems is challenging, and the bootstrap is inconsistent (Sen et al., 2010). Modified problem-specific procedures have been proposed (Cattaneo et al., 2020).

Contributions. This manuscript studies alternative inference based on two classical ideas: *risk inversion* and *sample splitting*, which can be implemented without reference to the limiting distribution of an estimator. The confidence sets studied here are not new: risk inversion for irregular inference appears as early as Stein (1981). See Section S.1 of KT26 for further historical accounts on this subject. Sample splitting has seen renewed interest in several

recent works (Robins and van der Vaart, 2006; Wasserman et al., 2020; Kim and Ramdas, 2024). The primary contribution is a new analysis of these ideas under greater generality: (1) a systematic dimension- and complexity-agnostic validity guarantee; (2) diameter bounds under mild conditions; (3) applications to challenging inference problems. Together, these results advance honest and adaptive inference for stochastic optimization.

Remark 1. *Sample splitting also appears in double machine learning (DML) (Chernozhukov et al., 2018), but its role there is fundamentally different. At its core, DML relies on influence-function expansions and asymptotic normality, whereas the framework here is designed precisely to avoid such assumptions.*

Organization. The manuscript is organized as follows. Section 2 defines three confidence sets to be studied; Section 3 presents general results for dimension-agnostic honest validity; Section 4 provides diameter bounds; Section 5 provides applications to high-dimensional mean, regularized regression, and Manski's discrete choice model with both theoretical and empirical results. Concluding remarks appear in Section 6. This manuscript is a condensed version of Takatsu and Kuchibhotla (2026), henceforth referred to as KT26. Formal proofs, additional references, and additional results appear in KT26.

2 Construction for General Stochastic Optimization

The intuition behind the construction is simple. Any $\theta_0 \in \theta(P)$ always belongs to the oracle set $\{\theta \in \Theta : \mathbb{M}_P(\theta) \leqslant \mathbb{M}_P(\widehat{\theta})\}$ for any $\widehat{\theta} \in \Theta$. This set is infeasible since $\mathbb{M}_P$ is unknown. Given an estimator of the objective function $\theta \mapsto \widehat{\mathbb{M}}(\theta)$, Vogel (2008) consider sets of the form $\{\theta \in \Theta : \widehat{\mathbb{M}}(\theta) \leqslant \widehat{\mathbb{M}}(\widehat{\theta}) + \gamma_{\alpha,N}\}$. This construction involves two estimators: $\widehat{\mathbb{M}}(\theta)$ and $\widehat{\theta}$. The validity of this set depends on the complexity of Θ when $\widehat{\mathbb{M}}(\theta)$ and $\widehat{\theta}$ are constructed from the same data. Sample splitting reduces this dependence, even under dependent observations. We formalize the idea below.

Sample splitting. Let $N \geqslant 1$ and partition $\{1, \dots, N\}$ as:

$$I_1 = \{1, \dots, n\}, \quad I_{\text{gap}} = \{n+1, \dots, n+r\}, \quad \text{and} \quad I_2 = \{n+r+1, \dots, N\}, \tag{4}$$

where $n, r \geqslant 0$. Throughout the rest of the manuscript, we assume that r satisfies $N = 2n + r$ so $|I_1| = |I_2| = n$, but the framework allows more general partitions. Write $D_\ell = \{Z_i : i \in I_\ell\}$ for $\ell = 1, 2$. The gap I_{gap} is discarded to reduce dependence between D_1 and D_2 under mixing; setting $r = 0$ suffices under independence. From D_1, construct *any* estimator $\widehat{\theta}_1 := \widehat{\theta}_1(D_1) \in \Theta$. From D_2, construct an estimated criterion, $\widehat{\mathbb{M}}_2(\theta) := \widehat{\mathbb{M}}_2(\theta; D_2)$.

2.1 Definition of confidence sets

Fixing $\alpha \in (0, 1)$, we define three confidence sets.

Median set. The *median set* uses a zero threshold:

$$\widehat{\mathrm{CI}}_N^{\texttt{Med}} := \left\{\theta \in \Theta : \widehat{\mathbb{M}}_2(\theta) - \widehat{\mathbb{M}}_2(\widehat{\theta}_1) \leqslant 0\right\}. \tag{5}$$

The CLT-based set. Let $\widehat{V}_{\theta,\theta'}$ be any estimator of the variance of $\widehat{\mathbb{M}}_2(\theta) - \widehat{\mathbb{M}}_2(\theta')$ for $\theta, \theta' \in \Theta$. The *CLT-based set* inflates the threshold by a normal quantile:

$$\widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,\alpha} := \left\{\theta \in \Theta : \widehat{\mathbb{M}}_2(\theta) - \widehat{\mathbb{M}}_2(\widehat{\theta}_1) \leqslant z_{1-\alpha}\widehat{V}^{1/2}_{\theta,\widehat{\theta}_1}\right\} \quad \text{where} \quad \mathbb{P}(N(0,1) \leqslant z_q) = q. \tag{6}$$

HulC. The *Hull-based confidence set* (Kuchibhotla et al., 2024), or HulC, uses set aggregation. Take $B := \lceil \log_2(1/\alpha) \rceil$, and partition $\{1, \ldots, N\}$ into $B+1$ consecutive bins $I_1, I^{(1)}_{\mathrm{gap}}, I^{(1)}_2, \ldots, I^{(B)}_{\mathrm{gap}}, I^{(B)}_2$. For simplicity, we assume that $|I^{(j)}_{\mathrm{gap}}| = r$ for all j with $|I_1| = |I^{(j)}_2| = m$ satisfying $N = Br + (B+1)m$. The framework allows more general partitions. Construct $\widehat{\theta}_1$ using $\{Z_i : i \in I_1\}$. Then, construct the median set using an estimated criterion $\widehat{\mathbb{M}}^{(j)}_2(\theta)$ based on $\{Z_i : i \in I^{(j)}_2\}$ for $1 \leqslant j \leqslant B$, with $\widehat{\theta}_1$ shared across the B constructions. The final set is a union of B such sets

$$\widehat{\mathrm{CI}}^{\mathtt{Hull}}_{N,\alpha} := \bigcup_{j=1}^{B} \widehat{\mathrm{CI}}^{\mathtt{Med},(j)}_{m}. \tag{7}$$

The number of aggregated sets B is determined by α alone with no hidden constants and does not diverge with N; for instance, $B = 4$ at $\alpha = 0.1$. The CLT-based set requires estimating $\widehat{V}_{\theta,\widehat{\theta}_1}$, while HulC avoids variance estimation entirely. The forthcoming section establishes that the CLT-based set and HulC satisfy (3) at level α; the median set is valid at level $1/2$. We present below a simple result comparing the three sets.

Theorem 1. *The following deterministic results hold.*

$$\bigcup_{\alpha \in [1/2,1]} \widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,\alpha} = \widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,1/2} = \widehat{\mathrm{CI}}^{\mathtt{Med}}_{N} = \bigcap_{\alpha \in [\alpha_0, 1/2]} \widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,\alpha} \quad \forall \alpha_0 \leqslant \frac{1}{2}. \tag{8}$$

If $\widehat{\Theta}_2 = \arg\min_{\theta \in \Theta} \widehat{\mathbb{M}}_2(\theta)$, *then*

$$\{\widehat{\theta}_1\} \cup \widehat{\Theta}_2 \subseteq \widehat{\mathrm{CI}}^{\mathtt{Med}}_{N} \subseteq \widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,\alpha} \quad \forall \alpha \leqslant \frac{1}{2}. \tag{9}$$

In particular,

$$\begin{gathered}
\mathrm{MC}_P(\widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,1/2-\delta}) \leqslant \mathrm{MC}_P(\widehat{\mathrm{CI}}^{\mathtt{Med}}_{N}) \leqslant \mathrm{MC}_P(\widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,1/2+\delta}) \quad \forall \delta \in [0, 1/2] \\
\mathrm{Diam}(\{\widehat{\theta}_1\} \cup \widehat{\Theta}_2) \leqslant \mathrm{Diam}(\widehat{\mathrm{CI}}^{\mathtt{Med}}_{N}) \leqslant \mathrm{Diam}(\widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,1/2-\delta}) \quad \forall \delta \in [0, 1/2] \quad \textit{and} \\
\mathrm{Diam}(\widehat{\mathrm{CI}}^{\mathtt{Hull}}_{N,\alpha}) \leqslant 2 \max_{1 \leqslant j \leqslant B} \mathrm{Diam}(\widehat{\mathrm{CI}}^{\mathtt{Med},(j)}_{m}),
\end{gathered} \tag{10}$$

where the diameter of a set A in $(\Theta, \|\cdot\|)$ is defined as $\mathrm{Diam}_{\|\cdot\|}(A) := \sup\{\|a-b\| : a, b \in A\}$. *The result holds for general norm $\|\cdot\|$ and we write* $\mathrm{Diam}(\cdot) = \mathrm{Diam}_{\|\cdot\|}(\cdot)$.

Theorem 1 follows from the definitions and provides deterministic relationships among the three confidence sets. We frequently refer to these geometric properties.

Remark 2. *An alternative construction based on concentration inequalities, such as Hoeffding's, Bernstein's or Bennett's inequality, is possible under additional tail structures of* $\widehat{\mathbb{M}}_2(\theta) - \widehat{\mathbb{M}}_2(\widehat{\theta}_1)$*; see Section S.10 of KT26.*

3 Honest Coverage Guarantees

To state the validity results, we introduce several key objects.

Mixing coefficient. For a sequence $\{Z_t\}_{t=1}^{\infty}$, define

$$\beta(k) := \sup_{i \geqslant 1} \mathbb{E}\left[\sup\left\{\left|\mathbb{P}(A \mid \mathcal{F}_{-\infty}^{i}) - \mathbb{P}(A)\right| \,:\, A \in \mathcal{F}_{i+k}^{\infty}\right\}\right] \tag{11}$$

where $\mathcal{F}_i^j$ is the σ-algebra generated by $\{Z_i, \ldots, Z_j\}$. Under independence, $\beta(k) = 0$ for all $k \geqslant 0$, and under m-dependence (Hoeffding and Robbins, 1948), $\beta(k) = 0$ for all $k \geqslant m$. Many relevant processes, including ARMA/GARCH models, satisfy $\beta(k) \to 0$ as $k \to \infty$; see Bradley (2005). Mixing sequences can be non-stationary (Fryzlewicz and Rao, 2011).

Curvature and MSE. For $\theta \in \Theta$, define the curvature and MSE as

$$\begin{aligned} \mathbb{C}_P(\theta) &:= \mathbb{M}_P(\theta) - \inf_{\theta_0 \in \theta(P)} \mathbb{M}_P(\theta_0) \quad \text{and} \\ \mathbb{V}_P(\theta) &:= \sup_{\theta_0 \in \theta(P)} \mathbb{E}_P[|(\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta) - (\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta_0)|^2]. \end{aligned} \tag{12}$$

The difficulty of estimating the minimizer is characterized by the behavior of $\theta \mapsto \Delta_P(\theta)$, where

$$\Delta_P(\theta) := \mathbb{C}_P(\theta)/\sqrt{\mathbb{V}_P(\theta)} \geqslant 0. \tag{13}$$

Larger $\Delta_P(\theta)$ corresponds to easier estimation. Estimation is easier when $\mathbb{C}_P(\theta)$ is large and $\mathbb{V}_P(\theta)$ is small; see Figures 1a and 1b. These quantities are defined for fixed $\theta \in \Theta$; evaluated at random $\widehat{\theta}_1 \in \Theta$, they become random variables. We write

$$\widehat{\mathbb{C}}_P = \mathbb{C}_P(\widehat{\theta}_1), \quad \widehat{\mathbb{V}}_P = \mathbb{V}_P(\widehat{\theta}_1) \quad \text{and} \quad \widehat{\Delta}_P = \widehat{\mathbb{C}}_P/\widehat{\mathbb{V}}_P^{1/2}. \tag{14}$$

A special instance of our framework is M-estimation, where

$$\mathbb{M}_P(\theta) = \mathbb{E}_P[\ell(Z;\theta)] \quad \text{and} \quad \widehat{\mathbb{M}}_2(\theta) = n^{-1}\sum_{i \in I_2} \ell(Z_i;\theta). \tag{15}$$

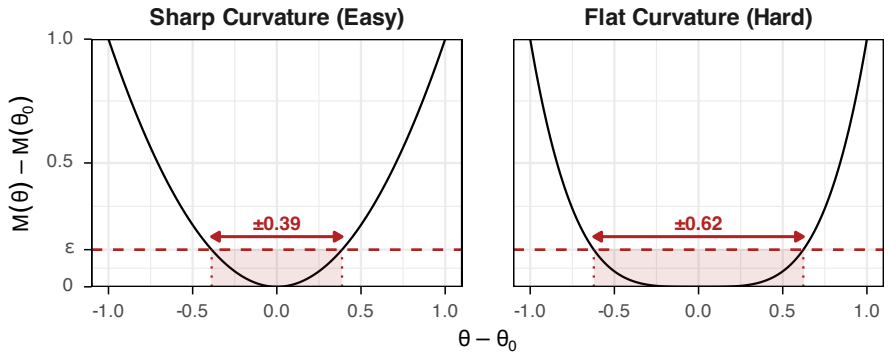
To provide intuition, we compute $\widehat{\mathbb{C}}_P$, $\widehat{\mathbb{V}}_P$, and $\widehat{\Delta}_P$ for two simple M-estimation problems.

Example 1. ***Mean estimation.*** *Let $Z_1, \ldots, Z_N$ have mean θ_0 and covariance Σ where $\ell(Z_i;\theta) = \|Z_i - \theta\|_{\Sigma^{-1}}^2$. Under independence,*

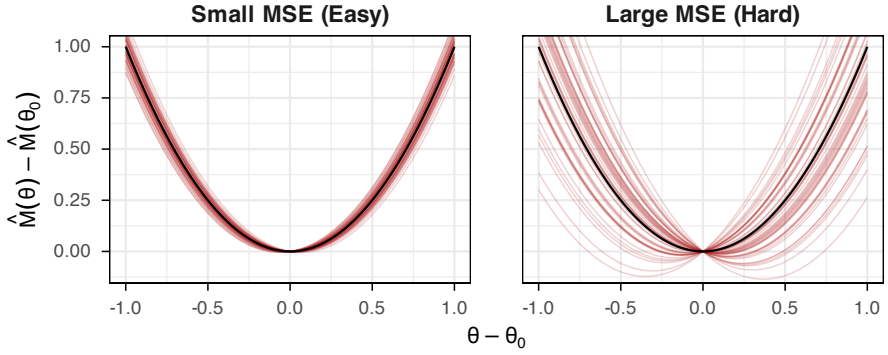
$$\mathbb{C}_P(\theta) = \|\theta - \theta_0\|_{\Sigma^{-1}}^2, \quad \mathbb{V}_P(\theta) = \frac{4\|\theta - \theta_0\|_{\Sigma^{-1}}^2}{n}, \quad \textit{and} \quad \Delta_P(\theta) = \frac{\sqrt{n}\|\theta - \theta_0\|_{\Sigma^{-1}}}{2}. \tag{16}$$

Maximum likelihood. *Let θ_0 minimize the KL divergence to the working parametric model; that is, $\ell(Z_i;\theta) = -\log p(Z_i;\theta)$. Under standard regularity conditions and independence, locally near θ_0,*

$$\mathbb{C}_P(\theta) \asymp \frac{\|\theta - \theta_0\|_{\mathcal{I}_P}^2}{2}, \quad \mathbb{V}_P(\theta) \asymp \frac{\|\theta - \theta_0\|_{\mathcal{I}_P}^2}{n}, \quad \textit{and} \quad \Delta_P(\theta) \asymp \frac{\sqrt{n}\|\theta - \theta_0\|_{\mathcal{I}_P}}{2}, \tag{17}$$



(a) Large (left) and small (right) curvature near θ_0. The shaded region identifies $\{\theta : \mathbb{M}_P(\theta) - \mathbb{M}_P(\theta_0) \leqslant \varepsilon\}$ for a margin $\varepsilon \geqslant 0$. Large curvature concentrates this region tightly around θ_0, so a modest ε rules out distant parameters. Small curvature means the criterion is nearly flat: one must take ε much closer to zero to achieve the same localization.



(b) Small (left) and large (right) MSE near θ_0. When MSE is small, $\widehat{\mathbb{M}}_2$ is a better estimator of the population objective $\mathbb{M}_P$. Hence, the corresponding empirical minimizer is more concentrated near θ_0.

Figure 1: Pictorial description of the roles of curvature and MSE for $\theta(P) = \{\theta_0\}$.

where $\mathcal{I}_P$ denotes the Fisher information matrix.

The curvature bound holds under dependence without modification. The variance under β-mixing follows from the covariance inequality for β-mixing sequences (Rio, 2017, Eq. 1.12b). In both cases, $\widehat{\Delta}_P \asymp \sqrt{n}\|\widehat{\theta}_1 - \theta_0\|/2$ in the appropriate norm. The ratio $\widehat{\Delta}_P$ is large when $\widehat{\theta}_1$ converges slowly, and tends to zero when $\widehat{\theta}_1$ converges strictly faster than $\sqrt{n}$.

Proposition 1. *If $\widehat{\theta}_1$ is inconsistent for $\theta(P)$, i.e., $\widehat{\mathbb{C}}_P \neq o_p(1)$, and $\theta \mapsto \widehat{\mathbb{M}}_2(\theta)$ converges in L_2 to $\theta \mapsto \mathbb{M}_P(\theta)$ for each θ, then $\widehat{\Delta}_P \xrightarrow{p} \infty$.*

Proof. This follows directly from the definition $\widehat{\Delta}_P = \widehat{\mathbb{C}}_P/\widehat{\mathbb{V}}_P^{1/2}$. □

3.1 Honest validity for hull-based confidence sets

Theorem 2. *The following bounds for miscoverage (3) hold:*

$$\begin{aligned} \mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}}) &\leqslant \mathbb{E}_P\left[\min\left\{1, \frac{1}{\widehat{\Delta}_P^2}\right\}\right] + \beta(r) \quad \textit{and} \\ \mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{Hull}}) &\leqslant \mathbb{E}_P[p_m^{\log_2(1/\alpha)}] + \log_2(2/\alpha)\beta(r) \\ &\leqslant \alpha \times \mathbb{E}_P[\exp\left(2\log_2(2/\alpha)\{p_m - 1/2\}_+\right)] + \log_2(2/\alpha)\beta(r). \end{aligned} \tag{18}$$

where $p_m = \sup_{\theta_0 \in \theta(P)} \mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}_m^{\texttt{Med}} \mid D_1)$.

Theorem 3. *Assume $\widehat{\mathbb{M}}_2(\theta)$ is unbiased, i.e., $\mathbb{E}_P[\widehat{\mathbb{M}}_2(\theta)] = \mathbb{M}_P(\theta)$. Then*

$$\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}}) \leqslant \mathbb{E}_P\left[\frac{1}{1+\widehat{\Delta}_P^2}\right] + \beta(r). \tag{19}$$

Both results are general: they are finite-sample and do not require uniqueness of $\theta(P)$, independence of observations, or any structure on $\mathbb{M}_P$. In particular, no smoothness of $\mathbb{M}_P$ is assumed, so the results apply to constrained optimization where $\mathbb{M}_P$ may be discontinuous on the boundary of Θ. The only requirement is finite second moment of $\widehat{\mathbb{M}}_2 - \mathbb{M}_P$ so that $\widehat{\Delta}_P$ is well-defined, and even this can be relaxed; see Remark 3 below.

The miscoverage bounds depend on $\widehat{\Delta}_P$. When $\widehat{\Delta}_P \xrightarrow{p} \infty$ (see, for instance, Proposition 1) the median set and HulC are valid at *any* level α since $\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}})$ and $\mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{Hull}})$ both tend to zero, meaning they are conservative. Validity of HulC at level α requires the median set to be valid at level $1/2$. This holds whenever $\widehat{\Delta}_P^2 \geqslant 2$ under Theorem 2 and $\widehat{\Delta}_P^2 \geqslant 1$ under Theorem 3.

Theorems 2 and 3 establish honest validity (3) for general settings whenever $\widehat{\Delta}_P$ is sufficiently large. A limitation is that validity is not implied for small $\widehat{\Delta}_P$. Section 3.2 establishes validity for all $\widehat{\Delta}_P \geqslant 0$, under additional moment assumptions.

Remark 3. *Theorems 2 and 3 rely only on Chebyshev's and Cantelli's inequalities and therefore require finite variance. Under finite $1+\delta$ moments, i.e., $\mathbb{E}_P[|(\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta) - (\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta_0)|^{1+\delta}] < \infty$, analogous bounds follow from Markov's inequality with a minor modification to the proof.*

3.2 Honest validity for CLT-based sets

We next establish validity of the CLT-based set. First, we place the following assumption.

Assumption 1 (Non-uniform Gaussian Approximation)**.** *Let* $\bar{\Phi}(t) = \mathbb{P}(N(0,1) \geqslant t)$*. For any* $\theta, \theta' \in \Theta$*, there exists a universal constant* $C \geqslant 1$*, a* P*-dependent function* $K_P(\theta, \theta')$*, a scaling* $\widetilde{V}_{\theta,\theta'}$ *(possibly random), and* $s \geqslant 0$*, such that*

$$\left| \mathbb{P}\left(\frac{(\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta) - (\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta')}{\widetilde{V}_{\theta,\theta'}^{1/2}} \geqslant t \right) - \bar{\Phi}(t) \right| \leqslant \frac{C K_P(\theta, \theta')}{\sqrt{n}(1 + |t|^s)}. \tag{20}$$

This assumption requires asymptotic normality of the *estimated objective* $\widehat{\mathbb{M}}_2(\theta) - \widehat{\mathbb{M}}_2(\theta')$, not of the minimizer. This distinction is important: the former requires far weaker assumptions. Since Assumption 1 concerns a univariate quantity, the approximation error decays as $n \to \infty$ regardless of the dimension of Θ, in contrast to asymptotic normality of the minimizer, which typically requires low-dimensional Θ.

Assumption 1 is satisfied for many estimators $\widehat{\mathbb{M}}_2$; results for general statistics are available with $s = 0$ under dependent data (Bentkus et al., 1997), and $s = 3$ under independent data (Chen and Shao, 2007); see also Haeusler and Joos (1988); Hafouta (2022). For random scaling $\widetilde{V}_{\theta,\hat{\theta}_1}$, self-normalization results are also extensive (Bentkus and Götze, 1996; Bentkus et al., 1997; Fan and Shao, 2018).

For M-estimation (15), Assumption 1 holds under classical conditions. Suppose

$$\mathrm{Var}_P(\ell(Z;\theta) - \ell(Z;\theta')) \geqslant C_1, \quad \mathbb{E}_P[|\ell(Z;\theta) - \ell(Z;\theta')|^3] \leqslant C_2, \quad \beta(r) \leqslant C_3 \exp(-C_4 r).$$

Then under β-mixing, Assumption 1 holds with $K_P(\theta,\theta') = (\log n)^2$, $\widetilde{V}_{\theta,\theta'} = \mathrm{Var}_P\big(\widehat{\mathbb{M}}_2(\theta) - \widehat{\mathbb{M}}_2(\theta')\big)$, and $s = 0$ (Tikhomirov, 1981, Theorem 2); under an additional condition $\mathbb{E}_P[|\ell(Z;\theta) - \ell(Z;\theta')|^{4+\delta}] \leqslant C_5$ for some $\delta > 0$, Assumption 1 holds with $K_P(\theta,\theta') = (\log n)^3$ and $s = 4$ (Tikhomirov, 1981, Theorem 4). For IID observations, Assumption 1 holds with

$$K_P(\theta,\theta') = \frac{\mathbb{E}_P[|\ell(Z;\theta) - \ell(Z;\theta')|^3]}{\mathrm{Var}_P\{\ell(Z;\theta) - \ell(Z;\theta')\}^{3/2}}, \quad \widetilde{V}_{\theta,\theta'} = \mathrm{Var}_P\big(\widehat{\mathbb{M}}_2(\theta) - \widehat{\mathbb{M}}_2(\theta')\big), \quad \text{and} \quad s = 3.$$

Theorem 4. *Suppose Assumption 1 holds with* $\widetilde{V}_{\theta,\theta'} = \mathrm{Var}_P\big(\widehat{\mathbb{M}}_2(\theta) - \widehat{\mathbb{M}}_2(\theta')\big)$ *for* $s \geqslant 0$*. Also suppose* $\widehat{\mathbb{M}}_2(\theta)$ *is unbiased for* $\mathbb{M}_P(\theta)$ *for all* $\theta \in \Theta$*. Then,*

$$\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}}) \leqslant \sup_{\theta_0 \in \theta(P)} \mathbb{E}_P\left[\bar{\Phi}(\hat{\Delta}_P) + \frac{C K_P(\theta_0, \hat{\theta}_1)}{\sqrt{n}(1 + |\hat{\Delta}_P|^s)} \right] + \beta(r). \tag{21}$$

Suppose instead Assumption 1 holds with $\widetilde{V}_{\theta,\theta'} = \widehat{V}_{\theta,\theta'}$ *used in* (6) *and* $s = 0$*. Then*

$$\mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{CLT}}) \leqslant \alpha + \frac{C}{\sqrt{n}} \sup_{\theta_0 \in \theta(P)} \mathbb{E}_P[K_P(\theta_0, \hat{\theta}_1)] + \beta(r) \quad \textit{for all} \quad \alpha \in (0,1). \tag{22}$$

Note that for $\alpha \leqslant 1/2$, it holds that $\widehat{\mathrm{CI}}_N^{\texttt{Med}} \subseteq \widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{CLT}}$ and hence $\mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{CLT}}) \leqslant \mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}})$. In this case, $\mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{CLT}})$ tends to zero whenever $\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}})$ tends to zero. Bound (22) can be improved using (21) and Theorem 1.

Since $\widehat{\Delta}_P \geqslant 0$ almost surely, $\bar{\Phi}(\widehat{\Delta}_P) \leqslant 1/2$. Writing the bound (21) as $1/2+\mathrm{Rem}_{P,N}+\beta(r)$, Theorems 2 and 4 give

$$\begin{aligned}
\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}}) &\leqslant \frac{1}{2} + \mathrm{Rem}_{P,N} + \beta(r) \\
\mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{CLT}}) &\leqslant \alpha + \frac{C}{\sqrt{n}} \sup_{\theta_0 \in \theta(P)} \mathbb{E}_P[K_P(\theta_0, \widehat{\theta}_1)] + \beta(r) \\
\mathrm{MC}_P(\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{Hull}}) &\leqslant \alpha \times \mathbb{E}_P[\exp(2\log_2(2/\alpha)\mathrm{Rem}_{P,N})] + \log_2(2/\alpha)\beta(r).
\end{aligned} \tag{23}$$

All three sets therefore satisfy (3) as long as the remainder terms are negligible. Unlike Theorems 2 and 3, validity holds for all $\widehat{\Delta}_P \geqslant 0$. When $s > 0$, large $\widehat{\Delta}_P$ further accelerates the decay of the remainder term.

Remark 4. *Assumption 1 is stated at rate $n^{-1/2}$, which often requires strong moment assumptions. Under weaker moment assumptions, analogous results hold at a slower rate (Katz, 1963; Tikhomirov, 1981; Haeusler and Joos, 1988).*

3.3 Summary of validity results and illustration

Assuming $\theta(P) = \{\theta_0\}$ for brevity, Theorems 2 to 4 together give

$$\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}}) \leqslant \begin{cases} \mathbb{E}_P[\bar{\Phi}(\widehat{\Delta}_P)] + \mathrm{Rem}_{P,N} + \beta(r) \\ \mathbb{E}_P[1/(1+\widehat{\Delta}_P^2)] + \beta(r) \\ \mathbb{E}_P[\min\{1, 1/\widehat{\Delta}_P^2\}] + \beta(r). \end{cases} \tag{24}$$

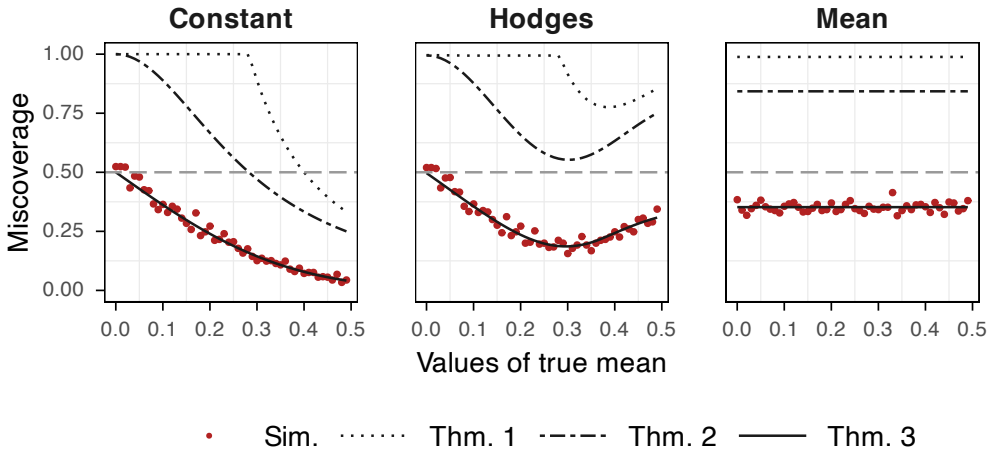
Since $\mathbb{E}_P[\bar{\Phi}(\widehat{\Delta}_P)] \leqslant \mathbb{E}_P[1/(1+\widehat{\Delta}_P^2)] \leqslant \mathbb{E}_P[\min\{1, 1/\widehat{\Delta}_P^2\}]$, each theorem provides a progressively tighter bound under progressively stronger assumptions. When $\widehat{\Delta}_P \xrightarrow{p} 0$, $\mathbb{E}_P[\bar{\Phi}(\widehat{\Delta}_P)] \to 1/2$, and Theorem 4 recovers an exact bound for the miscoverage. The following examples illustrate the role of $\widehat{\Delta}_P$ and the tightness of each bound.

Example 2 (Gaussian mean)**.** *Let $N = 100$, $n = 50$, $r = 0$, and $Z_1, \ldots, Z_{100}$ IID from $N(\theta_0, 1)$. We study the median set for θ_0 by letting $\mathbb{M}_P(\theta) = \mathbb{E}_P[(Z-\theta)^2]$ and $\widehat{\mathbb{M}}_2(\theta) = n^{-1}\sum_{i\in I_2}(Z_i - \theta)^2$. Three choices of $\widehat{\theta}_1$ are considered: (1) the constant estimator $\widehat{\theta}_1 = 0$; (2) Hodges' estimator $\widehat{\theta}_1 = \bar{Z}_n \mathbf{1}\{|\bar{Z}_n| \geqslant n^{-1/4}\}$ where $\bar{Z}_n = n^{-1}\sum_{i\in I_1} Z_i$; and (3) the sample mean $\widehat{\theta}_1 = \bar{Z}_n$. Assumption 1 holds and Theorem 4 gives*

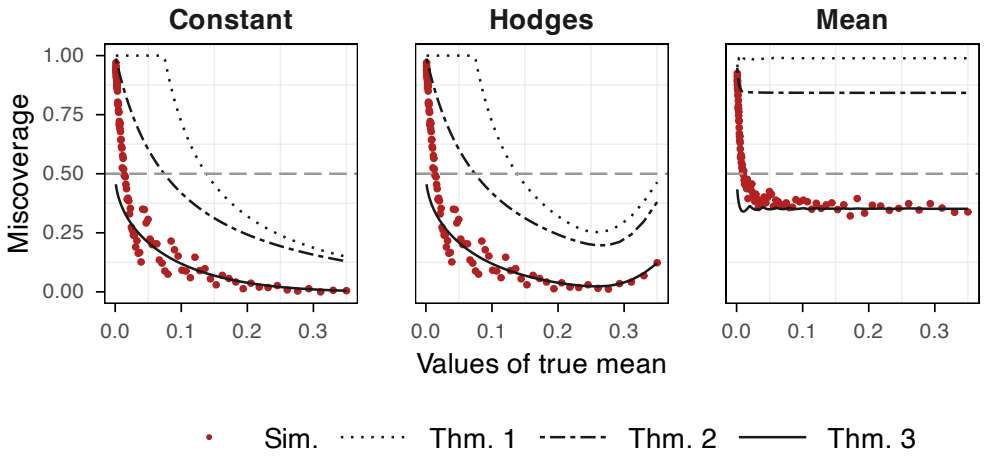
$$\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}}) \leqslant \mathbb{E}_P[\bar{\Phi}(\widehat{\Delta}_P)] + \mathbb{E}_P\left[\frac{C\sigma^{3/2}}{\sqrt{n}(1+\widehat{\Delta}_P^3)}\right]. \tag{25}$$

Direct calculation gives $\widehat{\Delta}_P^2 = n(\widehat{\theta}_1 - \theta_0)^2/4$ and, in fact, $\widehat{\Delta}_P^2 \sim \chi_1^2/4$ when $\widehat{\theta}_1$ is the sample mean.

Figure 2a displays empirical miscoverage of $\widehat{\mathrm{CI}}_N^{\texttt{Med}}$ over 200 replications (shown in dots) and analytic values for $\mathbb{E}_P[\min\{1, 1/\widehat{\Delta}_P^2\}]$, $\mathbb{E}_P[1/(1+\widehat{\Delta}_P^2)]$ and $\mathbb{E}_P[\bar{\Phi}(\widehat{\Delta}_P)]$ (labeled Thm. 1, 2, 3) for $\theta_0 \in [0, 0.5]$. We observe Theorem 4 tracks the miscoverage closely for all θ_0 and $\widehat{\theta}_1$,



(a) Empirical miscoverage under $N(\theta_0, 1)$. Since Assumption 1 holds, Theorem 4 tracks the miscoverage closely, confirming validity at level 1/2 across all settings. When $\widehat{\theta}_1$ is super-efficient (e.g., Hodges' estimator near $\theta_0 = 0$), miscoverage approaches 1/2. When $\widehat{\theta}_1$ performs poorly (e.g., constant estimator for $\theta_0 > 0$), miscoverage tends to zero.



(b) Empirical miscoverage under Bern(θ_0). For θ_0 near zero, Assumption 1 fails. Therefore, Theorem 4 is not a valid upper bound and miscoverage can exceed 1/2. For θ_0 bounded away from zero, Theorem 4 tracks the miscoverage closely. Since $\widehat{\mathbb{M}}_2$ is unbiased, Theorem 3 remains a valid finite-sample bound throughout.

Figure 2: Empirical miscoverage of $\widehat{\mathrm{CI}}_N^{\mathtt{Med}}$ over 200 replications, shown in dots, alongside analytic bounds from Thm.2, 3, and 4.

with $\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\mathtt{Med}}) \leqslant 1/2$ *up to the remainder. Theorems 2 and 3 provide valid but loose upper bounds. The role of* $\widehat{\Delta}_P$ *is also transparent. When* $\widehat{\theta}_1$ *converges faster than* $\sqrt{n}$*, i.e., super-efficient,* $\widehat{\Delta}_P \xrightarrow{p} 0$ *and miscoverage approaches* $1/2$*. This is visible for Hodges' estimator near* $\theta_0 = 0$*. When* $\widehat{\theta}_1$ *performs poorly,* $\widehat{\Delta}_P \xrightarrow{p} \infty$ *and miscoverage tends to zero, visible for the constant estimator when* $\theta_0 > 0$*.*

Example 3 (Bernoulli mean)**.** *Consider the identical setup to Example 2 with* $\{Z_1, \ldots, Z_{100}\}$ *IID* $\mathrm{Bern}(\theta_0)$ *for* $\theta_0 \in [0, 0.35]$*. Figure 2b displays the results. For small* θ_0*, the Poisson limit applies and Assumption 1 fails; Figure 2b confirms that Theorem 4 is not a uniformly valid upper bound and empirical miscoverage of* $\widehat{\mathrm{CI}}_N^{\mathtt{Med}}$ *exceeds* $1/2$ *in this regime. For* θ_0 *bounded away from zero, Assumption 1 holds and Theorem 4 tracks the miscoverage closely. Since* $\widehat{\mathbb{M}}_2$ *is unbiased regardless of whether* θ_0 *is small, Theorem 3 provides a valid finite-sample bound.*

We conclude by summarizing Theorems 2 to 4 in Table 1 below. All results hold for general stochastic optimization under dependent observations, without restriction to M-estimation or any specific structure on $\mathbb{M}_P$. The validity is established by either checking $\widehat{\Delta}_P$ is large or Assumption 1 holds, both of which require case-by-case analysis.

	Bound	Assumption	Median validity	Asympt. Exactness
Thm. 2	$\mathbb{E}_P[\min\{1, 1/\widehat{\Delta}_P^2\}]$	Finite variance	$(2 - \widehat{\Delta}_P^2)_+ \xrightarrow{p} 0$	No
Thm. 3	$\mathbb{E}_P[1/(1 + \widehat{\Delta}_P^2)]$	Finite variance; Unbiased $\widehat{\mathbb{M}}_2(\cdot)$	$(1 - \widehat{\Delta}_P^2)_+ \xrightarrow{p} 0$	No
Thm. 4	$\mathbb{E}_P[\bar{\Phi}(\widehat{\Delta}_P)]$	Assumption 1	Any $\widehat{\Delta}_P \geqslant 0$	Yes

Table 1: Validity results for the median set $\widehat{\mathrm{CI}}_N^{\mathtt{Med}}$. The fourth column refers to the asymptotic tightness of the bound on miscoverage.

4 Convergence Rates of the Confidence Sets

This section establishes non-asymptotic bounds on the diameter of the confidence sets, defined as $\mathrm{Diam}_{\|\cdot\|}(\cdot)$ in Theorem 1. The deterministic lower bound in Theorem 1 implies that the diameter cannot shrink faster than the estimation rate of $\widehat{\theta}_2 \in \arg\min_{\theta\in\Theta} \widehat{\mathbb{M}}_2(\theta)$. Hence, the bound depends on the complexity of Θ, in contrast to the dimension-agnostic validity established in Section 3. For problems where $\widehat{\theta}_2$ is minimax-optimal, a matching upper bound is established below. For M-estimation (15), $\widehat{\theta}_2$ is known to be sub-optimal in several constrained or high-dimensional settings (Kur et al., 2024; Aolaritei et al., 2025), and the diameter bounds will reflect this sub-optimality. A possible remedy is to modify $\widehat{\mathbb{M}}_2$; the results below are general and do not assume a particular form of $\widehat{\mathbb{M}}_2$.

We first obtain the simple lower bound.

Proposition 2. *Fix $\alpha \leqslant 1/2$. When $\widehat{\mathrm{CI}}_N^{\mathtt{Med}}$ is valid as established in Section 3, then with probability greater than $1-2\alpha$, $\mathrm{Diam}_{\|\cdot\|}(\widehat{\mathrm{CI}}_{N,\alpha}^{\mathtt{CLT}}) \geqslant \mathrm{Diam}_{\|\cdot\|}(\theta(P))$. The same holds for $\widehat{\mathrm{CI}}_{N,\alpha}^{\mathtt{Hull}}$.*

Proof. This follows by applying the union bound on two points in $\theta(P)$. □

Proposition 2 implies that the confidence set will not shrink to a singleton set unless $\theta(P)$ is itself a singleton. Hence, we assume $\theta(P) = \{\theta_0\}$ to establish a diameter bound below.

4.1 Upper bounds for diameters

Assume $\theta(P) = \{\theta_0\}$. The following conditions are imposed at θ_0.

(A1) There exist constants $c_0, \gamma \geqslant 0$ such that $\mathbb{C}_P(\theta) \geqslant c_0\|\theta - \theta_0\|^{1+\gamma}$ for all $\theta \in \Theta$.

(A2) There exists $\eta < 1 + \gamma$ such that

$$\mathbb{E}_P^*\left[\sup_{\|\theta-\theta_0\|<\delta} |(\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta) - (\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta_0)|\right] \leqslant \delta^\eta g(n), \tag{26}$$

where $\mathbb{E}_P^*[\cdot]$ denotes outer expectation.

(A3) The initial estimator $\widehat{\theta}_1$ satisfies $\mathbb{E}_P[|(\widehat{\mathbb{M}}_2-\mathbb{M}_P)(\widehat{\theta}_1)-(\widehat{\mathbb{M}}_2-\mathbb{M}_P)(\theta_0)| \mid \widehat{\theta}_1] \leqslant h(n)$.

(A4) There exists $\nu < 1 + \gamma$ such that

$$\begin{gathered}\mathbb{E}_P^*\left[\sup_{\|\theta-\theta_0\|<\delta} |\widehat{V}_{\theta,\theta_0}^{1/2} - V_{\theta,\theta_0}^{1/2}|\right] \leqslant \delta^\nu g_{\mathtt{var}}(n) \quad \text{and} \\ \mathbb{E}_P[|\widehat{V}_{\widehat{\theta}_1,\theta_0}^{1/2} - V_{\widehat{\theta}_1,\theta_0}^{1/2}| \mid \widehat{\theta}_1] \leqslant h_{\mathtt{var}}(n).\end{gathered} \tag{27}$$

(A1) quantifies curvature and implies that θ_0 is a strong global minimizer of $\mathbb{M}_P$ (Drusvyatskiy and Lewis, 2013). **(A2)** is a maximal inequality (van der Vaart and Wellner, 1996, Section 2.3.1) reflecting the local complexity of Θ; the outer expectation handles possible measurability issues. For M-estimation (15) under independence, this is controlled via symmetrization and Rademacher complexity; under β-mixing, Dehling and Philipp (2002) is a classical reference. **(A3)** quantifies the convergence rate of $\widehat{\theta}_1$. **(A4)** places analogous conditions on the variance estimator and is only needed for the CLT-based set. Randomness of $h(\cdot)$ and $h_{\mathtt{var}}(\cdot)$ is suppressed in notation. **(A1)** and **(A2)** are standard in M-estimation theory (van der Vaart and Wellner, 1996, Theorem 3.2.5) and appear in Kim and Pollard (1990) to distinguish regular from irregular estimators.

Theorem 5. *Assume **(A1)**, **(A2)**, and **(A3)**. For any $\varepsilon > 0$, there exists C_ε depending on ε and c_0 such that*

$$\mathrm{Diam}_{\|\cdot\|}(\widehat{\mathrm{CI}}_N^{\mathtt{Med}}) \leqslant C_\varepsilon\{g(n)^{1/(1+\gamma-\eta)} + h(n)^{2/(1+\gamma)}\} \tag{28}$$

with probability greater than $1-\varepsilon-\beta(r)$. *Under* ***(A4)*** *in addition, there exists* $C_{\varepsilon,\alpha}$ *depending on* ε, α, *and* c_0 *such that*

$$\begin{aligned}&\mathrm{Diam}_{\|\cdot\|}(\widehat{\mathrm{CI}}^{\mathtt{CLT}}_{N,\alpha})\\&\quad\leqslant C_{\varepsilon,\alpha}\{g(n)^{1/(1+\gamma-\eta)}+h(n)^{2/(1+\gamma)}+g_{\mathtt{var}}(n)^{1/(1+\gamma-\nu)}+h_{\mathtt{var}}(n)^{2/(1+\gamma)}\}\end{aligned}\tag{29}$$

with probability greater than $1-\varepsilon-\beta(r)$.

Theorem 5, via Theorem 1, extends to HulC under **(A1)**, **(A2)**, and **(A3)** alone, without **(A4)**. We remark on two features. First, Theorem 5 does not assume independence, specific structure on $\mathbb{M}_P$, or particular choices of $\widehat{\theta}_1$, $\widehat{\mathbb{M}}_2$, $\widehat{V}_{\theta,\theta'}$. Second, the diameter bound depends on γ and η despite the confidence sets being constructed without prior knowledge of them, so the sets *adapt* automatically to the unknown curvature and local complexity of Θ.

Remark 5. *When* $\widehat{V}_{\theta,\theta_0}$ *is a sample variance,* ***(A4)*** *requires a control over squared empirical processes. For uniformly bounded processes, the contraction principle (Ledoux and Talagrand, 2013, Theorem 4.12) transfers the bound in* ***(A2)*** *to* ***(A4)****. For heavy-tailed processes,* ***(A4)*** *typically requires stronger moment conditions. HulC could be considered advantageous relative to the CLT-based set in that it does not require* ***(A4)****.*

Remark 6. *For brevity, the dependence of* g *and* $g_{\mathtt{var}}$ *on* δ *is suppressed in* ***(A2)*** *and* ***(A4)****. Slowly varying dependence on* δ*, including poly-logarithmic factors, is permitted; see Theorems 12 and 13 of KT26.*

Remark 7. *Although* ***(A1)*** *is stated globally, it typically holds only locally near* θ_0*. Through a slightly more involved argument,* ***(A1)*** *is relaxed to a local condition in Section 5.3 of KT26.*

Remark 8. *The dependence on* α *is made explicit in Theorem 13 of KT26, where it enters as a multiplicative factor of* $\log(1/\alpha)$ *on all terms, including the complexity-dependent ones, i.e.,* g *and* $g_{\mathtt{var}}$*. The resulting confidence set is therefore* not *sub-Gaussian in the sense of Lugosi and Mendelson (2019). Constructing confidence sets with sub-Gaussian widths within this framework remains an open problem.*

5 Statistical Applications

This section provides statistical applications illustrating the results through both theory and simulation. Throughout, we focus on high-dimensional or irregular M-estimation (15) with unique minimizer $\theta(P)=\{\theta_0\}$ under IID observations. Theoretical results are stated for $\widehat{\mathrm{CI}}^{\mathtt{Med}}_N$; simulation results are provided for all three confidence sets. We adopt $N=2n$, $r=0$ with an even sample split. All proofs are presented in KT26.

5.1 High-dimensional mean

Setting. Consider IID observations $Z_1,\dots,Z_N\in\mathbb{R}^d$ with mean θ_0 and covariance Σ_P. Let $\mathbb{M}_P(\theta)=\mathbb{E}_P\|Z_i-\theta\|_2^2$ and $\widehat{\mathbb{M}}_2(\theta)=n^{-1}\sum_{i\in I_2}\|Z_i-\theta\|_2^2$. As the initial estimator $\widehat{\theta}_1$, we use the sample mean. A similar procedure was studied in Kim and Ramdas (2024).

Validity and diameter.

Theorem 6. *Let* $Z^\circ = \Sigma_P^{-1/2}(Z-\theta_0)$ *and* $\widetilde{\Delta}_P = \sqrt{n}\|\widehat{\theta}_1-\theta_0\|_2/(2\|\Sigma_P\|_{\mathrm{op}}^{1/2})$. *There exists a universal constant* $C>0$ *such that*

$$\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\mathtt{Med}}) \leqslant \frac{1}{2} + C\min\left\{\frac{1}{(1+\widetilde{\Delta}_P^2)}, \frac{\sup_{u\in\mathbb{S}^{d-1}}\mathbb{E}_P[|\langle Z^\circ,u\rangle|^3]}{\sqrt{n}(1+\widetilde{\Delta}_P^3)}\right\} \tag{30}$$

Validity holds under either of two conditions. First, $\widetilde{\Delta}_P \xrightarrow{p} \infty$, which holds when $\|\widehat{\theta}_1-\theta_0\|_2 \gg \|\Sigma_P\|_{\mathrm{op}}^{1/2}/\sqrt{n}$; under finite variance, this is easily satisfied for $\mathrm{tr}(\Sigma_P) \gg \|\Sigma_P\|_{\mathrm{op}}$. Second, the CLT remainder vanishes when $\mathbb{E}_P[|\langle Z^\circ,u\rangle|^3] < \infty$ for all $u\in\mathbb{S}^{d-1}$, i.e., L_3-L_2 norm equivalence. This condition is widely employed in high-dimensional statistics. This can be relaxed to $L_{2+\delta}$-L_2 norm equivalence for heavy-tailed observations.

Theorem 7. *For any* $n\geqslant 1$, *it holds that*

$$\mathrm{Diam}_{\|\cdot\|_2}(\widehat{\mathrm{CI}}_N^{\mathtt{Med}}) = O_P\left(\sqrt{\frac{\mathrm{tr}(\Sigma_P)}{n}} + \|\widehat{\theta}_1-\theta_0\|_2\right). \tag{31}$$

The rate $\sqrt{\mathrm{tr}(\Sigma_P)/n}$ is minimax-optimal for mean estimation, confirming that the diameter adapts to the intrinsic complexity of the problem.

Remark 9. *Neither Theorem 6 nor Theorem 7 requires* Θ *to be finite-dimensional. The derivations rely on* Σ_P *being a covariance operator on an inner product space. The framework can also be extended beyond inner product spaces, i.e., Fréchet mean problems. The analysis is left to future work.*

Improvement and computation. When the curvature $\mathbb{C}_P(\theta)$ is known or estimable, one can construct tighter *curvature-corrected* sets:

$$\begin{aligned}
\widehat{\mathrm{CI}}_N^{\mathtt{Med,UCB}} &:= \left\{\theta\in\Theta:\ \widehat{\mathbb{M}}_2(\theta)-\widehat{\mathbb{M}}_2(\widehat{\theta}_1) \leqslant -\widehat{\mathbb{C}}_2(\theta,\widehat{\theta}_1)\right\},\\
\widehat{\mathrm{CI}}_{N,\alpha}^{\mathtt{CLT,UCB}} &:= \left\{\theta\in\Theta:\ \widehat{\mathbb{M}}_2(\theta)-\widehat{\mathbb{M}}_2(\widehat{\theta}_1) \leqslant -\widehat{\mathbb{C}}_2(\theta,\widehat{\theta}_1) + z_{1-\alpha}\widehat{V}_{\theta,\widehat{\theta}_1}^{1/2}\right\},
\end{aligned} \tag{32}$$

where $\widehat{\mathbb{C}}_2(\theta,\widehat{\theta}_1)$ is a plug-in estimator of $\mathbb{M}_P(\theta)-\mathbb{M}_P(\widehat{\theta}_1)$ based on D_2, and UCB stands for upper confidence bound. HulC constructed from $\widehat{\mathrm{CI}}_N^{\mathtt{Med,UCB}}$ is denoted $\widehat{\mathrm{CI}}_{N,\alpha}^{\mathtt{Hull,UCB}}$; see Section 6.2 of KT26 for details.

For mean estimation, $\widehat{\mathbb{C}}_2(\theta,\widehat{\theta}_1) = \|\theta-\widehat{\theta}_1\|_2^2$, so no estimation is required. The sets have closed-form geometry and are easy to compute. Let $\widehat{\theta}_2 = \arg\min_{\theta\in\mathbb{R}^d}\widehat{\mathbb{M}}_2(\theta)$, i.e., the sample mean on D_2. The median set is a ball centered at $\widehat{\theta}_2$ with radius $\|\widehat{\theta}_2-\widehat{\theta}_1\|_2$; after curvature correction, it becomes a ball centered at $(\widehat{\theta}_1+\widehat{\theta}_2)/2$ with radius $\|\widehat{\theta}_2-\widehat{\theta}_1\|_2/2$. HulC is the union of B balls with shared $\widehat{\theta}_1$ and $\widehat{\theta}_2^{(j)} = \arg\min_{\theta\in\mathbb{R}^d}\widehat{\mathbb{M}}_2^{(j)}(\theta)$ constructed from each bin. Figure 3 illustrates this for $B=4$. An efficiently computable superset of the CLT-based set is described in Section 7.1 of KT26.

Simulation. We compare the CLT-based set and HulC against the asymptotic Wald interval based on the chi-squared distribution, the bootstrap and the multiplier bootstrap, all applied

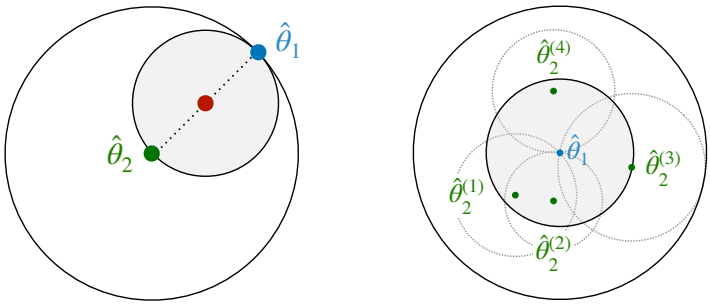


Figure 3: Geometry of the confidence sets for mean estimation. Left: the large sphere is $\widehat{\mathrm{CI}}_N^{\texttt{Med}}$, and the gray sphere is $\widehat{\mathrm{CI}}_N^{\texttt{Med,UCB}}$. Right: HulC is the union of balls (dotted circles); the large sphere is a simple superset of $\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{Hull}}$, and the gray region is the corresponding superset of $\widehat{\mathrm{CI}}_{N,\alpha}^{\texttt{Hull,UCB}}$ when $B = 4$.

to the L_2-norm. With $N = 300$ and $d \in \{2, \ldots, 200\}$, we generate $X_i \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.1^{|i-j|}$. We implement the CLT-based set and HulC at $\alpha = 0.1$ and $\alpha = 0.5$, with and without curvature correction. Coverage and ratios of effective radius (i.e., geometric mean of d radii) relative to the Wald interval are recorded over 500 replications; see S.9.1 of KT26 for further details.

Figure 4a displays the estimated coverage of all confidence sets. The Wald interval undercovers severely as d increases, while all other methods remain valid. Figure 4b displays the estimated effective radius ratio relative to the Wald interval. The curvature-corrected median set achieves an effective radius comparable to the bootstrap methods despite having an exact closed-form expression based on only two sample means (Figure 3). The displayed effective radius for the CLT-based set and HulC are larger than those of the bootstrap methods; this is because the ratio plotted corresponds to computable upper bounds, and the actual confidence sets are smaller but harder to evaluate exactly.

5.2 High-dimensional regression with L_1 penalty

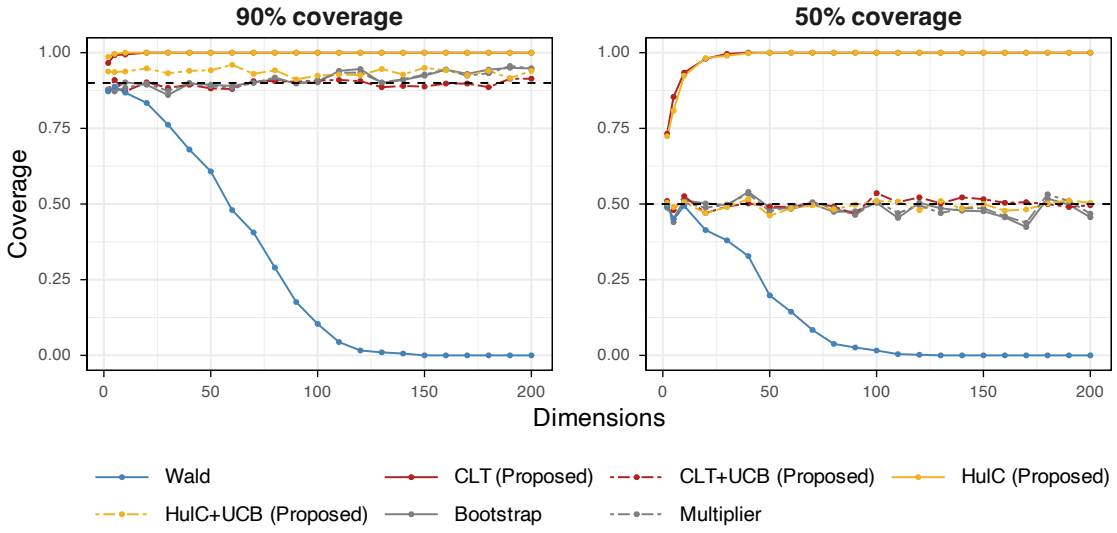
Setting. Consider IID observations $(X_1^\top, Y_1)^\top, \ldots, (X_N^\top, Y_N)^\top \in \mathbb{R}^d \times \mathbb{R}$ with positive definite $\Gamma_P := \mathbb{E}_P[XX^\top]$. For $\lambda \geqslant 0$, consider the optimization objective:

$$\mathbb{M}_P(\theta; \lambda) = \mathbb{E}_P[(Y - \theta^\top X)^2] + \lambda\|\theta\|_1 \quad \text{and} \quad \widehat{\mathbb{M}}_2(\theta; \lambda) = n^{-1} \sum_{i \in I_2} (Y_i - \theta^\top X_i)^2 + \lambda\|\theta\|_1. \quad (33)$$

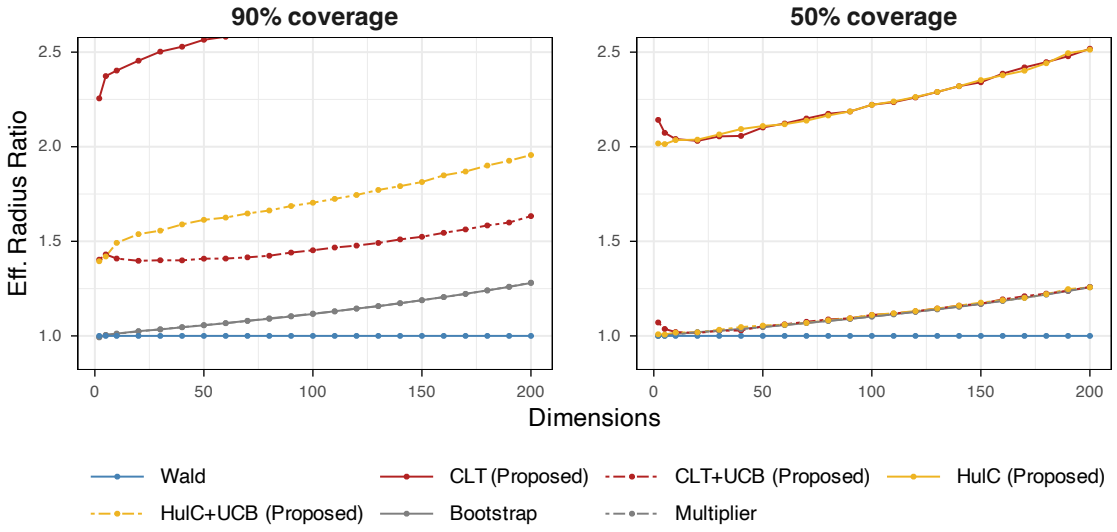
We study inference for $\theta_{0,\lambda} = \arg\min_{\theta \in \mathbb{R}^d} \mathbb{M}_P(\theta; \lambda)$. When $\lambda = 0$, the minimizer $\theta_0 \equiv \theta_{0,0}$ exists even if $\mathbb{E}_P[Y_i|X_i]$ is not linear, but in general $\theta_0 \neq \theta_{0,\lambda}$ for $\lambda > 0$. The median set is constructed from $\widehat{\mathbb{M}}_2(\theta; \lambda)$ for fixed $\lambda \geqslant 0$.

Validity and diameter. Denote $\|U\|_{L^q} = (\mathbb{E}_P \|U\|^q)^{1/q}$.

(B1) There exist $q_x \geqslant 2$ and $L \geqslant 1$ such that $\sup_{u \in \mathbb{S}^{d-1}} \|u^\top \Gamma_P^{-1/2} X\|_{L^{q_x}} \leqslant L$.



(a) Estimated coverage of the proposed sets, the Wald interval, and two bootstrap methods at the 90% and 50% nominal levels, as dimension increases.



(b) Average effective radius ratio of each confidence set relative to the Wald interval, as dimension increases.

Figure 4: Coverage and diameter bounds for high-dimensional mean.

(B1) imposes an L_{q_x}-L_2 norm equivalence on covariates, allowing for heavy-tailed distributions and forgoing sub-Gaussian or sub-exponential assumptions. This condition is common and mild in high-dimensional regression.

Theorem 8. *Assume* ***(B1)*** *with* $q_x = 4$. *Let* $\Sigma_P := \mathrm{Cov}_P(X)$ *and* $\varepsilon_\lambda := Y - \theta_{0,\lambda}^\top X$. *Suppose there exist* $\bar{\sigma}, \underline{\sigma}, \eta > 0$ *such that* $\underline{\sigma}^2 \leqslant \mathbb{E}_P[\varepsilon_\lambda^2] \leqslant \bar{\sigma}^2$ *and* $\eta \leqslant \lambda_{\min}(\Gamma_P^{-1/2} \Sigma_P \Gamma_P^{-1/2})$. *Define*

$$H_P = \mathrm{Cov}_P(X\varepsilon_\lambda) \quad \textit{and} \quad W^\circ = H_P^{-1/2}(X\varepsilon_\lambda - \mathbb{E}_P[X\varepsilon_\lambda]), \tag{34}$$

$$\begin{aligned} R_n &= \inf_{\delta > 0} \left\{ \frac{2L^2\delta}{\underline{\sigma}\sqrt{\eta}} + \mathbb{P}(\|\widehat{\theta}_1 - \theta_{0,\lambda}\|_{\Gamma_P} > \delta) \right\} + \frac{\sup_{u \in \mathbb{S}^{d-1}} \mathbb{E}_P[|\langle W^\circ, u\rangle|^3]}{\sqrt{n}(1 + \widetilde{\Delta}_P^3)} \\ \widetilde{\Delta}_P^2 &= \frac{n\|\widehat{\theta}_1 - \theta_{0,\lambda}\|_{\Gamma_P}^2}{4\bar{\sigma}^2 + 2L^4\|\widehat{\theta}_1 - \theta_{0,\lambda}\|_{\Gamma_P}^2}. \end{aligned} \tag{35}$$

There exists a universal constant $C > 0$ *such that*

$$\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}}) \leqslant \frac{1}{2} + C \min\left\{ \frac{1}{(1 + \widetilde{\Delta}_P^2)}, R_n \right\}. \tag{36}$$

This result holds for any $\lambda \geqslant 0$. As in high-dimensional mean, validity holds under either of two conditions. First, $\widetilde{\Delta}_P \xrightarrow{p} \infty$, which holds when $\|\widehat{\theta}_1 - \theta_{0,\lambda}\|_{\Gamma_P} \gg \bar{\sigma}/\sqrt{n}$ and $n \gg L^4$; for OLS this holds when *both* n and d are large regardless of their relative growth. In particular, validity holds for an inconsistent $\widehat{\theta}_1$. Second, $\|\widehat{\theta}_1 - \theta_{0,\lambda}\|_{\Gamma_P} \xrightarrow{p} 0$ together with $\mathbb{E}_P[|\langle W^\circ, u\rangle|^3] < \infty$ for all $u \in \mathbb{S}^{d-1}$, i.e., L_3-L_2 norm equivalence of the centered score, which can be relaxed to $L_{2+\delta}$-L_2 norm equivalence.

Theorem 9. *Assume* $\lambda = 0$ *and* ***(B1)*** *with* $q_x > 2$. *Suppose there exists a constant* $\bar{\sigma} > 0$ *such that* $\mathbb{E}_P[(Y - \theta_{0,\lambda}^\top X)^2] \leqslant \bar{\sigma}^2$. *Then for* $n \geqslant C_{q_x} d$, *with* C_{q_x} *depending only on* q_x,

$$\mathrm{Diam}_{\|\cdot\|_{\Gamma_P}}\left(\widehat{\mathrm{CI}}_N^{\texttt{Med}}\right) = O_P\left(\sqrt{\frac{\bar{\sigma}^2 d}{n}} + \|\widehat{\theta}_1 - \theta_0\|_{\Gamma_P} \right). \tag{37}$$

The rate $\sqrt{\bar{\sigma}^2 d/n}$ is minimax-optimal (Mourtada, 2022, Theorem 1). Existing confidence sets, such as Chang et al. (2023), establish a similar diameter bound under $d = o(n^{2/3})$ by a one-step bias-corrected method. A dimension-agnostic confidence set with minimax diameter rate under $C_{q_x} d \leqslant n$ appears to be new.

For $\lambda > 0$, we establish a deterministic bound that exhibits adaptivity to sparse submodels. Throughout, $\widehat{\theta}_1$ is treated fixed, that is, all statements are conditional on D_1. We follow the geometric framework of Negahban et al. (2012). For $S \subseteq \{1, \ldots, d\}$, its complement S^c, and $x \in \mathbb{R}^d$, x_S denotes x restricted to coordinates in S. Given $\widehat{\theta}_1$ and S, define the restricted star-shaped set:

$$\mathcal{C}(\widehat{\theta}_1, S) := \left\{ \widehat{\theta}_1 + u \in \mathbb{R}^d \,:\, \|u_{S^c}\|_1 \leqslant 3\|u_S\|_1 + 4\|(\widehat{\theta}_1)_{S^c}\|_1 \right\}. \tag{38}$$

We then introduce the following assumption:

(B2) Fix $\widehat{\theta}_1$ and denote $\widehat{\Gamma}_2 = n^{-1}\sum_{i\in I_2} X_i X_i^\top$. There exist a constant $\kappa_S > 0$ and a function τ such that

$$u^\top \widehat{\Gamma}_2 u \geqslant \kappa_S \|u\|_2^2 - \tau(n,d)\|u\|_1^2 \quad \text{for all} \quad \widehat{\theta}_1 + u \in \mathcal{C}(\widehat{\theta}_1, S). \tag{39}$$

(B2) is a version of the restricted eigenvalue (RE) condition of Loh and Wainwright (2012, Definition 1). For sub-Gaussian X_i, it holds with high probability whenever $n \gtrsim \mathrm{card}(S)\log d$ with $\kappa_S \approx \lambda_{\min}(\mathbb{E}_P[X_S X_S^\top])$. See Lecué and Mendelson (2017) for RE conditions under weak moment assumptions.

Theorem 10. *Fix $\widehat{\theta}_1$. For any λ satisfying $\lambda \geqslant 4n^{-1}\|\sum_{i\in I_2}(Y_i - \widehat{\theta}_1^\top X_i)X_i\|_\infty$,*

$$\mathrm{Diam}_{\|\cdot\|_2}(\widehat{\mathrm{CI}}_N^{\mathtt{Med}}) \leqslant \inf_S \left\{ \frac{3\lambda\sqrt{\mathrm{card}(S)}}{\kappa_S} + \sqrt{\frac{4\lambda\|(\widehat{\theta}_1)_{S^c}\|_1}{\kappa_S}} \right\} + \|\widehat{\theta}_1 - \theta_{0,\lambda}\|_2 \tag{40}$$

*where the infimum is over all $S \subseteq \{1,\dots,d\}$ for which **(B2)** holds.*

When $\theta_{0,\lambda}$ and $\widehat{\theta}_1$ are well-approximated by an s-sparse vector, $(\widehat{\theta}_1)_{S^c}$ is small for $S = \mathrm{supp}(\theta_{0,\lambda})$. Choosing $\lambda \asymp \sqrt{\log d/n}$ and $\widehat{\theta}_1$ as the Lasso, the diameter bound scales like $\sqrt{s\log d/n}$. More generally, whenever $\widehat{\theta}_1$ adapts to an approximately sparse model, the resulting confidence set inherits the same adaptivity.

The condition $\lambda \geqslant 4n^{-1}\|\sum_{i\in I_2}(Y_i - \widehat{\theta}_1^\top X_i)X_i\|_\infty$ must be verified. This is satisfied with high probability if the distributions of $\varepsilon_\lambda X$ and X are sub-Gaussian. We defer a full account on the randomness from $\widehat{\theta}_1$ to future work.

Improvement and computation. The curvature satisfies $\mathbb{C}_P(\theta) \geqslant \|\theta - \theta_{0,\lambda}\|_{\Gamma_P}^2$ with equality at $\lambda = 0$. The curvature-corrected sets (32) can be constructed by letting $\widehat{\mathbb{C}}_2(\theta, \widehat{\theta}_1) = \|\theta - \widehat{\theta}_1\|_{\widehat{\Gamma}_2}^2$. At $\lambda = 0$, the resulting sets have the same closed-form ball geometry as in Figure 3, with radii measured in $\|\cdot\|_{\widehat{\Gamma}_2}$ rather than $\|\cdot\|_2$. All sets and their diameters are therefore efficiently computable.

Simulation. We conduct two sets of experiments.

First, set $\lambda = 0$. The CLT-based set and HulC are compared against the asymptotic Wald interval based on chi-squared, the bootstrap, and the multiplier bootstrap on the L_2-norm. With $N = 300$ and $d \in \{10,\dots,120\}$, we generate $X_i \sim \mathcal{N}(0,\Sigma)$ with $\Sigma_{i,j} = 0.1^{|i-j|}$ and $Y_i = \theta_0^\top X_i + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, |\theta_0^\top X_i| + 0.5)$. The CLT-based set and HulC are implemented at $\alpha = 0.1$ and $\alpha = 0.5$, with and without curvature correction, using OLS as $\widehat{\theta}_1$. Coverage and effective radius (i.e., geometric mean of radii) ratios relative to the Wald interval are recorded over 500 replications; see S.9.2 of KT26 for further details.

Figure 5a displays the estimated coverage of all confidence sets. The Wald interval undercovers severely as d increases. The multiplier bootstrap also undercovers for large d, while Efron's bootstrap is conservative, consistent with empirically reported findings (El Karoui and Purdom, 2018). The CLT-based set and HulC are valid but conservative; with curvature correction, they achieve close to nominal levels. Figure 5b displays the estimated effective radius ratio relative to the Wald interval. The curvature-corrected median set is comparable in effective radius to the bootstrap methods and, notably, smaller than Efron's bootstrap.

Next, we conduct experiments with $\lambda > 0$ and verify adaptivity to sparse submodels. With $N \in \{300, 500, 1000, 2000, 3000, 5000\}$ and $d = 5000$, we generate $X_i \sim \mathcal{N}(0, I_d)$ and $Y_i \sim \mathcal{N}(\theta_0^\top X_i, 1)$ where θ_0 is s-sparse with $s = \lfloor n^\gamma \rfloor$ and $\gamma \in \{0.2, 0.4, 0.6, 0.8\}$. The set $\widehat{\mathrm{CI}}_N^{\texttt{Med,UCB}}$ is implemented with $\lambda = \sqrt{\log d/n}$ and $\widehat{\theta}_1$ as the Lasso at the same λ. The radius in a direction $u \in \mathbb{S}^{d-1}$ is estimated as the largest $h \geqslant 0$ such that $\widehat{\theta}_1 + uh \in \widehat{\mathrm{CI}}_N^{\texttt{Med,UCB}}$; the maximum radius over 100 random directions is recorded. The empirical coverage and median radius over 500 replications are recorded. Figure 6 shows that the confidence set remains valid throughout, though conservative. The right panel shows that the median radius shrinks at the different rates depending on γ, confirming adaptivity to the sparsity level.

5.3 Manski's discrete choice model

Setting. Consider IID observations $(X_1^\top, Y_1)^\top, \dots (X_N^\top, Y_N)^\top \in \mathbb{R}^d \times \{-1, 1\}$ from

$$Y_i := \mathrm{sgn}(\theta_0^\top X_i + \varepsilon_i) \quad \text{where} \quad \mathrm{sgn}(t) = 2\mathbf{1}\{t \geqslant 0\} - 1 \quad \text{and} \quad \mathrm{med}(\varepsilon_i | X_i) = 0. \tag{41}$$

Manski (1975) shows that $\theta_0 = \arg\min_{\theta \in \mathbb{S}^{d-1}} \mathbb{E}_P[-Y \,\mathrm{sgn}(\theta^\top X)]$. We construct the median set by letting $\mathbb{M}_P(\theta) = \mathbb{E}_P[-Y \,\mathrm{sgn}(\theta^\top X)]$ and $\widehat{\mathbb{M}}_2(\theta) = n^{-1} \sum_{i \in I_2} -Y_i \,\mathrm{sgn}(\theta^\top X_i)$.

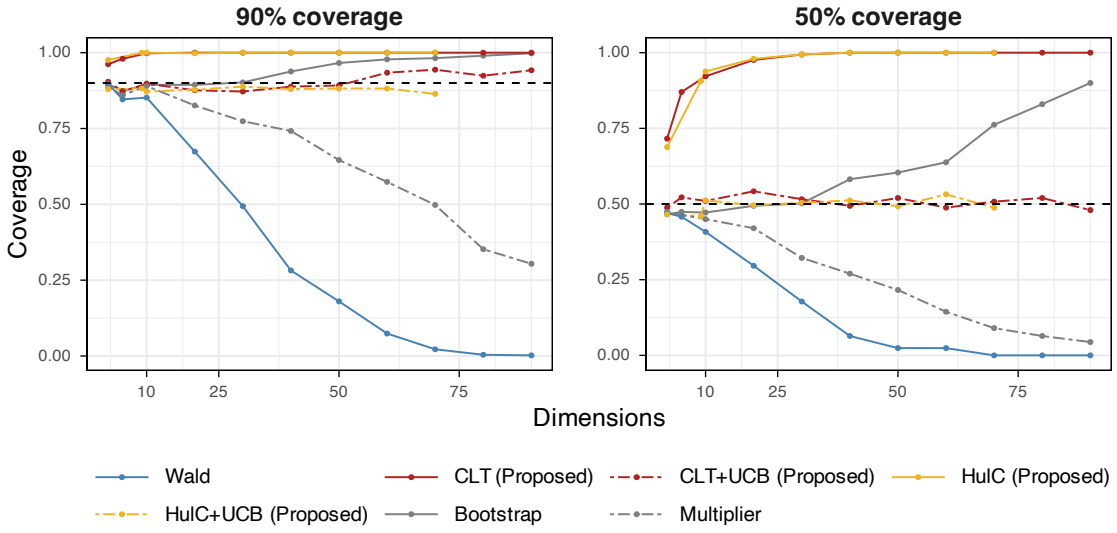
Validity and diameter. The following distributional assumptions are standard (Mukherjee et al., 2021);

- **(B3)** There exists $c_1 > 0$ such that $c_1 \|\theta - \theta_0\|_2 \leqslant \mathbb{P}_P\left(\mathrm{sgn}(\theta^\top X_i) \neq \mathrm{sgn}(\theta_0^\top X_i)\right)$ for all $\theta \in \mathbb{S}^{d-1}$.
- **(B4)** Let $\eta_P(x) := \mathbb{P}_P(Y = 1 \mid X = x)$. There exist $C_0 > 0$, $t^* \in (0, 1/2)$, and $\gamma > 0$, such that $\mathbb{P}_P\left(|\eta_P(X) - 1/2| < t\right) \leqslant C_0 t^{1/\gamma}$ for all $0 < t < t^*$.
- **(B5)** There exist $C_1, C_2 > 0$ such that $\mathbb{P}_P\left(\mathrm{sgn}(\theta^\top X_i) \neq \mathrm{sgn}(\theta_0^\top X_i)\right) \leqslant C_1 \|\theta - \theta_0\|_2$ for all $\theta \in \mathbb{S}^{d-1}$. Additionally, for some $\delta \in (0, (t^*)^{1+1/\gamma})$ and for all $\|\theta - \theta_0\|_2 \leqslant \delta$, $\mathbb{P}_P\left(\mathrm{sgn}(\theta^\top X_i) \neq \mathrm{sgn}(\theta_0^\top X_i)\right) \leqslant C_2 \|\theta - \theta_0\|_2^{1+\gamma}$.

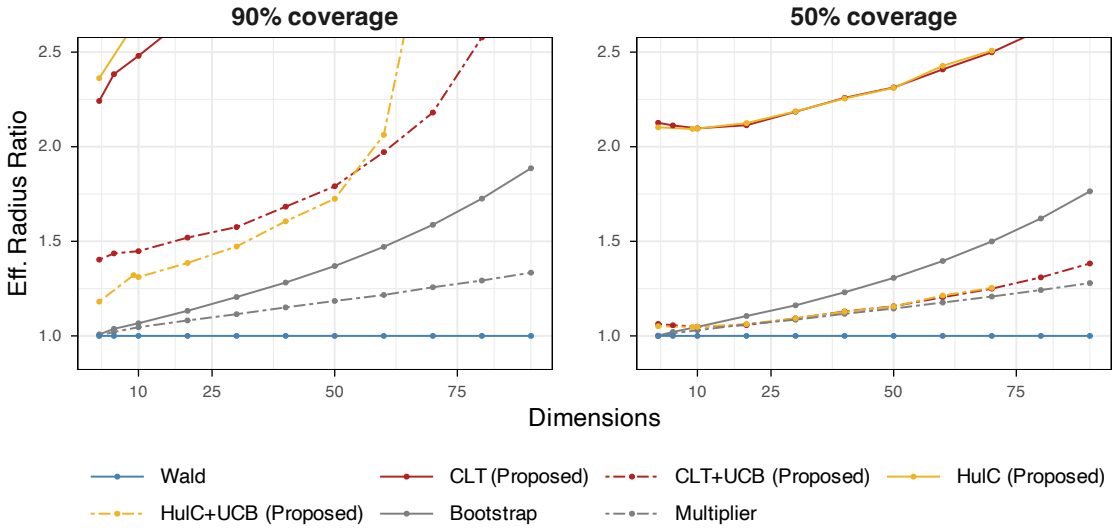
(B3) relates the covariate distribution to the geometry of the parameter space. **(B4)** is the low noise (margin) assumption (Mammen and Tsybakov, 1999), quantifying the deviation of the conditional class probability from 1/2 near the decision boundary; as $\gamma \to 0$, the boundary becomes well-separated, representing the most favorable case for estimation. **(B3)** and **(B4)** together give the lower bound $\mathbb{C}_P(\theta) \gtrsim \|\theta - \theta_0\|^{1+\gamma}$; the additional **(B5)** gives $\mathbb{C}_P(\theta) \asymp \|\theta - \theta_0\|^{1+\gamma}$ locally. Existing inference methods for this problem require stronger distributional assumptions on X and $\eta_P(X)$, e.g., Patra et al. (2018, Assumption C3).

Theorem 11. *Assume **(B3)**. There exists a universal constant $C > 0$ such that*

$$\mathrm{MC}_P(\widehat{\mathrm{CI}}_N^{\texttt{Med}}) \leqslant \frac{1}{2} + \mathbb{E}_P\left[\min\left\{1, \frac{C}{\sqrt{c_1 n \|\widehat{\theta}_1 - \theta_0\|_2}}\right\}\right]. \tag{42}$$



(a) Estimated coverage of the proposed sets, the Wald interval, and two bootstrap methods at the 90% and 50% nominal levels, as dimension increases.



(b) Average effective radius ratio of each confidence set relative to the Wald interval, as dimension increases.

Figure 5: Coverage and diameter bounds for high-dimensional OLS.

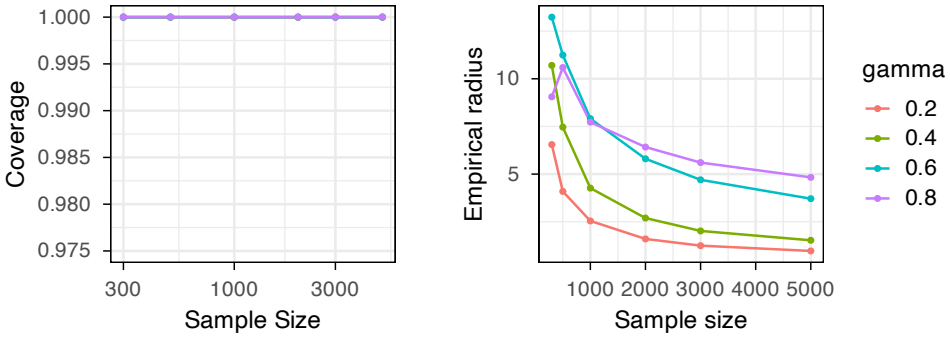


Figure 6: The coverage and estimated radius of the confidence set with varying levels of sparsity $s = n^\gamma$. The results for $\widehat{\mathrm{CI}}_N^{\mathtt{Med,UCB}}$ are displayed. Left: The empirical coverage shows that the confidence set is valid, but conservative. Right: The estimated radius shows that the confidence set indeed adapts to the sparsity of the population target.

Validity requires $n\|\widehat{\theta}_1 - \theta_0\|_2 \xrightarrow{p} \infty$. When $\widehat{\theta}_1$ is a minimax-optimal estimator which converges at rate $(d/n)^{1/(1+2\gamma)}$ (Mukherjee et al., 2019, Theorem 3.1) or slower, the validity holds as n or d grows. Validity also holds when $\widehat{\theta}_1$ is inconsistent, e.g., the smoothed maximum score estimator (Horowitz, 1992) with fixed bandwidth.

Theorem 12. *Assume* ***(B3)***, ***(B4)***, ***(B5)***. *For all n large enough*[1],

$$\mathrm{Diam}_{\|\cdot\|_2}\big(\widehat{\mathrm{CI}}_N^{\mathtt{Med}}\big) = O_P\left(\left(\frac{d\log(n/d)}{n}\right)^{1/(1+2\gamma)} + \|\widehat{\theta}_1 - \theta_0\|_2\right). \tag{43}$$

The rate matches the minimax rate (Mukherjee et al., 2019, Theorem 3.4) (up to a logarithmic factor), confirming rate adaptivity through γ.

Computation. The confidence set has complex geometry due to the non-smooth objective. We estimate the diameter via a geodesic-based sampling algorithm: for a direction $v \in \mathbb{R}^d$ with $v \perp \widehat{\theta}_1$, define a geodesic $\gamma(\phi) = \cos(\phi)\widehat{\theta}_1 + \sin(\phi)v$ through $\widehat{\theta}_1$ and locate the boundary of $\widehat{\mathrm{CI}}_N^{\mathtt{Med}}$ by bisection. Repeating over many directions estimates the diameter; see Figure 7 and Section S.9.3 of KT26 for details.

Simulation. We compare the CLT-based set and HulC against subsampling with estimated rate (Bertail et al., 1999) and the bootstrap (included as a reference despite its known inconsistency (Sen et al., 2010)). Given $N = 200$ and $d \in \{10, \ldots, 90\}$ ($d \leqslant 50$ for subsampling and bootstrap), we generate $X_i \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{i,j} = 0.1^{|i-j|}/d$, and

$$\mathbb{P}(Y_i = 1 \mid X_i) = \Phi(\mathrm{sgn}(\theta_0^\top X_i) \cdot |\theta_0^\top X_i|^{\gamma-1}), \tag{44}$$

which approximately satisfies **(B3)**–**(B5)** with parameter γ. Numerical studies use $\gamma \in \{1/2, 1, 2\}$, corresponding to expected convergence rates, $(d/N)^{1/2}$, $(d/N)^{1/3}$, and $(d/N)^{1/5}$ up to logarithmic factors. We use the smoothed maximum score estimator (Horowitz, 1992) as $\widehat{\theta}_1$

[1] If **(B4)** holds globally, we only require $n \geqslant Cd$ for some constant C.

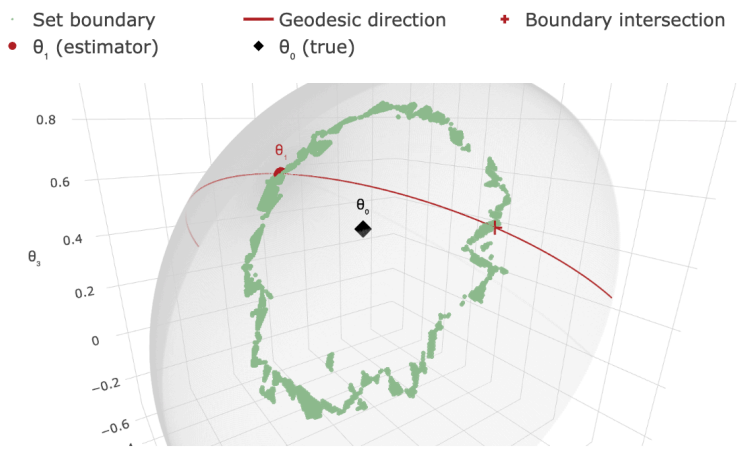


Figure 7: Geodesic-based diameter estimation on $\mathbb{S}^2$. The green region is the boundary of the confidence set; the red line is one geodesic through $\widehat{\theta}_1$ and its intersection with the boundary is found by bisection.

since Manski's maximum score estimator is not computationally feasible in large dimensions. The rest of the details can be found in Section 9 of KT26.

Figure 8a displays the estimated coverage of all confidence sets constructed at the 90% nominal level. Both CLT-based set and HulC remain valid throughout, yet they become more conservative as d grows. On the other hand, the validity of the comparative methods depends significantly on dimension and curvature.

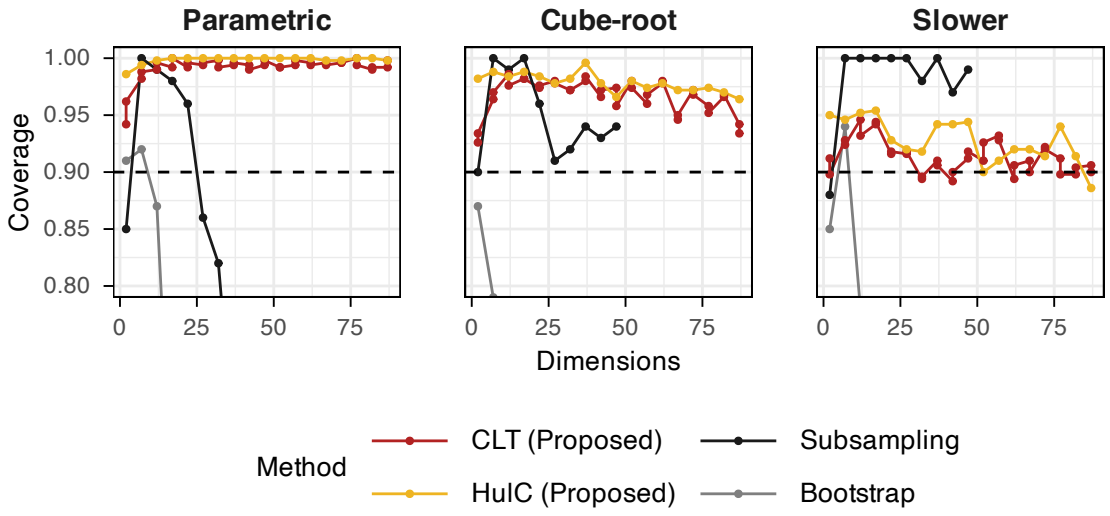
Figure 8b displays the average estimated diameter of each confidence set. The slope of each line corresponds to the exponent in the rate of convergence, with theoretical values $1/2, 1/3$, and $1/5$ for $\gamma = 1/2, 1, 2$ respectively. The slope shown for the CLT-based method is computed based on $d \leqslant 50$ via linear regression. The observed rate of the diameter scales similarly to the theoretical rates.

6 Concluding Remarks

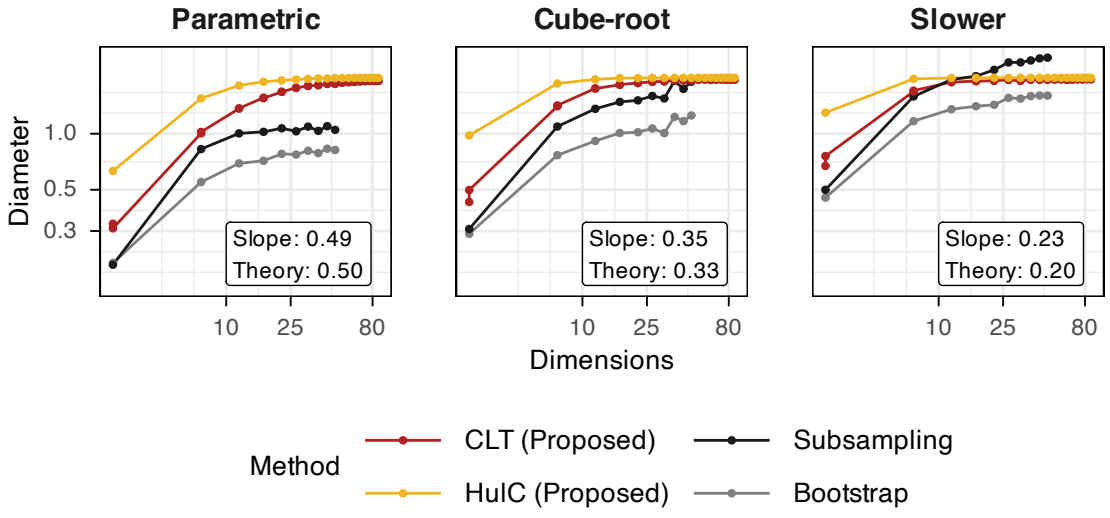
This manuscript studies a general framework for constructing confidence sets for solutions to stochastic optimization, based on risk inversion and sample splitting. A key feature of the framework is dimension-agnostic validity, which remains in force in high-dimensional and irregular settings where standard asymptotic theory breaks down. The framework is analyzed through a unified treatment of validity, conservativeness, and diameter bounds.

The theoretical results are illustrated through three applications: high-dimensional mean estimation, high-dimensional linear regression with and without L_1 penalization, and Manski's discrete choice model. In each case, the proposed sets achieve minimax-optimal or near-optimal diameter rates. Several guarantees appear to be new: dimension-agnostic validity with $\sqrt{d/N}$ diameter rate for misspecified OLS under weak moment conditions; honest inference for population Lasso programs without model correctness; and an adaptive diameter bound for Manski's discrete choice model that reflects the unknown margin parameter.

Several directions remain open. First, *semiparametric problems*: when the target is a



(a) Estimated coverage of the proposed confidence sets and two sampling methods at the 90% nominal level, as a function of dimension.



(b) Average estimated diameter of confidence sets on a log–log scale, with theoretical rate shown. Observed slopes closely match the theoretical rates.

Figure 8: Coverage and diameter bounds for the high-dimensional Manski problem.

functional of a higher-dimensional nuisance parameter estimated in a first stage, it would be of interest to understand whether the framework can recover the elbow between parametric and nonparametric rates. Second, *removing sample splitting*: it remains open whether splitting can be avoided using tools from algorithmic stability or differential privacy, which offer alternative decoupling mechanisms without discarding data. Third, *sub-Gaussian widths*: constructing confidence sets within this framework that achieve sub-Gaussian diameter bounds in the sense of Lugosi and Mendelson (2019) remains open.

7 Selected Proofs

Proof of Theorem 2. The first bound follows from Chebyshev's inequality and Berbee's coupling lemma (Thm 1, KT26). For the second bound, iterative Berbee's coupling lemma gives $\mathrm{MC}_P(\widehat{\mathrm{CI}}^{\texttt{Hull}}_{N,\alpha}\kappa) \leqslant \mathbb{E}_P[p^B_{m,\kappa}] + B\beta(r)$ (Thm 5, KT26). Applying $\ln(1+x) \leqslant x$ yields

$$\ln p^B \leqslant B \ln \frac{1}{2} + 2B\left(p - \frac{1}{2}\right)_+ \quad \text{where} \quad B = \lceil \log_2(1/\alpha) \rceil,$$

and exponentiating gives the result. □

Proof of Theorem 3. The result follows from Cantelli's inequality and Berbee's coupling lemma (Thm 2, KT26). □

Proof of Theorem 4. We provide the result for $\alpha = 1/2$. On D_1 and fixed $\theta_0 \in \theta(P)$,

$$\begin{aligned}\mathbb{P}(\theta_0 \notin \widehat{\mathrm{CI}}^{\texttt{CLT}}_{N,\alpha,r}|D_1) &= \mathbb{P}\left(\widehat{\mathbb{M}}_2(\theta_0) - \widehat{\mathbb{M}}_2(\widehat{\theta}_1) \geqslant 0 | D_1\right) \\ &= \mathbb{P}\left(\frac{(\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\theta_0) - (\widehat{\mathbb{M}}_2 - \mathbb{M}_P)(\widehat{\theta}_1)}{V^{1/2}_{\theta_0,\hat{\theta}_1}} \geqslant \hat{\Delta}_P | D_1\right) \leqslant \bar{\Phi}(\hat{\Delta}_P) + \frac{CK_P(\theta_0,\widehat{\theta}_1)}{\sqrt{n}(1+|\hat{\Delta}_P|^3)}\end{aligned}$$

where the first inequality follows from Assumption 1. The result is obtained by taking expectation over D_1 during which we apply Berbee's coupling lemma at the expense of $\beta(r)$ term. See KT26 for a formal argument. □

Proof of Theorem 5. The bound (28) is a special case of Theorem 12 of KT26. Setting r_n such that

$$r_n^{-2} r_n^{2\eta/(1+\gamma)} g(n) \asymp 1 \implies r_n \asymp \{g(n)\}^{\frac{1+\gamma}{2(1+\gamma-\eta)}},$$

the final rate is given by $r_n^{2/(1+\gamma)} = g(n)^{1/(1+\gamma-\eta)}$. Setting $s_n = h(n)$,

$$\mathbb{P}_P(|(\widehat{\mathbb{M}}_2 - \mathbb{M}_2)(\widehat{\theta}_1) - (\widehat{\mathbb{M}}_2 - \mathbb{M}_2)(\theta_0)| \geqslant \varepsilon^{-1} h(n) \mid \widehat{\theta}_1) \leqslant \varepsilon.$$

This yields the term $h(n)^{2/(1+\gamma)}$ in (28). The bound (29) follows analogously from Theorem 13 of KT26. □

Proof of Theorem 10. Denote $\widehat{\mathcal{L}}_2(\theta) = n^{-1}\sum_{i\in I_2}(Y_i - \theta^\top X_i)^2$ and introduce the short-hand notation $u_S = \Pi_S(u)$ where $\Pi_S(u) := \arg\min_{v\in S}\|u - v\|_2$. Then it follows that

$$\begin{aligned}
\widehat{\mathrm{CI}}_N^{\mathtt{Med}} &= \left\{\theta \in \Theta \,:\, \widehat{\mathcal{L}}_2(\theta) + \lambda\|\theta\|_1 \leqslant \widehat{\mathcal{L}}_2(\widehat{\theta}_1) + \lambda\|\widehat{\theta}_1\|_1\right\} \\
&\subseteq \left\{\widehat{\theta}_1 + u \in \Theta \,:\, -\frac{\lambda}{2}\{\|u_S\|_1 + \|u_{S^c}\|_1\} \leqslant -\lambda\{\|u_{S^c}\|_1 - \|u_S\|_1 - 2\|\Pi_{S^c}(\widehat{\theta}_1)\|_1\}\right\} \\
&= \left\{\widehat{\theta}_1 + u \in \Theta \,:\, \|u_{S^c}\|_1 \leqslant 3\|u_{S_S}\|_1 + 4\|\Pi_{S^c}(\widehat{\theta}_1)\|_1\right\}
\end{aligned}$$

where the inclusion follows from (Negahban et al., 2012, Lemma 3) under **(B2)** and $\lambda \geqslant 4n^{-1}\|\sum_{i\in I_2}(Y_i - \widehat{\theta}_1^\top X_i)X_i\|_\infty$. Denoting $\theta = \widehat{\theta}_1 + u$, for any $\theta \in \widehat{\mathrm{CI}}_N^{\mathtt{Med}}$ it holds that

$$\begin{aligned}
&\widehat{\mathcal{L}}_2(\theta) + \lambda\|\theta\|_1 \leqslant \widehat{\mathcal{L}}_2(\widehat{\theta}_1) + \lambda\|\widehat{\theta}_1\|_1 \\
&\quad\implies \kappa_S\|u\|_2^2 - \tau_S^2(\widehat{\theta}_1) \leqslant 3\lambda\|u_S\|_1 + 4\lambda\|\Pi_{S^c}(\widehat{\theta}_1)\|_1 - \lambda\|u_{S^c}\|_1 \\
&\quad\implies 0 \leqslant -\kappa_S\left(\|u\|_2^2 - \frac{3\lambda\sqrt{s}}{\kappa_S}\|u\|_2 - \frac{4\lambda\|\Pi_{S^c}(\widehat{\theta}_1)\|_1}{\kappa_S}\right)
\end{aligned}$$

where we used $\|u\|_1 \leqslant \sqrt{s}\|u\|_2$ for all $u \in S$. This inequality cannot hold for $\|u\|_2$ large, and for such $\|u\|_2$, $\widehat{\theta}_1 + u$ will not be in the confidence set. Specifically, we can show that $\widehat{\theta}_1 + u \notin \widehat{\mathrm{CI}}_N^{\mathtt{Med}}$ if

$$\|u\|_2 \geqslant \frac{3\lambda\sqrt{s}}{\kappa_S} + \sqrt{\frac{4\lambda\|\Pi_{S^c}(\widehat{\theta}_1)\|}{\kappa_S}}.$$

Hence we deduce that $\widehat{\mathrm{CI}}_N^{\mathtt{Med}}$ is contained in a ball, centered at $\widehat{\theta}_1$ with radius given by the RHS term of the above display. □

References

Aolaritei, L., Jordan, M. I., Pathak, R., and Ulichney, A. (2025). Revisiting mean estimation over L_p balls: Is the MLE optimal? *arXiv preprint arXiv:2506.10354*.

Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *The Annals of Mathematical Statistics*, 27(4):1115–1122.

Bentkus, V., Götze, F., and Tikhomoirov, A. (1997). Berry-Esseen bounds for statistics of weakly dependent samples. *Bernoulli*, 3(3):329–349.

Bentkus, V. and Götze, F. (1996). The Berry-Esseen bound for student's statistic. *The Annals of Probability*, 24(1):491–503.

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.

Bradley, R. C. (2005). Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144.

Cattaneo, M. D., Jansson, M., and Ma, X. (2019). Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies*, 86(3):1095–1122.

Cattaneo, M. D., Jansson, M., and Nagasawa, K. (2020). Bootstrap-based inference for cube root asymptotics. *Econometrica*, 88(5):2203–2219.

Chang, W., Kuchibhotla, A. K., and Rinaldo, A. (2023). Inference for projection parameters in linear regression: beyond $d = o(n^{1/2})$. *arXiv preprint arXiv:2307.00795*.

Chen, L. H. and Shao, Q.-M. (2007). Normal approximation for nonlinear statistics using a concentration inequality approach. *Bernoulli*, 13(2):581–599.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Dehling, H. and Philipp, W. (2002). *Empirical process techniques for dependent data*. Springer.

Drusvyatskiy, D. and Lewis, A. S. (2013). Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential. *SIAM Journal on Optimization*, 23(1):256–267.

El Karoui, N. and Purdom, E. (2018). Can we trust the bootstrap in high-dimensions? the case of linear models. *Journal of Machine Learning Research*, 19(5):1–66.

Fan, X. and Shao, Q.-M. (2018). Berry–Esseen bounds for self-normalized martingales. *Communications in Mathematics and Statistics*, 6(1):13–27.

Fryzlewicz, P. and Rao, S. S. (2011). Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli*, 17(1):320–346.

Haeusler, E. and Joos, K. (1988). A nonuniform bound on the rate of convergence in the martingale central limit theorem. *The Annals of Probability*, 16(4):1699–1720.

Hafouta, Y. (2022). Non-uniform berry-esseen theorem and Edgeworth expansions with applications to transport distances for weakly dependent random variables. *arXiv preprint arXiv:2210.07204*.

Hoeffding, W. and Robbins, H. (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal*, 15(3):773–780.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–531.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 5(1):221–233.

Katz, M. L. (1963). Note on the berry-esseen theorem. *The Annals of Mathematical Statistics*, 34(3):1107–1108.

Kim, I. and Ramdas, A. (2024). Dimension-agnostic inference using cross U-statistics. *Bernoulli*, 30(1):683–711.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219.

Knight, K. (1998). Limiting distributions for L_1 regression estimators under general conditions. *The Annals of Statistics*, 26(2):755–770.

Kuchibhotla, A. K., Balakrishnan, S., and Wasserman, L. (2024). The HulC: confidence regions from convex hulls. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(3):586–622.

Kur, G., Gao, F., Guntuboyina, A., and Sen, B. (2024). Convex regression in multidimensions: Suboptimality of least squares estimators. *The Annals of Statistics*, 52(6):2791–2815.

Lecué, G. and Mendelson, S. (2017). Sparse recovery under weak moment assumptions. *Journal of the European Mathematical Society*, 19(3):881–904.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces*. Springer.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664.

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, 21(1):255–285.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of econometrics*, 3(3):205–228.

Mourtada, J. (2022). Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178.

Mukherjee, D., Banerjee, M., and Ritov, Y. (2019). Non-standard asymptotics in high dimensions: Manski's maximum score estimator revisited. *arXiv preprint arXiv:1903.10063*.

Mukherjee, D., Banerjee, M., and Ritov, Y. (2021). Optimal linear discriminators for the discrete choice model in growing dimensions. *The Annals of Statistics*, 49(6):3324–3357.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557.

Patra, R. K., Seijo, E., and Sen, B. (2018). A consistent bootstrap procedure for the maximum score estimator. *Journal of Econometrics*, 205(2):488–507.

Pötscher, B. M. (2002). Lower risk bounds and properties of confidence sets for ill-posed estimation problems with applications to spectral density and persistence estimation, unit roots, and estimation of long memory parameters. *Econometrica*, 70(3):1035–1065.

Rio, E. (2017). *Asymptotic theory of weakly dependent random processes*, volume 80. Springer.

Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253.

Sen, B., Banerjee, M., and Woodroofe, M. (2010). Inconsistency of bootstrap: the grenander estimator. *The Annals of Statistics*, 38(4):1953–1977.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Takatsu, K. and Kuchibhotla, A. K. (2026). Honest inference for stochastic optimization. *arXiv preprint arXiv:2501.07772*.

Tikhomirov, A. N. (1981). On the convergence rate in the central limit theorem for weakly dependent random variables. *Theory of Probability & Its Applications*, 25(4):790–809.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van der Vaart, A. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Vogel, S. (2008). Universal confidence sets for solutions of optimization problems. *SIAM Journal on Optimization*, 19(3):1467–1488.

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