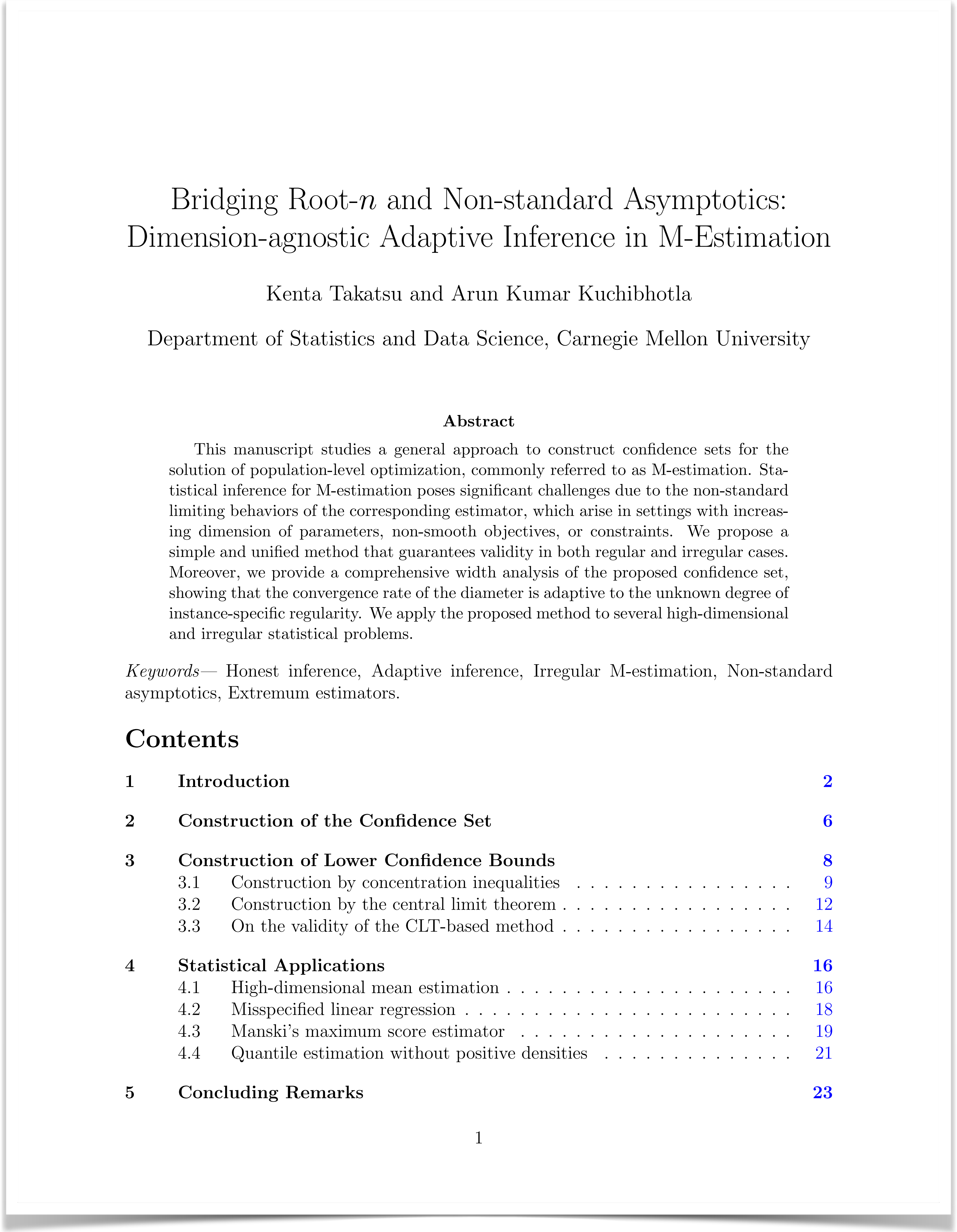
Dimension-agnostic inference for M-estimation

With Arun Kumar Kuchibhotla

Kenta Takatsu
Carnegie Mellon University


Bridging Root-n and Non-standard Asymptotics: Dimension-agnostic Adaptive Inference in M-Estimation

Kenta Takatsu and Arun Kumar Kuchibhotla

Department of Statistics and Data Science, Carnegie Mellon University

Abstract

This manuscript studies a general approach to construct confidence sets for the solution of population-level optimization, commonly referred to as M-estimation. Statistical inference for M-estimation poses significant challenges due to the non-standard limiting behaviors of the corresponding estimator, which arise in settings with increasing dimension of parameters, non-smooth objectives, or constraints. We propose a simple and unified method that guarantees validity in both regular and irregular cases. Moreover, we provide a comprehensive width analysis of the proposed confidence set, showing that the convergence rate of the diameter is adaptive to the unknown degree of instance-specific regularity. We apply the proposed method to several high-dimensional and irregular statistical problems.

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

Contents

1. Motivation and Settings

Inference on Median

Observe IID real-valued R.V.s $X_1, \ldots, X_n$ and estimate $\mu = \mathrm{med}(X)$.

Use the sample median as an estimator $\bar{\mu}_n = X_{(\lceil n/2 \rceil)}$.

Inference on Median

Observe IID real-valued R.V.s $X_1, \ldots, X_n$ and estimate $\mu = \mathrm{med}(X)$.

Use the sample median as an estimator $\bar{\mu}_n = X_{(\lceil n/2 \rceil)}$.

How do we construct a confidence interval?

Inference on Median

Observe IID real-valued R.V.s $X_1, \ldots, X_n$ and estimate $\mu = \text{med}(X)$.

Use the sample median as an estimator $\bar{\mu}_n = X_{(\lceil n/2 \rceil)}$.

How do we construct a confidence interval?

1. $n^{1/2}(\bar{\mu}_n - \mu) \xrightarrow{d} N\left(0, 1/(4f^2(\mu))\right)$ where $f(\mu) > 0$ is density at median.

Inference on Median

Observe IID real-valued R.V.s $X_1, \ldots, X_n$ and estimate $\mu = \text{med}(X)$.

Use the sample median as an estimator $\bar{\mu}_n = X_{(\lceil n/2 \rceil)}$.

How do we construct a confidence interval?

1. $n^{1/2}(\bar{\mu}_n - \mu) \xrightarrow{d} N\left(0, 1/(4f^2(\mu))\right)$ where $f(\mu) > 0$ is density at median.

2. $n^{1/(2\beta)}(\bar{\mu}_n - \mu) \xrightarrow{d} W(\beta)$ where $W(\beta)$ is not Gaussian (Smirnov, 1952) and β is the Hölder smoothness of CDF at μ (When $\beta = 1$, it reduces to the first case).

Inference on Median

Observe IID real-valued R.V.s $X_1, \ldots, X_n$ and estimate $\mu = \mathrm{med}(X)$.

Use the sample median as an estimator $\bar{\mu}_n = X_{(\lceil n/2 \rceil)}$.

How do we construct a confidence interval?

1. $n^{1/2}(\bar{\mu}_n - \mu) \xrightarrow{d} N\left(0, 1/(4f^2(\mu))\right)$ where $f(\mu) > 0$ is density at median.

2. $n^{1/(2\beta)}(\bar{\mu}_n - \mu) \xrightarrow{d} W(\beta)$ where $W(\beta)$ is not Gaussian (Smirnov, 1952) and β is the Hölder smoothness of CDF at μ (When $\beta = 1$, it reduces to the first case).

Inference based on asymptotic normality requires $f(\mu) > 0$; otherwise, we need to know the precise value of β.

Many estimators display similar irregular behaviors where “limiting” distributions are non-standard, complicated or unknown (without additional assumptions).

ex) Median / quantile.

ex) Shape-restricted estimators and classification (Kim and Pollard, 1990).

ex) High-dimensional problems.

Many estimators display similar irregular behaviors where "limiting" distributions are non-standard, complicated or unknown (without additional assumptions).

ex) Median / quantile.

ex) Shape-restricted estimators and classification (Kim and Pollard, 1990).

ex) High-dimensional problems.

Statistical inference for these irregular problems are known to be challenging.

Asymptotic normality does *not* hold. Bootstrap is usually inconsistent.

Typically, you have to change the methods between regular and irregular cases.

Many estimators display similar irregular behaviors where “limiting” distributions are non-standard, complicated or unknown (without additional assumptions).

ex) Median / quantile.

ex) Shape-restricted estimators and classification (Kim and Pollard, 1990).

ex) High-dimensional problems.

Statistical inference for these irregular problems are known to be challenging.

Asymptotic normality does *not* hold. Bootstrap is usually inconsistent.

Typically, you have to change the methods between regular and irregular cases.

Q. Can we come up with a single confidence set that remains valid for for both regular and irregular settings?

Observe $X_1, \ldots, X_n \in \mathcal{X}$ from a distribution $P \in \mathcal{P}$. Let Θ be a space of parameters.

$$\theta_P = \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}[m(X; \theta)] \text{ and } \widehat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m(X_i; \theta).$$

ex) Regression, maximum likelihood, cross-validation.

ex) Quantile estimation, shape-restricted estimation, classification.

ex) High-dimensional problems where the "dimension" d of Θ grows with n.

Observe $X_1, \ldots, X_n \in \mathscr{X}$ from a distribution $P \in \mathscr{P}$. Let Θ be a space of parameters.

$$\theta_P = \underset{\theta \in \Theta}{\operatorname{argmin}}\, \mathbb{E}[m(X;\theta)] \text{ and } \widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n} m(X_i;\theta).$$

ex) Regression, maximum likelihood, cross-validation.

ex) Quantile estimation, shape-restricted estimation, classification.

ex) High-dimensional problems where the “dimension” d of Θ grows with n.

The goal is to construct a confidence set $\mathrm{CI}_{n,\alpha}$ for θ_P with the following properties:

1. Uniform validity over $\mathscr{P}$, including both regular and irregular cases, i.e., $\inf_{P \in \mathscr{P}} \mathbb{P}(\theta_P \in \mathrm{CI}_{n,\alpha}) \geq 1 - \alpha$
2. Optimal and adaptive convergence rates of the diameter of $\mathrm{CI}_{n,\alpha}$.

2. Proposed Procedure

Observe $X_1, \ldots, X_{2n} \in \mathcal{X}$ from a distribution $P \in \mathcal{P}$. Let Θ be a space of parameters.

The parameter of interest is $\theta_P := \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}[m(X;\theta)]$.

1. Split data $\{X_i\}_{i=1}^{2n}$ into two sets D_1 and D_2. Using D_1, construct *any* estimator $\widehat{\theta}$.
2. For each $\theta \in \Theta$ and $X_i \in D_2$, define $\xi_i := m(X_i;\theta) - m(X_i;\widehat{\theta})$.
3. Let $\bar{\xi}$ and $\widehat{\sigma}^2$ be the sample mean and variance of ξ_i computed on D_2.
4. The confidence set is defined as $\mathrm{CI}_{n,\alpha} := \{\bar{\xi} \le n^{-1/2} z_{1-\alpha} \widehat{\sigma} \ : \theta \in \Theta\}$ where $z_{1-\alpha}$ is $1-\alpha$ quantile of the standard Normal.

Observe $(X_1, Y_1), \ldots, (X_{2n}, Y_{2n}) \in \mathbb{R}^d \times \mathbb{R}$ from a distribution $P \in \mathscr{P}$. Let Θ be a space of parameters.

Define $\theta_P := \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}[\{Y - \theta(X)\}^2]$.

1. Split data $\{(X_i, Y_i)\}_{i=1}^{2n}$ into two sets D_1 and D_2. Using D_1, construct *any* estimator $\widehat{\theta}$.
2. For each $\theta \in \Theta$ and $(X_i, Y_i) \in D_2$, define $\xi_i := \{Y_i - \theta(X_i)\}^2 - \{Y_i - \widehat{\theta}(X_i)\}^2$.
3. Let $\bar{\xi}$ and $\widehat{\sigma}^2$ be the sample mean and variance of ξ_i based on D_2.
4. The confidence set is defined as $\mathrm{CI}_{n,\alpha} := \{\bar{\xi} \le n^{-1/2} z_{1-\alpha} \widehat{\sigma} \ : \theta \in \Theta\}$.

Observe $(X_1, Y_1), \ldots, (X_{2n}, Y_{2n}) \in \mathbb{R}^d \times \mathbb{R}$ from a distribution $P \in \mathscr{P}$. Let Θ be a space of parameters.

Define $\theta_P := \underset{\theta \in \Theta}{\operatorname{argmin}} \, \mathbb{E}[\{Y - \theta(X)\}^2 + \lambda(\theta)]$.

1. Split data $\{(X_i, Y_i)\}_{i=1}^{2n}$ into two sets D_1 and D_2. Using D_1, construct *any* estimator $\widehat{\theta}$.
2. For each $\theta \in \Theta$ and $(X_i, Y_i) \in D_2$, define $\xi_i := \{Y_i - \theta(X_i)\}^2 - \{Y_i - \widehat{\theta}(X_i)\}^2$.
3. Let $\bar{\xi}$ and $\widehat{\sigma}^2$ be the sample mean and variance of ξ_i based on D_2.
4. The confidence set is defined as $\mathrm{CI}_{n,\alpha} := \{\bar{\xi} + \lambda(\theta) - \lambda(\widehat{\theta}) \le n^{-1/2} z_{1-\alpha} \widehat{\sigma} \ : \theta \in \Theta\}$.

Why does this work?

Observe that θ_P is a minimizer and $\mathbb{E}[m(X;\theta_P)] - \mathbb{E}[m(X;\widehat{\theta})] \leq 0$ for any $\widehat{\theta} \in \Theta$.

Why does this work?

Observe that θ_P is a minimizer and $\mathbb{E}[m(X;\theta_P)] - \mathbb{E}[m(X;\hat{\theta})] \leq 0$ for any $\hat{\theta} \in \Theta$.

A non-actionable but valid confidence set is $\left\{ \mathbb{E}[m(X;\theta) - m(X;\hat{\theta})] \leq 0 \ : \ \theta \in \Theta \right\}$.

Why does this work?

Observe that θ_P is a minimizer and $\mathbb{E}[m(X;\theta_P)] - \mathbb{E}[m(X;\hat{\theta})] \leq 0$ for any $\hat{\theta} \in \Theta$.

A non-actionable but valid confidence set is $\left\{\mathbb{E}[m(X;\theta) - m(X;\hat{\theta})] \leq 0 \ : \ \theta \in \Theta\right\}$.

The proposed confidence set is $\left\{n^{-1}\sum [m(X_i;\theta) - m(X_i;\hat{\theta})] \leq \gamma_{n,\alpha} \ : \ \theta \in \Theta\right\}$ where $\gamma_{n,\alpha} \to 0$ is an appropriate cutoff to guarantee validity. From earlier, we define $\xi_i := m(X_i;\theta) - m(X_i;\hat{\theta})$, and we can use the central limit theorem (CLT) for the t-statistics of $\{\xi_i\}$ to obtain $\gamma_{n,\alpha}$.

Why does this work?

Observe that θ_P is a minimizer and $\mathbb{E}[m(X;\theta_P)] - \mathbb{E}[m(X;\hat{\theta})] \leq 0$ for any $\hat{\theta} \in \Theta$.

A non-actionable but valid confidence set is $\left\{ \mathbb{E}[m(X;\theta) - m(X;\hat{\theta})] \leq 0 \ : \ \theta \in \Theta \right\}$.

The proposed confidence set is $\left\{ n^{-1} \sum [m(X_i;\theta) - m(X_i;\hat{\theta})] \leq \gamma_{n,\alpha} \ : \ \theta \in \Theta \right\}$ where $\gamma_{n,\alpha} \to 0$ is an appropriate cutoff to guarantee validity. From earlier, we define $\xi_i := m(X_i;\theta) - m(X_i;\hat{\theta})$, and we can use the central limit theorem (CLT) for the t-statistics of $\{\xi_i\}$ to obtain $\gamma_{n,\alpha}$.

This approach is agnostic to the dimension of Θ and the choice of the estimator $\hat{\theta}$.

Why does this work?

Observe that θ_P is a minimizer and $\mathbb{E}[m(X;\theta_P)] - \mathbb{E}[m(X;\widehat{\theta})] \le 0$ for any $\widehat{\theta} \in \Theta$.

A non-actionable but valid confidence set is $\left\{\mathbb{E}[m(X;\theta) - m(X;\widehat{\theta})] \le 0 \ : \ \theta \in \Theta\right\}$.

The proposed confidence set is $\left\{n^{-1}\sum[m(X_i;\theta) - m(X_i;\widehat{\theta})] \le \gamma_{n,\alpha} \ : \ \theta \in \Theta\right\}$ where $\gamma_{n,\alpha} \to 0$ is an appropriate cutoff to guarantee validity. From earlier, we define $\xi_i := m(X_i;\theta) - m(X_i;\widehat{\theta})$, and we can use the central limit theorem (CLT) for the t-statistics of $\{\xi_i\}$to obtain $\gamma_{n,\alpha}$.

This approach is agnostic to the dimension of Θ and the choice of the estimator $\widehat{\theta}$.

This idea is not new (Vogel, 2008), but we use sample-splitting to estimate $\widehat{\theta}$.

3. Theoretical Properties

Uniform Validity

Recall $\xi_i := m(X_i; \theta) - m(X_i; \widehat{\theta})$. Let $\bar{\xi}$ and $\widehat{\sigma}^2$ be the sample mean and variance of ξ_i.
The CI is defined as $\mathrm{CI}_{n,\alpha} := \{\bar{\xi} \leq n^{-1/2} z_{1-\alpha} \widehat{\sigma} \ : \theta \in \Theta\}$.

Uniform Validity

Recall $\xi_i := m(X_i; \theta) - m(X_i; \widehat{\theta})$. Let $\bar{\xi}$ and $\widehat{\sigma}^2$ be the sample mean and variance of ξ_i.
The CI is defined as $\mathrm{CI}_{n,\alpha} := \{\bar{\xi} \leq n^{-1/2} z_{1-\alpha} \widehat{\sigma} \ : \theta \in \Theta\}$.

Define $\xi_{P,i} := m(X_i; \theta_P) - m(X_i; \widehat{\theta})$ and similarly for $\bar{\xi}_P$ and $\widehat{\sigma}^2_P$.

Uniform Validity

Recall $\xi_i := m(X_i; \theta) - m(X_i; \widehat{\theta})$. Let $\bar{\xi}$ and $\widehat{\sigma}^2$ be the sample mean and variance of ξ_i.
The CI is defined as $\mathrm{CI}_{n,\alpha} := \{\bar{\xi} \leq n^{-1/2} z_{1-\alpha} \widehat{\sigma} \ : \theta \in \Theta\}$.

Define $\xi_{P,i} := m(X_i; \theta_P) - m(X_i; \widehat{\theta})$ and similarly for $\bar{\xi}_P$ and $\widehat{\sigma}_P^2$.

Denote the Kolmogorov-Smirnov distance by $\Delta_{n,P} := \sup_{t \in \mathbb{R}} \left| \mathbb{P}(n^{1/2}(\bar{\xi}_P - \mathbb{E}[\bar{\xi}_P]) / \widehat{\sigma}_P \leq t) - \Phi(t) \right|$
where $\Phi(t)$ is the CDF of the standard Normal.

Uniform Validity

Recall $\xi_i := m(X_i; \theta) - m(X_i; \widehat{\theta})$. Let $\bar{\xi}$ and $\widehat{\sigma}^2$ be the sample mean and variance of ξ_i.
The CI is defined as $\mathrm{CI}_{n,\alpha} := \{\bar{\xi} \le n^{-1/2} z_{1-\alpha} \widehat{\sigma} \ : \theta \in \Theta\}$.

Define $\xi_{P,i} := m(X_i; \theta_P) - m(X_i; \widehat{\theta})$ and similarly for $\bar{\xi}_P$ and $\widehat{\sigma}_P^2$.

Denote the Kolmogorov-Smirnov distance by $\Delta_{n,P} := \sup_{t \in \mathbb{R}} \left| \mathbb{P}(n^{1/2}(\bar{\xi}_P - \mathbb{E}[\bar{\xi}_P])/\widehat{\sigma}_P \le t) - \Phi(t) \right|$
where $\Phi(t)$ is the CDF of the standard Normal.

Theorem 4

For any $n \ge 1$, it holds $\inf_{P \in \mathscr{P}} \mathbb{P}(\theta_P \in \mathrm{CI}_{n,\alpha}) \ge 1 - \alpha - \sup_{P \in \mathscr{P}} \Delta_{n,P}$.

We show conditions under which $\Delta_{n,P} \to 0$ as $n \to \infty$ uniformly over $\mathscr{P}$. This holds under mild assumptions on P, including the cases traditionally considered "irregular".

Convergence Rates

Let $\mathrm{Diam}_{\|\cdot\|}(A) := \sup\{\|a-b\| \; : \; a, b \in A\}$.

Theorem 8 (informal)

For all $\theta \in \Theta$,

(1) Curvature condition: $\mathbb{E}[m(X;\theta) - m(X;\theta_P)] \geq c_1\|\theta - \theta_P\|^{1+\beta}$ for some $\beta \geq 0$.

(2) Variance condition: $\mathrm{Var}[m(X;\theta) - m(X;\theta_P)] \leq c_2\|\theta - \theta_P\|^{2\eta}$ for some $\eta < 1 + \beta$.

Then, we have $\mathrm{Diam}_{\|\cdot\|}(\mathrm{CI}_{n,\alpha}) = O_P(n^{-1/(2+2\beta-2\eta)} + r_n^{1/(1+\beta)} + s_n^{1/(1+\beta)})$.

Convergence Rates

Let $\mathrm{Diam}_{\|\cdot\|}(A) := \sup\{\|a-b\| \,:\, a,b \in A\}$.

Theorem 8 (informal)

For all $\theta \in \Theta$,

(1) Curvature condition: $\mathbb{E}[m(X;\theta) - m(X;\theta_P)] \geq c_1\|\theta - \theta_P\|^{1+\beta}$ for some $\beta \geq 0$.

(2) Variance condition: $\mathrm{Var}[m(X;\theta) - m(X;\theta_P)] \leq c_2\|\theta - \theta_P\|^{2\eta}$ for some $\eta < 1+\beta$.

Then, we have $\mathrm{Diam}_{\|\cdot\|}(\mathrm{CI}_{n,\alpha}) = O_P(n^{-1/(2+2\beta-2\eta)} + r_n^{1/(1+\beta)} + s_n^{1/(1+\beta)})$.

r_n depends on the complexity of Θ and the moments of the local envelope $\sup_{\|\theta-\theta_P\|<\delta} |m(X;\theta) - m(X;\theta_P)|$

Convergence Rates

Let $\mathrm{Diam}_{\|\cdot\|}(A) := \sup\{\|a-b\| \ : \ a, b \in A\}$.

Theorem 8 (informal)

For all $\theta \in \Theta$,

(1) Curvature condition: $\mathbb{E}[m(X;\theta) - m(X;\theta_P)] \geq c_1\|\theta - \theta_P\|^{1+\beta}$ for some $\beta \geq 0$.

(2) Variance condition: $\mathrm{Var}[m(X;\theta) - m(X;\theta_P)] \leq c_2\|\theta - \theta_P\|^{2\eta}$ for some $\eta < 1+\beta$.

Then, we have $\mathrm{Diam}_{\|\cdot\|}(\mathrm{CI}_{n,\alpha}) = O_P(n^{-1/(2+2\beta-2\eta)} + r_n^{1/(1+\beta)} + s_n^{1/(1+\beta)})$.

s_n depends on the quality of the initial estimator $\widehat{\theta}$.

Convergence Rates

Let $\mathrm{Diam}_{\|\cdot\|}(A) := \sup\{\|a - b\| \ : \ a, b \in A\}$.

Theorem 8 (informal)

For all $\theta \in \Theta$,

(1) Curvature condition: $\mathbb{E}[m(X;\theta) - m(X;\theta_P)] \geq c_1\|\theta - \theta_P\|^{1+\beta}$ for some $\beta \geq 0$.

(2) Variance condition: $\mathrm{Var}[m(X;\theta) - m(X;\theta_P)] \leq c_2\|\theta - \theta_P\|^{2\eta}$ for some $\eta < 1 + \beta$.

Then, we have $\mathrm{Diam}_{\|\cdot\|}(\mathrm{CI}_{n,\alpha}) = O_P(n^{-1/(2+2\beta-2\eta)} + r_n^{1/(1+\beta)} + s_n^{1/(1+\beta)})$.

Regular (smooth) problems correspond to $\eta = \beta = 1$, leading to $O_P(n^{-1/2})$.

So-called "cube-root" problems correspond to $\eta = 1/2$ and $\beta = 1$, leading to $O_P(n^{-1/3})$.

Median estimation corresponds to $\eta = 1$, leading to $O_P(n^{-1/(2\beta)})$.

4. Statistical Applications

Assumption-lean Regression

Observe $(X_1, Y_1)\ldots, (X_{2n}, Y_{2n}) \in \mathbb{R}^d \times \mathbb{R}$ from a distribution $P \in \mathscr{P}$:

$$Y_i = \theta_P^\top X_i + \varepsilon_i \text{ where } \mathbb{E}[\varepsilon_i X_i] = 0 \text{ and } \mathbb{E}[\varepsilon_i^2 \,|\, X_i] = \sigma_i^2.$$

The target parameter is $\theta_P := \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}[(Y - \theta^\top X)^2]$.

- θ_P exists without making the linearity assumption for $\mathbb{E}[Y_i \,|\, X_i]$.
- We allow $d \to \infty$ as $n \to \infty$

We define the confidence set as $\mathrm{CI}_{n,\alpha} := \{\bar{\xi} \leq n^{-1/2} z_{1-\alpha} \widehat{\sigma} \;:\; \theta \in \mathbb{R}^d\}$

where $\xi_i := \{Y_i - \theta^\top X_i\}^2 - \{Y_i - \widehat{\theta}^\top X_i\}^2$.

Assumption-lean Regression

Denote the gram matrix $\Gamma_P := \mathbb{E}[XX^\top]$.

Theorem 11 (informal)

Assume

(A1) $0 < \underline{\lambda} \leq \lambda_{\min}(\Gamma_P), \lambda_{\max}(\Gamma_P) \leq \bar{\lambda} < \infty$.

(A2) $0 < \underline{\sigma} \leq \sigma_i \leq \bar{\sigma} < \infty$ for all $1 \leq i \leq n$.

(A3) There exists $L \geq 1$ such that $\left(\mathbb{E}[|t^\top X|^4]\right)^{1/4} \leq L\left(\mathbb{E}[|t^\top X|^2]\right)^{1/2}$ for all $t \in \mathbb{S}^{d-1}$.

(A4) The CLT holds for the R.V.s $\langle t, \epsilon_i X_i \rangle$ (e.g., the Lindeberg condition).

(A5) The initial estimator $\widehat{\theta}$ is uniformly consistent for all $P \in \mathscr{P}$.

Then, $\liminf_{n\to\infty} \inf_{P\in\mathscr{P}} \mathbb{P}(\theta_P \in \mathrm{CI}_{n,\alpha}) \geq 1 - \alpha$.

Assumption-lean Regression

Denote the gram matrix $\Gamma_P := \mathbb{E}[XX^\top]$.

Theorem 12 (informal)

Assuming (A1)-(A3) and assume $\widehat{\theta}$ is the ordinary least squares (OLS).

When $d \log^2(d) \leq n$,

$$\mathrm{Diam}_{\|\cdot\|}(\mathrm{CI}_{n,\alpha}) = O_P\left(\sqrt{\overline{\sigma}^2 \mathrm{tr}(\Gamma_P)/n}\right)$$

Assumption-lean Regression

Denote the gram matrix $\Gamma_P := \mathbb{E}[XX^\top]$.

Theorem 12 (informal)

Assuming (A1)-(A3) and assume $\hat{\theta}$ is the ordinary least squares (OLS).

When $d \log^2(d) \leq n$,

$$\mathrm{Diam}_{\|\cdot\|}(\mathrm{CI}_{n,\alpha}) = O_P\left(\sqrt{\bar{\sigma}^2 \mathrm{tr}(\Gamma_P)/n}\right)$$

This result is minimax-optimal (Mourtada, 2022).

5. Summary

We introduced a general framework to perform statistical inference on the solution of optimization problems (M-estimation) based on sample-splitting.

We introduced a general framework to perform statistical inference on the solution of optimization problems (M-estimation) based on sample-splitting.

Despite irregular nature of the problem, we provide the confidence set that remains uniformly valid for both regular and irregular cases.

We introduced a general framework to perform statistical inference on the solution of optimization problems (M-estimation) based on sample-splitting.

Despite irregular nature of the problem, we provide the confidence set that remains uniformly valid for both regular and irregular cases.

The validity is agnostic to the dimension of the problem and the choice of the initial estimator.

We introduced a general framework to perform statistical inference on the solution of optimization problems (M-estimation) based on sample-splitting.

Despite irregular nature of the problem, we provide the confidence set that remains uniformly valid for both regular and irregular cases.

The validity is agnostic to the dimension of the problem and the choice of the initial estimator.

The width of the confidence set converges adaptively to the local geometric features of optimization (curvature and variance).

This talk focused on the CLT-based approach but the manuscript also develops an approach based on concentration inequalities.

This talk focused on the CLT-based approach but the manuscript also develops an approach based on concentration inequalities.

More examples in the manuscript (Manski’s estimator / quantile estimation) demonstrate the validity and optimality of the proposed method in some challenging inference problems.

Thank You

Smirnov, N. V. (1952). Limit distributions for the terms of a variational series. *American Mathematical Society Translations, 6*.

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

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

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.*