Randomness-free Study of M-estimators

NBK Inequalities

Arun Kumar Kuchibhotla

The Wharton School,
University of Pennsylvania.

03 July, 2019

Outline

Introduction

Let's Remember Cramér

- Suppose $Z_1, \ldots, Z_n$ are observations and we consider estimtor $\hat{\theta}$ that satisfies

$$\sum_{i=1}^{n} \psi(Z_i, \hat{\theta}_n) = 0.$$

- MLE, OLS, GLMs and many more estimators are all obtained this way.
- The classical proof of Cramér (1946) proves the Bahadur representation:

$$\sqrt{n}(\hat{\theta} - \theta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbb{E}[\dot{\psi}(Z_1, \theta)])^{-1} \psi(Z_i, \theta) + o_p(1),$$

  under some conditions including $Z_1, \ldots, Z_n$ are iid and smoothness of ψ.
- The proof is based on Taylor series expansion (a deterministic tool):

$$0 = \sum_{i=1}^{n} \psi(Z_i, \hat{\theta}_n) \;\approx\; \sum_{i=1}^{n} \psi(Z_i, \theta) + \sum_{i=1}^{n} \dot{\psi}(Z_i, \theta)(\hat{\theta} - \theta).$$

Let's Remember Cramér

- Suppose $Z_1, \ldots, Z_n$ are observations and we consider estimtor $\hat{\theta}$ that satisfies

$$\sum_{i=1}^{n} \psi(Z_i, \hat{\theta}_n) = 0.$$

- MLE, OLS, GLMs and many more estimators are all obtained this way.
- The classical proof of Cramér (1946) proves the Bahadur representation:

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbb{E}[\dot{\psi}(Z_1, \theta)])^{-1} \psi(Z_i, \theta) + o_p(1),$$

under some conditions including $Z_1, \ldots, Z_n$ are iid and smoothness of ψ.

- The proof is based on Taylor series expansion (a deterministic tool):

$$0 = \sum_{i=1}^{n} \psi(Z_i, \hat{\theta}_n) \approx \sum_{i=1}^{n} \psi(Z_i, \theta) + \sum_{i=1}^{n} \dot{\psi}(Z_i, \theta)(\hat{\theta} - \theta).$$

Do we need Z_i independent or even random? What is θ?

Importance of Bahadur Representation

- Bahadur representation is more important than asymptotic normality.
- It implies asymptotic normality of estimators and Bahadur representation is one of the most popular ways of proving asymptotic normality.
- Bahadur representation is closed under smooth transformations and under addition: (This does *not* hold for asym. normality in general)
  - If $\hat{\theta}_1, \ldots, \hat{\theta}_d$ satisfy the representation, then for any smooth function $f(\cdot, \cdot, \ldots, \cdot)$, we have

    $$\sqrt{n}(f(\hat{\theta}_1, \ldots, \hat{\theta}_d) - f(\theta_1, \ldots, \theta_d)) = n^{-1/2} \sum_{i=1}^{n} \psi_f(Z_i) + o_p(1),$$

    for some function $\psi_f(\cdot)$.
  - If $\hat{\theta}_1, \hat{\theta}_2$ satisfy the representation with $\texttt{Inf}_1$ and $\texttt{Inf}_2$ as influence functions, then

    $$\sqrt{n}(\alpha_1\hat{\theta}_1 + \alpha_2\hat{\theta}_2 - \alpha_1\theta_1 - \alpha_2\theta_2) = n^{-1/2} \sum_{i=1}^{n} [\alpha_1 \texttt{Inf}_1(Z_i) + \alpha_2 \texttt{Inf}_2(Z_i)] + o_p(1).$$

- It is also important for validity of bootstrap/resampling procedures.

Importance of Bahadur Representation

- Bahadur representation is more important than asymptotic normality.
- It implies asymptotic normality of estimators and Bahadur representation is one of the most popular ways of proving asymptotic normality.
- Bahadur representation is closed under smooth transformations and under addition: (This does *not* hold for asym. normality in general)
  - If $\hat{\theta}_1, \ldots, \hat{\theta}_d$ satisfy the representation, then for any smooth function $f(\cdot, \cdot, \ldots, \cdot)$, we have

$$\sqrt{n}(f(\hat{\theta}_1, \ldots, \hat{\theta}_d) - f(\theta_1, \ldots, \theta_d)) = n^{-1/2} \sum_{i=1}^{n} \psi_f(Z_i) + o_p(1),$$

    for some function $\psi_f(\cdot)$.
  - If $\hat{\theta}_1, \hat{\theta}_2$ satisfy the representation with $\mathtt{Inf}_1$ and $\mathtt{Inf}_2$ as influence functions, then

$$\sqrt{n}(\alpha_1\hat{\theta}_1 + \alpha_2\hat{\theta}_2 - \alpha_1\theta_1 - \alpha_2\theta_2) = n^{-1/2} \sum_{i=1}^{n} [\alpha_1 \mathtt{Inf}_1(Z_i) + \alpha_2 \mathtt{Inf}_2(Z_i)] + o_p(1).$$

- It is also important for validity of bootstrap/resampling procedures.

Bahadur Representation $\Rightarrow$ Inference

NBK Inequalities: Linear Regression[1]

[1] K. (2018), Deterministic Inequalities for Smooth M-estimators. arXiv:1809.05172
Thanks to Mateo Wirth, Bikram Karmakar.

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.
- Following Cramér's proof, we get for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \beta), \quad \text{where} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.
- Following Cramér's proof, we get for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \beta), \quad \text{where} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

- This holds for any set of observations (with $\hat{\Sigma}$ invertible).

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i (Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.
- Following Cramér's proof, we get for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \beta), \quad \text{where} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

- This holds for any set of observations (with $\hat{\Sigma}$ invertible).
- Requires neither independence nor a (true linear) model.

Start with Linear Regression

- Consider regression data $Z_i := (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}, 1 \le i \le n$ and the OLS estimator

$$\hat{\beta} := \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1}^{n} X_i(Y_i - X_i^\top \hat{\beta}) = 0.$$

- Here $\psi(Z_i, \theta) = X_i(Y_i - X_i^\top \theta)$, linear in θ. Hence Taylor series is exact.
- Following Cramér's proof, we get for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \beta), \quad \text{where} \quad \hat{\Sigma} := \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top.$$

- If Z_i satisfy a version of LLN: $\hat{\Sigma} \approx \Sigma$ for some Σ, then for any $\beta \in \mathbb{R}^d$,

$$\sqrt{n}\,(\hat{\beta} - \beta) \;=\; (1 + o_p(1)) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Sigma^{-1} X_i (Y_i - X_i^\top \beta),$$

Note: Error is multiplicative not additive!!

Formal Result for OLS

For any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_{op}.$$

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- Inequality is a deterministic version of Bahadur representation.

Formal Result for OLS

For any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_{op}.$$

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- Inequality is a deterministic version of Bahadur representation.
- In some cases (e.g., subsampling/cross-validation) the flexibility of choosing arbitrary Σ, β comes in handy. Also note: $\mathcal{D}^{\Sigma} \approx 0$ is same as $\hat{\Sigma} \approx \Sigma$.

Formal Result for OLS

For any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_{op}.$$

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- Inequality is a deterministic version of Bahadur representation.
- In some cases (e.g., subsampling/cross-validation) the flexibility of choosing arbitrary Σ, β comes in handy. Also note: $\mathcal{D}^{\Sigma} \approx 0$ is same as $\hat{\Sigma} \approx \Sigma$.
- Requires no model assumptions, no randomness assumptions, no assumptions on d/n, no independence/dependence assumptions.

Formal Result for OLS

For any $\Sigma \in \mathbb{R}^{d\times d}$, set

$$\mathcal{D}^{\Sigma} := \|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_p\|_{op}.$$

Theorem (Inequality for OLS Estimator)

For any set of observations $Z_i = (X_i, Y_i)$, any $\Sigma \in \mathbb{R}^{d\times d}$ and any $\beta \in \mathbb{R}^d$, we have

$$\left\|\hat{\beta} - \beta - \frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma} \le \frac{\mathcal{D}^{\Sigma}}{(1-\mathcal{D}^{\Sigma})_+}\left\|\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}X_i(Y_i - X_i^{\top}\beta)\right\|_{\Sigma}.$$

- Inequality is a deterministic version of Bahadur representation.
- In some cases (e.g., subsampling/cross-validation) the flexibility of choosing arbitrary Σ, β comes in handy. Also note: $\mathcal{D}^{\Sigma} \approx 0$ is same as $\hat{\Sigma} \approx \Sigma$.
- Requires no model assumptions, no randomness assumptions, no assumptions on d/n, no independence/dependence assumptions.
- Implies optimal rates, finite sample tail bounds, Berry–Esseen bounds for $\hat{\beta}$.

Application 1: Leave-one-out Cross-Validation

Application 1: Leave-one-out Cross-Validation (LOOCV)

- The deterministic inequality can be readily used for simplifying LOOCV.
- For each $1 \le j \le n$, define

$$\hat{\beta}_{-j} := \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \sum_{i=1, i \neq j}^{n} (Y_i - X_i^\top \theta)^2 \quad \Leftrightarrow \quad \sum_{i=1, i \neq j}^{n} X_i (Y_i - X_i^\top \hat{\beta}_{-j}) = 0.$$

- In this case, it is intuitively clear that $\hat{\beta}_{-j}$ is close to $\hat{\beta}$.
- Note that $\hat{\Sigma}_{-j} \approx \hat{\Sigma}$ for any j, where $\hat{\Sigma}_{-j} = (n-1)^{-1} \sum_{i=1, i \neq j}^{n} X_i X_i^\top$.

Corollary (Deterministic Approximation of LOOCV)

If $n \ge 2$, then simultaneously, for all $1 \le j \le n$, we have

$$\left\| \hat{\beta}_{-j} - \hat{\beta} - \frac{\hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \hat{\beta})}{n} \right\|_{\hat{\Sigma}} \le \frac{2\mathfrak{D}/n}{(1 - 2\mathfrak{D}/n)_+} \left\| \frac{\hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \hat{\beta})}{n} \right\|_{\hat{\Sigma}},$$

where $\mathfrak{D} := 1 + \max_{1 \le j \le n} \|\hat{\Sigma}^{-1/2} X_j\|$. *(Hence* $\hat{\beta}_{-j} \approx \hat{\beta} + n^{-1} \hat{\Sigma}^{-1} X_i (Y_i - X_i^\top \hat{\beta})$.*)*

Application 2: Transformations of Response

Application 2: Transformations of Response

- In modeling, it is sometimes of interest to transform the response to match the assumptions like Gaussianity or homoscedasticity.
- Finding a "good" transformation involves some data snooping. Once again the inequality can be used to get a result for final estimator.
- Suppose $\mathcal{G}$ is a class of transformations under consideration and for each $g \in \mathcal{G}$, we have the OLS estimator

$$\hat{\beta}_g := \mathrm{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^n (g(Y_i) - X_i^\top \theta)^2.$$

For any $g \in \mathcal{G}$, define $\mathtt{Inf}_g(\theta) := n^{-1} \sum_{i=1}^n \Sigma^{-1} X_i (g(Y_i) - X_i^\top \theta)$.

Corollary (Bahadur Representation with Transformed Response)

For any set of observations $Z_i = (X_i, Y_i)$, any Σ, any $g \in \mathcal{G}$ and any $\beta_g \in \mathbb{R}^d$,

$$\left\| \hat{\beta}_g - \beta_g - \mathtt{Inf}_g(\beta_g) \right\|_\Sigma \le \frac{\mathcal{D}^\Sigma}{(1 - \mathcal{D}^\Sigma)_+} \| \mathtt{Inf}_g(\beta_g) \|_\Sigma.$$

In particular this holds for any random $\hat{g} \in \mathcal{G}$ chosen based on the data.

Application 3: Variable Selection

Application 3: Variable Selection

- More often than not, the set of covariates in a reported model is not the same as the set of covariates the analyst started with.
- Finding a "good" set of covariates involves some data snooping.
- Suppose $\mathcal{M}$ is a collection of models (set of covariates) and for each $M \in \mathcal{M}$, we have the OLS estimator

$$\hat{\beta}_M := \text{argmin}_{\theta \in \mathbb{R}^{|M|}} \sum_{i=1}^n (Y_i - X_{i,M}^\top \theta)^2.$$

Set for any $M \in \mathcal{M}$, $\texttt{Inf}_M(\theta) := n^{-1} \sum_{i=1}^n \Sigma_M^{-1} X_{i,M}(Y_i - X_{i,M}^\top \theta)$.

Corollary (Bahadur Representation with Variable Selection)

For any $M \in \mathcal{M}$, any Σ_M, and any $\beta_M \in \mathbb{R}^{|M|}$, we have

$$\left\| \hat{\beta}_M - \beta_M - \mathit{Inf}_M(\beta_M) \right\|_{\Sigma_M} \le \frac{\mathcal{D}_M^\Sigma}{(1 - \mathcal{D}_M^\Sigma)_+} \| \mathit{Inf}_M(\beta_M) \|_{\Sigma_M},$$

where $\mathcal{D}_M^\Sigma := \| \Sigma_M^{-1/2} \hat{\Sigma}_M \Sigma_M^{-1/2} - I_{|M|} \|_{op}$. In particular M can be random chosen based on the data.

NBK Inequalities: Smooth M-estimation[2]

[2]K. (2018), Deterministic Inequalities for Smooth M-estimators. arXiv:1809.05172
Thanks to Mateo Wirth, Bikram Karmakar.

Semi-local Convergence: **N**ewton-**K**antorovich Theorem

Consider a function $g(\cdot)$. Define $B(w^0, \eta; A) := \{w : \|w - w^0\|_A \leq \eta\}$.

Semi-local Convergence: **N**ewton-**K**antorovich Theorem

Consider a function $g(\cdot)$. Define $B(w^0, \eta; A) := \{w : \|w - w^0\|_A \leq \eta\}$.

If there exists $w^0 \in \mathbb{R}^q$ and $L > 0$ such that

$$\left\| \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} \ddot{g}(w) \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} - I_q \right\|_{op} \leq L\|w - w^0\|_{\ddot{g}(w^0)},$$

whenever $\|w - w^0\|_{\ddot{g}(w^0)} \leq (3L)^{-1}$, (ratio-type continuity condition) and

Semi-local Convergence: Newton-Kantorovich Theorem

Consider a function $g(\cdot)$. Define $B(w^0, \eta; A) := \{w : \|w - w^0\|_A \leq \eta\}$.

If there exists $w^0 \in \mathbb{R}^q$ and $L > 0$ such that

$$\left\| \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} \ddot{g}(w) \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} - I_q \right\|_{op} \leq L\|w - w^0\|_{\ddot{g}(w^0)},$$

whenever $\|w - w^0\|_{\ddot{g}(w^0)} \leq (3L)^{-1}$, (ratio-type continuity condition) and

$$\left\| \left[\ddot{g}(w^0)\right]^{-1} \dot{g}(w^0) \right\|_{\ddot{g}(w^0)} \leq \frac{2}{9L} \quad \text{("Close" to zero gradient at } w^0\text{).}$$

Then

Semi-local Convergence: **N**ewton-**K**antorovich Theorem

Consider a function $g(\cdot)$. Define $B(w^0, \eta; A) := \{w : \|w - w^0\|_A \le \eta\}$.

If there exists $w^0 \in \mathbb{R}^q$ and $L > 0$ such that

$$\left\| \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} \ddot{g}(w) \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} - I_q \right\|_{op} \le L\|w - w^0\|_{\ddot{g}(w^0)},$$

whenever $\|w - w^0\|_{\ddot{g}(w^0)} \le (3L)^{-1}$, (ratio-type Lipschitz condition) and

$$\left\| \left[\ddot{g}(w^0)\right]^{-1} \dot{g}(w^0) \right\|_{\ddot{g}(w^0)} \le \frac{2}{9L}, \quad \text{(“Close” to zero gradient at } w^0\text{).}$$

Then

Semi-local Convergence: Newton-Kantorovich Theorem

Consider a function $g(\cdot)$. Define $B(w^0, \eta; A) := \{w : \|w - w^0\|_A \le \eta\}$.

If there exists $w^0 \in \mathbb{R}^q$ and $L > 0$ such that

$$\left\| \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} \ddot{g}(w) \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} - I_q \right\|_{op} \le L\|w - w^0\|_{\ddot{g}(w^0)},$$

whenever $\|w - w^0\|_{\ddot{g}(w^0)} \le (3L)^{-1}$, (ratio-type Lipschitz condition) and

$$\left\| \left[\ddot{g}(w^0)\right]^{-1} \dot{g}(w^0) \right\|_{\ddot{g}(w^0)} \le \frac{2}{9L},$$ ("Close" to zero gradient at w^0).

Then $\exists$ a unique $w^\star \in B(w^0, r; \ddot{g}(w^0)) \ni \dot{g}(w^\star) = 0$ and

$$\left\| w^\star - \underbrace{\left[w^0 - \left(\ddot{g}(w^0)\right)^{-1} \dot{g}(w^0)\right]}_{\text{First Newton Iterate}} \right\|_{\ddot{g}(w^0)} \le \frac{9L}{4} \left\| \left[\ddot{g}(w^0)\right]^{-1} \dot{g}(w^0) \right\|_{\ddot{g}(w^0)}^2 .$$

Quadratic Convergence of Newton's Algorithm.

Semi-local Convergence: Newton-Kantorovich Theorem

Consider a function $g(\cdot)$. Define $B(w^0, \eta; A) := \{w : \|w - w^0\|_A \le \eta\}$.

If there exists $w^0 \in \mathbb{R}^q$ and $L > 0$ such that

$$\left\| \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} \ddot{g}(w) \left[\ddot{g}(w^0)\right]^{-\frac{1}{2}} - I_q \right\|_{op} \le L \|w - w^0\|_{\ddot{g}(w^0)},$$

whenever $\|w - w^0\|_{\ddot{g}(w^0)} \le (3L)^{-1}$, (ratio-type Lipschitz condition) and

$$\left\| \left[\ddot{g}(w^0)\right]^{-1} \dot{g}(w^0) \right\|_{\ddot{g}(w^0)} \le \frac{2}{9L},$$ ("Close" to zero gradient at w^0).

Then $\exists$ a unique $w^\star \in B(w^0, r; \ddot{g}(w^0)) \ni \dot{g}(w^\star) = 0$ and

$$\Bigg\| \underbrace{(w^\star - w^0)}_{\text{Estimation Err.}} + \underbrace{\left(\ddot{g}(w^0)\right)^{-1} \dot{g}(w^0)}_{\text{Influence function}} \Bigg\|_{\ddot{g}(w^0)} \le \frac{9L}{4} \Bigg\| \underbrace{\left[\ddot{g}(w^0)\right]^{-1} \dot{g}(w^0)}_{\text{Influence function}} \Bigg\|^2_{\ddot{g}(w^0)}.$$

Finite Sample bnd Bahadur Representation of M-estimator.

What have we gained?

- No randomness assumptions on the data; result is deterministic.

What have we gained?

- No randomness assumptions on the data; result is deterministic.
- No independence assumptions on observations.

What have we gained?

- No randomness assumptions on the data; result is deterministic.
- No independence assumptions on observations.
- No model assumptions. Allows study under misspecification.

What have we gained?

- No randomness assumptions on the data; result is deterministic.
- No independence assumptions on observations.
- No model assumptions. Allows study under misspecification.
- No asymptotics; everything holds at any finite sample size.

What have we gained?

- No randomness assumptions on the data; result is deterministic.
- No independence assumptions on observations.
- No model assumptions. Allows study under misspecification.
- No asymptotics; everything holds at any finite sample size.
- Bounds are in terms of $\ddot{g}$ and $\dot{g}$ that are averages if $g(\theta) = \sum_{i=1}^{n} \ell(Z_i, \theta)$. **Averages studied for more than a century under various settings.**

What have we gained?

- No randomness assumptions on the data; result is deterministic.
- No independence assumptions on observations.
- No model assumptions. Allows study under misspecification.
- No asymptotics; everything holds at any finite sample size.
- Bounds are in terms of $\ddot{g}$ and $\dot{g}$ that are averages if $g(\theta) = \sum_{i=1}^{n} \ell(Z_i, \theta)$. **Averages studied for more than a century under various settings.**
- Disadvantage: Requires smoothness on the function.

What have we gained?

- No randomness assumptions on the data; result is deterministic.
- No independence assumptions on observations.
- No model assumptions. Allows study under misspecification.
- No asymptotics; everything holds at any finite sample size.
- Bounds are in terms of $\ddot{g}$ and $\dot{g}$ that are averages if $g(\theta) = \sum_{i=1}^{n} \ell(Z_i, \theta)$. **Averages studied for more than a century under various settings.**
- Disadvantage: Requires smoothness on the function.

Under whatever dependence,

LLN for $\ddot{g}(w^0)$ and CLT for $\dot{g}(w^0)$ $\quad \Rightarrow \quad$ CLT for $w^\star - w^0$.

Application: Logistic/Poisson Regression

Application: Logistic/Poisson Regression

- For either $\psi(u) = \log(1 + \exp(u))$, Logistic or $\psi(u) = \exp(u)$ Poisson, let

$$\hat{\beta} := \text{argmin}_{\theta \in \mathbb{R}^d} \; L_n(\theta), \quad \text{where} \quad L_n(\theta) := \sum_{i=1}^{n} \left[\psi(X_i^\top \theta) - Y_i X_i^\top \theta\right],$$

- Define for any $\theta \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$, $\mathcal{D}^{\Sigma}(\theta) := \|\Sigma^{-1/2}\ddot{L}_n(\theta)\Sigma^{-1/2} - I_d\|_{op}$.

Theorem

For any $\beta \in \mathbb{R}^d$ and any $\Sigma \in \mathbb{R}^{d \times d}$, if

$$\max_{1 \le i \le n} \|\Sigma^{-1/2}X_i\| \times \|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma} \le 0.19(1 - \mathcal{D}^{\Sigma}(\beta))_+,$$

then

$$\frac{\|\hat{\beta}_n - \beta + \Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}}{\|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}} \le \frac{\mathcal{D}^{\Sigma}(\beta)}{(1 - \mathcal{D}^{\Sigma}(\beta))_+} + \frac{10 \max_i \|\Sigma^{-1/2}X_i\| \|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}}{(1 - \mathcal{D}^{\Sigma}(\beta))_+^2}.$$

Proves "CLT" if $\dim(X_i) = o(\sqrt{n})$.

Summary and Conclusions

Some Comments

- Deterministic inequalities as above proving Bahadur representation are what we call NBK (**N**ewton-**B**ahadur-**K**antarovich) inequalities.

Some Comments

- Deterministic inequalities as above proving Bahadur representation are what we call NBK (**N**ewton-**B**ahadur-**K**antarovich) inequalities.
- Following the result for logistic and Poisson regression, applications like cross-validation, transformations, variable selection as done for linear regression can be carried out easily.

Some Comments

- Deterministic inequalities as above proving Bahadur representation are what we call NBK (**N**ewton-**B**ahadur-**K**antarovich) inequalities.
- Following the result for logistic and Poisson regression, applications like cross-validation, transformations, variable selection as done for linear regression can be carried out easily.
- The additional assumption above comes from non-linearity of the estimating function which also leads to an additional term in the remainder.

Some Comments

- Deterministic inequalities as above proving Bahadur representation are what we call NBK (**N**ewton-**B**ahadur-**K**antarovich) inequalities.
- Following the result for logistic and Poisson regression, applications like cross-validation, transformations, variable selection as done for linear regression can be carried out easily.
- The additional assumption above comes from non-linearity of the estimating function which also leads to an additional term in the remainder.
- Newton-Kantarovich theorem was developed to study convergence of Newton iterates and it implies Bahadur representation.

Some Comments

- Deterministic inequalities as above proving Bahadur representation are what we call NBK (**N**ewton-**B**ahadur-**K**antarovich) inequalities.
- Following the result for logistic and Poisson regression, applications like cross-validation, transformations, variable selection as done for linear regression can be carried out easily.
- The additional assumption above comes from non-linearity of the estimating function which also leads to an additional term in the remainder.
- Newton-Kantarovich theorem was developed to study convergence of Newton iterates and it implies Bahadur representation.
- This thinking leads to some new first order expansion results for penalized/regularized estimators in high-dimensions.

Some Comments

- Deterministic inequalities as above proving Bahadur representation are what we call NBK (**N**ewton-**B**ahadur-**K**antarovich) inequalities.
- Following the result for logistic and Poisson regression, applications like cross-validation, transformations, variable selection as done for linear regression can be carried out easily.
- The additional assumption above comes from non-linearity of the estimating function which also leads to an additional term in the remainder.
- Newton-Kantarovich theorem was developed to study convergence of Newton iterates and it implies Bahadur representation.
- This thinking leads to some new first order expansion results for penalized/regularized estimators in high-dimensions.
- NBK inequalities are also proved for Cox proportional hazards model, Non-linear least squares, Equality constrained M-estimators among others.

Some Comments Contd.

- In order to apply NBK inequalities to complete the study of an estimator in any setting, one needs to choose Σ, β and bound the remainder terms in the inequalities.
- For $\hat{\beta}$ defined as a minimizer of $L_n(\cdot)$, a canonical choice of Σ, β is given by

$$\beta := \underset{\theta \in \mathbb{R}^d}{\text{argmin}}\ \mathbb{E}[L_n(\theta)] \quad \text{and} \quad \Sigma := \mathbb{E}[\ddot{L}_n(\beta)].$$

- For independent as well as a weakly dependent sub-Gaussian observations,

$$\max\{\mathcal{D}^{\Sigma}(\beta),\ \|\Sigma^{-1}\dot{L}_n(\beta)\|_{\Sigma}\} = O_p(\sqrt{d/n}),$$

  which implies optimal rates for Bahadur representation.
- In case of variable selection, we have

$$\max_{|M|\le k}\ \max\{\mathcal{D}^{\Sigma}_M(\beta_M),\ \|\Sigma_M^{-1}\dot{L}_n(\beta_M)\|_{\Sigma_M}\} = O_p(\sqrt{k\log(ed/k)/n}).$$

  This solves the post-selection inference problem with increasing dimension and much more.

Summary and Conclusions

- We have introduced the idea of studying estimators in a deterministic way.
- NBK inequalities solve almost all problems about an estimator in one shot:
  - They imply Berry–Esseen type bounds and hence (finite sample) normal approximation results can follow.
  - They allow for understanding the effects of increasing dependence between observations, increasing dimension.
- Importantly in the context of reproducibility, NBK inequalities allow study of estimators obtained after data snooping.
- In particular, it solves the problem of post-selection inference in a unified way and in the most general setting available till date.
- Further in the context of cross-validation/subsampling, NBK inequalities show how computation can be reduced at the expense of very small approximation error.
- Application of a (proximal) variant of Newton's method for penalized or constrained estimators leads to first order expansion results.

Summary and Conclusions

- We have introduced the idea of studying estimators in a deterministic way.
- NBK inequalities solve almost all problems about an estimator in one shot:
  - They imply Berry–Esseen type bounds and hence (finite sample) normal approximation results can follow.
  - They allow for understanding the effects of increasing dependence between observations, increasing dimension.
- Importantly in the context of reproducibility, NBK inequalities allow study of estimators obtained after data snooping.
- In particular, it solves the problem of post-selection inference in a unified way and in the most general setting available till date.
- Further in the context of cross-validation/subsampling, NBK inequalities show how computation can be reduced at the expense of very small approximation error.
- Application of a (proximal) variant of Newton's method for penalized or constrained estimators leads to first order expansion results.

Thanks!

Application: Post-selection Inference

- **Uniform linear representation** result allows us to claim

$$\max_{M\in\mathcal{M}} \|\hat{\beta}_M - \beta_M\|_\infty \;\approx\; \max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n}\psi_M(X_i, Y_i)\right\|_\infty,$$

  for some vector functions ψ_M.

- **High-dimensional CLT** implies

$$\max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n}\psi_M(X_i, Y_i)\right\|_\infty \;\overset{\mathcal{L}}{\approx}\; \max_{M\in\mathcal{M}} \|G_M\|_\infty,$$

  for some Gaussian process $(G_M)_{M\in\mathcal{M}}$.

- **Corresponding multiplier bootstrap** implies

$$\max_{M\in\mathcal{M}} \|\hat{\beta}_M - \beta_M\|_\infty \;\overset{\mathcal{L}}{\approx}\; \max_{M\in\mathcal{M}} \left\|\frac{1}{n}\sum_{i=1}^{n} g_i\hat{\psi}_M(X_i, Y_i)\right\|_\infty \quad \text{Cond. on } (X_i, Y_i),$$

  for $g_1, \ldots, g_n \sim N(0, 1)$ (iid).

PoSI Contd.

- To finish inference, need to compute

$$\max_{M \in \mathcal{M}} \left\| \frac{1}{n} \sum_{i=1}^{n} g_i \hat{\psi}_M(X_i, Y_i) \right\|_{\infty},$$

  for a given set of models $\mathcal{M}$.

- Number the models in $\mathcal{M}$ as $1, 2, \ldots, N$. We have

$$x_j \; := \; \left\| \frac{1}{n} \sum_{i=1}^{n} g_i \hat{\psi}_j(X_i, Y_i) \right\|_{\infty}.$$

- Need to compute (at least approximately)

$$\|x\|_{\infty} = \max_{1 \le j \le N} |x_j|,$$

  for the vector $x = (x_1, \ldots, x_N)$.

Maximum Computation[3]

- Observe that

$$\left(\frac{1}{N}\sum_{j=1}^{N} x_j^q\right)^{1/q} \le \|x\|_\infty \le N^{1/q}\left(\frac{1}{N}\sum_{j=1}^{N} x_j^q\right)^{1/q}.$$

- If W is a random variable drawn uniformly from $\{x_1, \ldots, x_N\}$, then

$$(\mathbb{E}[W^q])^{1/q} \le \|x\|_\infty \le N^{1/q}(\mathbb{E}[W^q])^{1/q}.$$

- Hence (multiplicatively) approximating the maximum is same as approximating the **expectation** of a random variable given access to independent draws.

How many draws required to find $\mathbb{E}[W^q]$ upto a factor of $(1 \pm \varepsilon)$?

[3] Joint work (in progress) with Junhui Cai

Summary

- We have shown how the **analysis of Newton's method** can be used to derive **finite sample results for M-estimators**.
- This idea allow "easier" study of constrained/penalized M-estimators.
- Connections to AMP??
- These results imply post-selection inference for various estimation procedures including **GLM**s, **Cox Model**, **NonLinear Least Squares**, **Equality Constrained MLE**.
- Realizing PoSI in practice requires solving a maximum problem.
- $$\text{PoSI} \rightarrow \text{Maximum Estimation} \rightarrow \text{Mean Estimation.}$$
- achievable sample complexity bounds for maximum??

Maximum Computation (Contd.)

- An estimator $\hat{E}_W$ of $\mathbb{E}[W] > 0$ is an (ε, δ) approximate if

$$\mathbb{P}\left(\left|\frac{\hat{E}_W}{\mathbb{E}[W]} - 1\right| \le \varepsilon\right) \;\ge\; 1 - \delta.$$

- If a random variable $W \ge 0$ is known to satisfy

$$\mathrm{Var}(W) \;\le\; L^2(\mathbb{E}[W])^2$$

then

$$n_{\varepsilon,\delta} \;\asymp\; \frac{2L^2}{\varepsilon^2}\log\left(\frac{1}{\sqrt{2\pi}\delta}\right).$$

- If a random variable $W \in [0, B]$ for some known B, then

$$n_{\varepsilon,\delta} \;\asymp\; C\max\left\{\frac{\mathrm{Var}(W)}{\varepsilon^2(\mathbb{E}[W])^2}, \frac{B}{\varepsilon\mathbb{E}[W]}\right\}\log\left(\frac{1}{\delta}\right),$$

for some universal constant $C > 0$.