Notes on Efficiency of Estimators

Xinghao Yu*

March, 2025

In this notes, we consider a parametric statistical model $\mathcal{P}$ defined as the collection of probability measures $\{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ on some measurable space $(\mathcal{X}, \mathcal{B})$ where Θ is an open subset of $\mathbb{R}^d$. For each $\theta \in \Theta$, let p_θ be a density of P_θ w.r.t. some dominating σ-finite measure μ.

1 Recap of QMD and LAN

The classical result that stating the asymptotic normality of MLE requiring the map $\theta \mapsto p_\theta$ to be twice continuously differentiable, this makes the theorem quite weak. To see this, consider the family of Laplace distributions with scale 1 and mean $\theta \in \mathbb{R}$:

$$p_\theta(y) = \frac{1}{2}\exp(-|y-\theta|)$$

The MLE estimator for θ is the median, but it's apparently that p_θ does not satisfy the C^2 assumption, yet the MLE estimator is asymptotically normal. According to David Pollard's comment on QMD, Le Cam (1970) noted that high-order differentiability can be relaxed. This motivates the development of more general smoothness conditions that guarantee the MLE to be asymptotic normal.

Definition 1 (QMD)**.** *A statistical model $\mathcal{P}$ is quadratic mean differentiable (QMD) at $\theta \in \Theta$ if there exists a measurable vector-valued function $\eta_\theta : \mathcal{X} \mapsto \mathbb{R}^d$ such that*

$$\int_{\mathcal{X}} \left(\sqrt{dP_{\theta+h}} - \sqrt{dP_\theta} - \frac{1}{2}h^\top \eta_\theta \sqrt{dP_\theta}\right)^2 = o(||h||^2)$$

In this case, the function η_θ is called score and mean zero, and the $I_\theta := \mathbb{E}_{P_\theta}[\eta_\theta \eta_\theta^\top]$ exists and is defined as the Fisher information at θ.

Why squared root of the density? It's not just an element in $\mathcal{L}^2$ space, it's an element with norm 1.

Lemma 1 (Section 19, David Pollard)**.** *Let $\{\delta_n\}$ be a sequence of constants tending to 0. Let ξ_0, ξ_1, ... be elements of norm 1 for which $\xi_n = \xi_0 + \delta_n W + r_n$, with W a fixed element and $||r_n|| = o(\delta_n)$. Then $< \xi_0, W >= 0$ and $< \xi_0, r_n >= -\frac{1}{2}\delta_n^2 ||W||^2 + o(\delta_n^2)$.*

Example [Laplace] The family of Laplace distribution is QMD, the points where the derivative of the log-likelihood does not exist form a Lebesgue null set, so η can be arbitrary on these points.

$$\eta_\theta(x) = \begin{cases} 1 & \text{if } x < \theta \\ -1 & \text{if } x > \theta \\ 0 & \text{if } x = \theta \end{cases}$$

will satisfy the definition of QMD.

*School of Economics, Singapore Management University

Example [Uniform] Consider $p_\theta = \text{Unif}[0,\theta]$. This is not QMD. Note that, for $\theta_0 > 0$ and $h > 0$,

$$\int_{\theta_0}^{\theta_0+h} \Big(\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - \frac{1}{2}\eta_{\theta_0}(x)h\sqrt{p_{\theta_0}(x)}\Big)^2 d\mu(x) = \int_{\theta_0}^{\theta_0+h} \Big(\frac{1}{\sqrt{\theta_0+h}}\Big)^2 d\mu(x) = \frac{h}{\theta_0+h} = O(h)$$

Therefore, a general MLE asymptotic normality theorem which didn't rely on twice continuous differentiability rather QMD can be stated as follow.

Theorem 1 (Thm.5.39 of vdV)**.** *Suppose $\mathcal{P}$ is QMD at θ. Let i.i.d. $X_1, \cdots, X_n \sim P_\theta \in \mathcal{P}$ with $\theta \in$ Int Θ. Suppose there exist a measurable function $K(x) \in \mathcal{L}^2(\theta)$ such that*

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \le K(x)||\theta_1 - \theta_2||$$

for all x and all θ_1, θ_2 in a neighborhood of θ. If I_θ is non-singular and the MLE estimator $\hat{\theta}_n^{ML}$ is consistent for θ, then

$$\sqrt{n}(\hat{\theta}_n^{ML} - \theta) \rightsquigarrow \mathcal{N}(0, I_\theta^{-1})$$

Definition 2 (LAN)**.** *For a sequence of models $\{\mathcal{P}_n\}_{n\in\mathbb{N}}$ defined as the sequence of collections of probability measures $\{P_{\theta,n} : \theta \in \Theta \subseteq \mathbb{R}^d\}$ on the corresponding sequence of measurable spaces $(\mathcal{X}_n, \mathcal{B}_n)$. It is locally asymptotic normal (LAN) at the point $\theta \in$ Int Θ if there exists a mapping $\Delta_n : \mathcal{X}_n \mapsto \mathbb{R}^d$ and matrix $K \succeq 0$ such that for all $h \in \mathbb{R}^d$ and n large enough that $\theta + h/\sqrt{n} \in \Theta$,*

$$\log \frac{dP_{\theta+h/\sqrt{n},n}}{dP_{\theta,n}} = h^\top \Delta_n - \frac{1}{2}h^\top K h + o_{P_{\theta,n}}(||h||)$$

and

$$\Delta_n \underset{P_{\theta,n}}{\rightsquigarrow} \mathcal{N}(0, K)$$

We call K the precision matrix for the sequence of models $\{\mathcal{P}_n\}_{n\in\mathbb{N}}$.

Proposition 1 (QMD Implies LAN)**.** *Suppose $\mathcal{P}$ is QMD at θ with socre η_θ. Let i.i.d. $X_1, \cdots, X_n \sim P_\theta \in \mathcal{P}$, for every $h \in \mathbb{R}^d$, as $n \to \infty$,*

$$\log \frac{dP_{\theta+h/\sqrt{n},n}}{dP_{\theta,n}}(X_1, \cdots, X_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h^\top \eta_\theta(X_i) - \frac{1}{2}h^\top I_\theta h + o_{P_{\theta,n}}(1)$$

Example [Local Asymptotic Analysis of MLE] Consider following set-up. Let i.i.d. $X_1, \cdots, X_n \sim P_\theta \in \mathcal{P}$, where $\mathcal{P}$ is QMD at θ with socre η_θ. Assume the MLE estimator $\hat{\theta}_n^{ML}$ exists. If we aim to study the limit distribution of $\sqrt{n}(\hat{\theta}_n^{ML} - \theta - \frac{h}{\sqrt{n}})$ under the alternative distribution $P_{\theta+h/\sqrt{n},n}$. Since h is fixed, it suffices to study $\sqrt{n}(\hat{\theta}_n^{ML} - \theta)$ under the alternative.

Step 1: Check contiguity $P_{\theta+h/\sqrt{n},n} \lhd P_{\theta,n}$. According to Prop.1, we have

$$\log \frac{dP_{\theta+h/\sqrt{n},n}}{dP_{\theta,n}} \underset{P_{\theta,n}}{\rightsquigarrow} \mathcal{N}(-\frac{1}{2}h^\top I_\theta h, h^\top I_\theta h)$$

By applying the Normal example of Le Cam's 1st Lemma, we can verify the desired result.

Step 2: Establish joint convergence of the MLE and the tilted measure as

$$\Big(\sqrt{n}(\hat{\theta}_n^{ML} - \theta),\ \log \frac{dP_{\theta+h/\sqrt{n},n}}{dP_{\theta,n}}\Big) \underset{P_{\theta,n}}{\rightsquigarrow} \mathcal{N}\Big(\begin{bmatrix} 0 \\ -\frac{1}{2}h^\top I_\theta h \end{bmatrix}, \begin{bmatrix} I_\theta^{-1} & h \\ h^\top & h^\top I_\theta h \end{bmatrix}\Big)$$

Step 3: Characterize the alternative limiting distribution by Le Cam's 3rd Lemma, that is

$$\sqrt{n}\Big(\hat{\theta}_n^{ML} - [\theta + \frac{h}{\sqrt{n}}]\Big) \underset{P_{\theta+\frac{h}{\sqrt{n}},n}}{\rightsquigarrow} \mathcal{N}(0, I_\theta^{-1}) \tag{1}$$

Theorem 2 (Thm.7.10 of vdV). *Let $\{\mathcal{P}_n\}_{n\in\mathbb{N}}$ be LAN at θ with non-singular K. If for every $h \in \mathbb{R}^d$*

$$Z_n \underset{P_{\theta+\frac{h}{\sqrt{n}},n}}{\rightsquigarrow} L_{\theta,h}$$

then

$$Z_n \underset{P_{\theta+\frac{h}{\sqrt{n}},n}}{\rightsquigarrow} Z$$

where Z is a randomized statistics in $\{\mathcal{N}(h, K^{-1})\}_{h\in\mathbb{R}^d}$.

2 Asymptotic Optimality

Motivation 1: In classical parametric theory, suppose that we are interested in estimating $\psi(\theta)$ based on the data $X_1, \cdots, X_n \sim P_\theta$, where $\psi : \Theta \mapsto \mathbb{R}^k$ is a known function. A natural question can be that: How can we justify the performance of an estimator $T_n := T(X_1, \cdots, X_n)$ for $\psi(\theta)$? From a finite sample perspective, we have one classical result for the best known lower bound on the performance of any unbiased estimator of $\psi(\theta)$.

Theorem 3 (Cramér-Rao Lower Bound). *Suppose the map $\psi : \Theta \mapsto \mathbb{R}^k$ is differentiable at $\theta \in \Theta$ with the $k \times d$ Jacobian matrix $\nabla_\theta\psi(\theta)$ whose ij-th element is given by $\frac{\partial}{\partial\theta_j}\psi_i(\theta)$. Then for any unbiased estimator T_n, i.e., $\mathbb{E}_{P_\theta}[T_n] - \psi(\theta) = 0$ for every $\theta \in \Theta$, it holds that*

$$cov_\theta(T_n) \succeq \nabla_\theta\psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta\psi(\theta)^\top$$

Then, in the asymptotic sense, for any asymptotically normal estimators T_n with

$$\sqrt{n}(T_n - \psi(\theta)) \rightsquigarrow \mathcal{N}(0, \Sigma_\theta)$$

for every θ. Can we analogously conclude that $\Sigma_\theta \succeq \nabla_\theta\psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta\psi(\theta)^\top$?
Note that, MLE estimator $T_n = \psi(\hat{\theta}_n^{ML})$ achieves the proposed lower bound, to see this, according to Thm.1 and delta method,

$$\sqrt{n}\big(\psi(\hat{\theta}_n^{ML}) - \psi(\theta)\big) \rightsquigarrow \mathcal{N}(0, \nabla_\theta\psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta\psi(\theta)^\top)$$

Motivation 2: When making statistical decision, in finite sample sense, we can use several optimality criterion such as admissibility and minimaxity to choose among all kinds of estimators.

Definition 3 (Admissibility). *Given a risk function $R(T_n, \psi(\theta)) = \mathbb{E}_{P_\theta}[L(T_n, \psi(\theta))]$ with proper loss function, we say an estimator T_n is inadmissible (in a class $\mathcal{C}$ of estimators) if T_n is dominated by another estimator. i.e., there exists another estimator in class $\tilde{T}_n \in \mathcal{C}$ with*

$$R(\tilde{T}_n, \psi(\theta)) \leq R(T_n, \psi(\theta)), \quad \forall\ \theta$$
$$R(\tilde{T}_n, \psi(\theta)) < R(T_n, \psi(\theta)), \quad \exists\ \theta$$

And we say an estimator (decision rule) T_n is admissible if there is no other estimators $\tilde{T}_n \in \mathcal{C}$ that dominates it.

Admissibility is a simple notion that simply checks, intuitively, that T_n is good somewhere. Next, we also have the notion of minimaxity, which intuitively checks that T_n is terrible nowhere.

Definition 4 (Minimaxity). *We say that T_n is minimax with respect to $\mathcal{C}$ if*

$$T_n \in \arg\min_{T\in\mathcal{C}} \max_{\theta\in\Theta} R(T, \psi(\theta))$$

In other words, T_n minimizes the global maximum risk for an unknown value of θ.

Then, can we extend these concepts under asymptotic setting to design "asymptotic optimal" estimators?

Definition 5 (Hodges' Estimator)**.** *Consider the Gaussian location model $X_1, \cdots, X_n \sim \mathcal{N}(\theta, 1)$ with $\theta \in \mathbb{R}$, the Hodges' estimator is defined as*

$$\tilde{\theta}_n = \begin{cases} \bar{X}_n & \text{if } |\bar{X}_n| \geq n^{-1/4}, \\ 0 & o/w \end{cases}$$

Note that, when $\theta = 0$, $\bar{X}_n$ converges to 0 at the rate $\sqrt{n}$ by CLT, so it converges faster than $n^{-1/4}$, which means $\tilde{\theta}_n$ equals to 0 w.h.p.. When outside a local neighborhood of $\theta = 0$, $\tilde{\theta}_n$ behaves like the MLE estimator $\hat{\theta}_n = \bar{X}_n$. Thus, for the asymptotic distribution

$$\sqrt{n}(\tilde{\theta}_n - \theta) \rightsquigarrow \begin{cases} 0 & \text{if } \theta = 0 \\ \mathcal{N}(0,1) & \text{if } \theta \neq 0 \end{cases}$$

At $\theta = 0$, it seems like $\tilde{\theta}_n$ is even more asymptotically efficient than the MLE estimator since the asymptotic variance for $\tilde{\theta}_n$ is 0 while the asymptotic variance for MLE is 1. From the risk perspective, consider squared loss function $L(u - v) = (u - v)^2$. Then we have

$$\lim_{n\to\infty} nR(\hat{\theta}_n, \theta) = \lim_{n\to\infty} n\mathbb{E}_{P_\theta}[(\bar{X}_n - \theta)^2] = 1$$

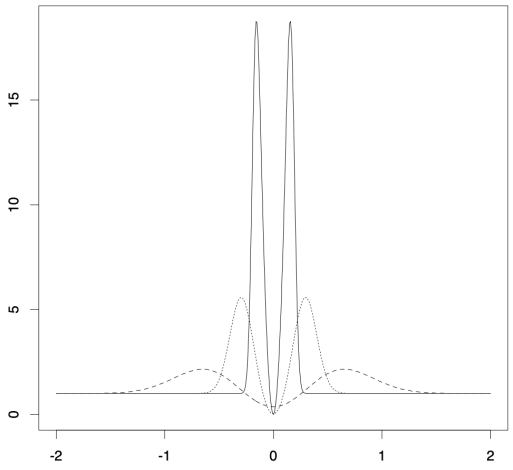
And for the Hodges' estimator, we have

$$\lim_{n\to\infty} nR(\tilde{\theta}_n, \theta) = \begin{cases} 0 & \text{if } \theta = 0 \\ 1 & \text{if } \theta \neq 0 \end{cases}$$

Hence by definition, the MLE estimator $\hat{\theta}_n$ is dominated by the Hodges' estimator $\tilde{\theta}_n$ asymptotically. But under finite sample situation, for the Hodges' estimator,

$$\sup_{\theta\in\mathbb{R}} R(\tilde{\theta}_n, \theta) \to \infty, \quad \text{as } n \to \infty$$

Thus, $\tilde{\theta}_n$ has bad minimax risk, it "buys" its better asymptotic performance at $\theta = 0$ at the expense of worse behavior for points "close" to zero, especially at $n^{-1/4}$.



Quadratic risk function of the Hodges estimator based on a sample of size 10 (dashed), 100 (dotted) and 1000 (solid) observations from the $N(\theta, 1)$-distribution.

The fix is to apply minimaxity to the local asymptotic risk. Suppose δ_n is $\sqrt{n}$-consistent for θ and consider squared loss function $L(u - v) = (u - v)^2$. Define

$$R_s(\delta_n, \theta, h) = \lim_{n\to\infty} n\mathbb{E}_{P_{\theta+\frac{h}{\sqrt{n}},n}}\left[\left(\delta_n - \theta - \frac{h}{\sqrt{n}}\right)^2\right]$$

Then the optimal minimax estimator δ_n in the sense of local asymptotic risk is defined as the minimizer of

$$\sup_h R_s(\delta_n, \theta, h)$$

for every θ.

Theorem 4 (Thm.8.3 of vdV)**.** *Suppose the statistical model $\mathcal{P}$ is QMD at $\theta \in \Theta$ with nonsingular Fisher information matrix I_θ. Suppose the map $\psi : \Theta \mapsto \mathbb{R}^k$ is differentiable at $\theta \in \Theta$ with the $k \times d$ Jacobian matrix $\nabla_\theta \psi(\theta)$. Let T_n be an estimator of $\psi(\theta)$. Suppose that*

$$\sqrt{n}\Big(T_n - \psi(\theta + \frac{h}{\sqrt{n}})\Big) \underset{P_{\theta+\frac{h}{\sqrt{n}},n}}{\rightsquigarrow} L_{\theta,h}$$

for all $h \in \mathbb{R}^d$. Then there exists a randomized function T of $X \sim \mathcal{N}(h, I_\theta^{-1})$ such that,

$$T - \nabla_\theta \psi(\theta) \cdot h \sim L_{\theta,h}$$

for all $h \in \mathbb{R}^d$.

Proof Sketch of Thm.4: Note that,

$$\sqrt{n}\Big(T_n - \psi(\theta + \frac{h}{\sqrt{n}})\Big) = \sqrt{n}\Big(T_n - \psi(\theta)\Big) + \sqrt{n}\Big(\psi(\theta) - \psi(\theta + \frac{h}{\sqrt{n}})\Big)$$

For the second term,

$$\sqrt{n}\Big(\psi(\theta) - \psi(\theta + \frac{h}{\sqrt{n}})\Big) = \sqrt{n}\frac{\psi(\theta) - \psi(\theta + \frac{h}{\sqrt{n}})}{h/\sqrt{n}}\frac{h}{\sqrt{n}} \to -\nabla_\theta \psi(\theta) \cdot h$$

Regarding the first term, just follows Thm.2. ■

For intuition of Thm.4, since the loss function is non-negative c.t.s., according to CMT and Portmanteau Lemma, we can wirte

$$\begin{aligned}
\sup_h R_s(\delta_n, \theta, h) &= \sup_h \lim_{n\to\infty} n\mathbb{E}_{P_{\theta+\frac{h}{\sqrt{n}},n}}\Big[\big(\delta_n - \theta - \frac{h}{\sqrt{n}}\big)^2\Big] \\
&= \sup_h \liminf_{n\to\infty} n\mathbb{E}_{P_{\theta+\frac{h}{\sqrt{n}},n}}\Big[\big(\delta_n - \theta - \frac{h}{\sqrt{n}}\big)^2\Big] \\
&\geq \sup_h \mathbb{E}_{L_{\theta,h}}\big[(T-h)^2\big] \\
&\geq \inf_T \sup_h \mathbb{E}_{L_{\theta,h}}\big[(T-h)^2\big] \\
&= \mathbb{E}[(X-h)^2] \\
&= I_\theta^{-1}
\end{aligned}$$

where the third line follows the Portmanteau Lemma and Thm.4 where $\nabla_\theta \psi(\theta) = I$, in this way we transform the calculation of local asymptotic minimax risk to the minimaxity of squared loss in a Gaussian location family, i.e., T is an estimator of the location of $\mathcal{N}(h, I_\theta^{-1})$; the second last line follows that, the estimator $T = \hat{h}^{ML} = X$ is minimax w.r.t. squared loss in the Gaussian location family, thus we know that $X - h \sim \mathcal{N}(0, I_\theta^{-1})$.

3 Asymptotic Convolution Theorem

To fix the asymptotic theorem on optimality, the first attempt is to restrict to a family of regular estimators.

Definition 6 (Regular Estimator)**.** *An estimator T_n is called regular at $\theta \in \Theta$ for estimating $\psi(\theta)$ if, for any $h \in \mathbb{R}^d$,*

$$\sqrt{n}\Big(T_n - \psi(\theta + \frac{h}{\sqrt{n}})\Big) \underset{P_{\theta+\frac{h}{\sqrt{n}},n}}{\rightsquigarrow} L_\theta$$

where L_θ is an arbitrary probability measure that does not depend on h.

Example [MLE] According to Eq.1 we have calculated in Section 1, it's regular at θ.
Example [Hodges] For the Gaussian location model i.i.d. $X_1, \cdots, X_n \sim \mathcal{N}(\theta, 1)$. Consider

$$\hat{\theta}_n^H = \begin{cases} \bar{X}_n & \text{if } |\bar{X}_n| \geq n^{-1/4}, \\ \epsilon \bar{X}_n & \text{o/w} \end{cases}$$

for arbitrary small $\epsilon > 0$. It's not regular at $\theta = 0$ since

$$\sqrt{n}\Big(\hat{\theta}_n^H - \frac{h}{\sqrt{n}}\Big) \underset{P_{\frac{h}{\sqrt{n}},n}}{\rightsquigarrow} \mathcal{N}\big(h(1-\epsilon), \epsilon^2\big)$$

To see this, let $\sqrt{n}(\bar{X}_n - \theta) \rightsquigarrow Z \sim \mathcal{N}(0,1)$, then for $\theta_n = \frac{h}{\sqrt{n}}$,

$$\begin{aligned}
\sqrt{n}(\hat{\theta}_n^H - \theta_n) &= \sqrt{n}(\bar{X}_n - \theta_n) \cdot \mathbf{1}_{\{|\bar{X}_n| \geq n^{-1/4}\}} + \sqrt{n}(\epsilon \bar{X}_n - \theta_n) \cdot \mathbf{1}_{\{|\bar{X}_n| < n^{-1/4}\}} \\
&= \sqrt{n}(\bar{X}_n - \theta_n) \cdot \mathbf{1}_{\{\sqrt{n}|\bar{X}_n - \theta_n + \theta_n| \geq n^{1/4}\}} + \big[\epsilon\sqrt{n}(\bar{X}_n - \theta_n) + \sqrt{n}\theta_n(\epsilon - 1)\big] \cdot \mathbf{1}_{\{\sqrt{n}|\bar{X}_n - \theta_n + \theta_n| < n^{1/4}\}} \\
&\rightsquigarrow Z \cdot \mathbf{1}_{\{|Z+h| \geq n^{1/4}\}} + \big[\epsilon Z + h(\epsilon - 1)\big] \cdot \mathbf{1}_{\{|Z+h| < n^{1/4}\}} \\
&= \epsilon Z + h(\epsilon - 1) \sim \mathcal{N}\big(h(1-\epsilon), \epsilon^2\big)
\end{aligned}$$

Theorem 5 (Hájek-Le Cam Convolution Theorem)**.** *Suppose the statistical model $\mathcal{P}$ is QMD at $\theta \in \Theta$ with nonsingular Fisher information matrix I_θ. Suppose the map $\psi : \Theta \mapsto \mathbb{R}^k$ is differentiable at $\theta \in \Theta$ with the $k \times d$ Jacobian matrix $\nabla_\theta \psi(\theta)$. If estimator T_n is regular at θ and*

$$\sqrt{n}(T_n - \psi(\theta)) \underset{P_{\theta,n}}{\rightsquigarrow} L_\theta$$

for θ, then there exists some probability measure M_θ such that

$$L_\theta = Z_\theta + \Delta_\theta \sim \mathcal{N}\Big(0, \nabla_\theta \psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta \psi(\theta)^\top\Big) * M_\theta$$

where $Z_\theta \sim \mathcal{N}\Big(0, \nabla_\theta \psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta \psi(\theta)^\top\Big)$ and $\Delta_\theta \sim M_\theta$ are independent.

The notation $*$ denotes the convolution operation between two distributions and should be interpreted as follows: If $X \sim P$ and $Y \sim Q$ and $X \perp Y$, then $X + Y \sim P * Q$.

Lemma 2 (Lemma 8.14 of vdV)**.** *Given i.i.d. data $X_1, \cdots, X_n \sim P_\theta \in \mathcal{P}$. Suppose $\mathcal{P}$ is QMD at $\theta \in \Theta$ with score η_θ and nonsingular Fisher information matrix I_θ. If T_n is a regular and asymptotic efficient estimator for estimating $\psi(\theta)$, then T_n is asymptotic linear with unique influence function*

$$\tilde{\ell}_\psi(\cdot) = \nabla_\theta \psi(\theta) I_\theta^{-1} \eta_\theta(\cdot)$$

such that,

$$\sqrt{n}(T_n - \psi(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\ell}_\psi(X_i) + o_p(1) \rightsquigarrow \mathcal{N}\Big(0, \nabla_\theta \psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta \psi(\theta)^\top\Big)$$

Example [Normal Distribution] Suppose i.i.d. $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Let $\hat{\mu}_n$ and $\hat{\sigma}^2$ denote the MLE estimators of μ and σ^2, respectively. Then we can show that, their i-th influence functions are given by $X_i - \mu$ and $(X_i - \mu)^2 - \sigma^2$, respectively.

Then, since Thm. 5 already formalize the notion that the MLE is optimal among the nice regular families. But how do superefficient estimators exist? The second fix is to show that the superefficient can only be achieved on a set of Lebesgue measure 0.

Theorem 6 (a.e. Convolution Theorem)**.** *Suppose that the statistical model $\mathcal{P}$ is QMD at $\theta \in \Theta$ with nonsingular Fisher information matrix I_θ. Suppose the map $\psi : \Theta \mapsto \mathbb{R}^k$ is differentiable at $\theta \in \Theta$ with the $k \times d$ Jacobian matrix $\nabla_\theta \psi(\theta)$. Let T_n be any estimator such that*

$$\sqrt{n}(T_n - \psi(\theta)) \underset{P_{\theta,n}}{\rightsquigarrow} L_\theta$$

for every θ, then there exists some probability measure M_θ such that

$$L_\theta = \mathcal{N}\Big(0, \nabla_\theta \psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta \psi(\theta)^\top\Big) * M_\theta$$

for Lebesgue almost every θ.

This theorem does not contradict the results of the Hodges' estimator. In that case, $\mathcal{P} = \{\mathcal{N}(\theta, 1) : \theta \in \mathbb{R}\}, \psi(\theta) = \theta$, and $\mathcal{N}\Big(0, \nabla_\theta \psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta \psi(\theta)^\top\Big) = \mathcal{N}(0, 1)$. For every $\theta \neq 0, \sqrt{n}(\tilde{\theta}_n - \theta) = \sqrt{n}(\bar{X}_n - \theta) \rightsquigarrow \mathcal{N}(0, 1)$, thus, the theorem is satisfied for M_θ the distribution with unit mass at 0.
In order to assert that $\mathcal{N}\Big(0, \nabla_\theta \psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta \psi(\theta)^\top\Big)$ is in fact the "best" limit distribution, we need the following lemma:

Lemma 3. *For any bowl-shaped loss function L on $\mathbb{R}^d$, every probability distribution M on $\mathbf{R}^d$, and every covariance matrix Σ,*

$$\int L(x) d\mathcal{N}(0, \Sigma) \leq \int L(x) d(\mathcal{N}(0, \Sigma) * M)$$

Thus, if "best" is measured by any bowl-shaped loss function (including squared error loss), then, under the assumptions of Thm. 6, MLE estimators are "best" for Lebesgue almost every θ.

4 Local Asymptotic Minimax Theorem

We can extend the derivation in Section 2 into a more general Theorem.

Theorem 7 (Local Asymptotic Minimax (LAM))**.** *Suppose that the statistical model $\mathcal{P}$ is QMD at $\theta \in \Theta$ with nonsingular Fisher information matrix I_θ. Suppose the map $\psi : \Theta \mapsto \mathbb{R}^k$ is differentiable at $\theta \in \Theta$ with the $k \times d$ Jacobian matrix $\nabla_\theta \psi(\theta)$. For every bowl-shaped loss function $L(\cdot)$, i.e., L is symmetric and quasi-convex, then for any sequence of estimators $\{T_n\}$ of $\psi(\theta)$,*

$$\lim_{c \to \infty} \liminf_{n \to \infty} \sup_{||h|| \leq c} \mathbb{E}_{P_{\theta + \frac{h}{\sqrt{n}}, n}} \Big[L\Big(\sqrt{n}\big(T_n - \psi(\theta + \frac{h}{\sqrt{n}})\big)\Big)\Big] \geq \mathbb{E}[L(Z)]$$

with $Z \sim \mathcal{N}\Big(0, \nabla_\theta \psi(\theta) \cdot I_\theta^{-1} \cdot \nabla_\theta \psi(\theta)^\top\Big)$.

The proof of LAM can be summarized with the following diagram,

$$\text{QMD Model} \xrightarrow{\text{Prop.1}} \text{LAN} \xrightarrow{\text{Thm.4}} \text{Gaussian Location Model} \xrightarrow{\text{Thm.8}} \text{LAM}$$

Consider the Gaussian location model where we have an unknown mean $\theta \in \mathbb{R}^d$ and observe $X \sim \mathcal{N}(\theta, \Sigma)$ with a known non-singular covariance Σ. Consider a bowl-shaped loss function L, a natural estimator of θ is $\hat{\theta} = X$. The following theorem provides a sense that $\hat{\theta}$ is minimax.

Theorem 8 (Anderson)**.** *For the proposed Gaussian location model and any bowl-shaped loss function L, it holds that*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}_\theta[L(\hat{\theta} - \theta)] = \mathbb{E}[L(Z)]$$

with $Z \sim \mathcal{N}(0, \Sigma)$.

Lemma 4 (Anderson's Lemma)**.** *Let $Z \sim \mathcal{N}(0, \Sigma)$ and L be bowl-shaped. Then*

$$\min_{x \in \mathbb{R}^d} \mathbb{E}[L(Z + x)] = \mathbb{E}[L(Z)]$$

Theorem 9 (Prépoka-Leindler or Functional Brunn-Minkowski). *Let $\lambda \in (0,1)$, and f, g, h be three non-negative real-valued functions on $\mathbb{R}^d$. If*

$$h(\lambda x + (1-\lambda)y) \geq f(x)^{\lambda} g(y)^{1-\lambda}, \quad \forall\ x, y \in \mathbb{R}^d$$

then

$$\int h(x)dx \geq \Big(\int f(x)dx\Big)^{\lambda}\Big(\int g(x)dx\Big)^{1-\lambda}$$

Example [Bernoulli Trial] Suppose $X_1, \cdots, X_n \sim \text{Bern}(p)$, where $p \in [0,1]$, we are interested in estimating $\sqrt{p}$, consider the squared loss function: $L(u,v) = (u-v)^2$. We have following correspondence,

$$I_p = \frac{1}{p(1-p)},\ \psi(\cdot) = \sqrt{\cdot},\ \nabla_p \psi(p) = \frac{1}{2\sqrt{p}}$$

Thus by LAM, for every $p_0 \in [0,1]$,

$$\lim_{h\to\infty} \liminf_{n\to\infty} \sup_{|p'-p_0| \leq \frac{h}{\sqrt{n}}} n\mathbb{E}_{p'}[(T_n - \sqrt{p'})^2] \geq \frac{(\nabla_p \psi(p)|_{p=p_0})^2}{I_{p_0}} = \frac{p_0(1-p_0)}{(2\sqrt{p_0})^2} = \frac{1-p_0}{4}$$

Translate this into a global minimax lower bound is

$$\inf_{T_n} \sup_{p\in[0,1]} \mathbb{E}_p[(T_n - \sqrt{p})^2] \geq \frac{1+o_n(1)}{4n}$$

Definition 7 (Shrinkage Estimators, Section 8.8 of vdV). *Let $X_1, \cdots, X_n \sim \mathcal{N}(\theta, I_k)$ where $k \geq 3$. Then the James-Stein estimator for θ is defined as*

$$\hat{\theta}_n^{JS} = \Big(1 - \frac{(k-2)}{n||\bar{X}_n||^2}\Big)\bar{X}_n$$

Let verify that $\hat{\theta}_n^{JS}$ is uniformly better than $\hat{\theta}_n^{ML}$ in single observation case. Consider $X \sim \mathcal{N}(\theta, I_k)$, and squared loss function $L(u-v) = ||u-v||^2$,

$$\begin{aligned}
R(\hat{\theta}_n^{JS}, \theta) &= \mathbb{E}[||\hat{\theta}_n^{JS} - \theta||^2] = \mathbb{E}\Big[||\Big(1 - \frac{(k-2)}{||X||^2}\Big)X - \theta||^2\Big] = \mathbb{E}\Big[||X - \theta - \frac{(k-2)}{||X||^2}X||^2\Big] \\
&= \mathbb{E}\Big[||X-\theta||^2\Big] + \mathbb{E}\Big[||\frac{(k-2)}{||X||^2}X||^2\Big] - \mathbb{E}\Big[2(k-2)\sum_{i=1}^{k}\frac{(X_i - \theta_i)X_i}{||X||^2}\Big] \\
&= k + (k-2)^2\mathbb{E}\Big[\frac{1}{||X||^2}\Big] - 2(k-2)\sum_{i=1}^{k}\mathbb{E}\Big[\frac{||X||^2 - 2X_i^2}{||X||^4}\Big] \\
&= k - (k-2)^2\mathbb{E}\Big[\frac{1}{||X||^2}\Big] \\
&< k = \mathbb{E}\Big[||X-\theta||^2\Big] = R(\hat{\theta}_n^{ML}, \theta)
\end{aligned}$$

when $k \geq 3$.