Chapter 4. An Introduction to Asymptotic Theory[*]

We introduce some basic asymptotic theory in this chapter, which is necessary to understand the asymptotic properties of the LSE. For more advanced materials on the asymptotic theory, see Dudley (1984), Shorack and Wellner (1986), Pollard (1984, 1990), Van der Vaart and Wellner (1996), Van der Vaart (1998), Van de Geer (2000) and Kosorok (2008). For reader-friendly versions of these materials, see Gallant (1987), Gallant and White (1988), Newey and McFadden (1994), Andrews (1994), Davidson (1994) and White (2001). Our discussion is related to Section 2.1 and Chapter 7 of Hayashi (2000), Appendix C and D of Hansen (2007) and Chapter 3 of Wooldridge (2010). In this chapter and the next chapter, $\|\cdot\|$ always means the Euclidean norm.

1 Five Weapons in Asymptotic Theory

There are five tools (and their extensions) that are most useful in asymptotic theory of statistics and econometrics. They are the weak law of large numbers (WLLN, or LLN), the central limit theorem (CLT), the continuous mapping theorem (CMT), Slutsky's theorem,[1] and the Delta method. We only state these five tools here; detailed proofs can be found in the techinical appendix.

To state the WLLN, we first define the convergence in probability.

Definition 1 *A random vector Z_n* converges in probability *to Z as $n \to \infty$, denoted as $Z_n \xrightarrow{p} Z$, if for any $\delta > 0$,*

$$\lim_{n\to\infty} P(\|Z_n - Z\| > \delta) = 0.$$

Although the limit Z can be random, it is usually constant. The probability limit of Z_n is often denoted as $\text{plim}(Z_n)$. If $Z_n \xrightarrow{p} 0$, we denote $Z_n = o_p(1)$. When an estimator converges in probability to the true value as the sample size diverges, we say that the estimator is **consistent**. This is a good property for an estimator to possess. It means that for any given distribution of the data, there is a sample size n sufficiently large such that the estimator will be arbitrarily close to the true value with high probability. Consistency is also an important preliminary step in establishing other important asymptotic approximations.

[*] Email: pingyu@hku.hk

[1] Eugen Slutsky (1880-1948) was a Russian/Soviet mathematical statistician and economist, who is also famous for the Slutsky equation in microeconomics.

Theorem 1 (WLLN) *Suppose $X_1, \cdots, X_n, \cdots$ are i.i.d. random vectors, and $E\left[\|X\|\right] < \infty$; then as $n \to \infty$,*

$$\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E\left[X\right].$$

The CLT tells more than the WLLN. To state the CLT, we first define the convergence in distribution.

Definition 2 *A random k vector Z_n* converges in distribution *to Z as $n \to \infty$, denoted as $Z_n \xrightarrow{d} Z$, if*

$$\lim_{n\to\infty} F_n(z) = F(z),$$

at all z where $F(\cdot)$ is continuous, where F_n is the cdf of Z_n and F is the cdf of Z.

Usually, Z is normally distributed, so all $z \in \mathbb{R}^k$ are continuity points of F. If Z_n converges in distribution to Z, then Z_n is **stochastically bounded** and we denote $Z_n = O_p(1)$. Rigorously, $Z_n = O_p(1)$ if $\forall \varepsilon > 0$, $\exists M_\varepsilon < \infty$ such that $P(\|Z_n\| > M_\varepsilon) < \varepsilon$ for any n. If $Z_n = o_p(1)$, then $Z_n = O_p(1)$. We can show that $o_p(1)+o_p(1) = o_p(1)$, $o_p(1)+O_p(1) = O_p(1)$, $O_p(1)+O_p(1) = O_p(1)$, $o_p(1)o_p(1) = o_p(1)$, $o_p(1)O_p(1) = o_p(1)$, and $O_p(1)O_p(1) = O_p(1)$.

Exercise 1 *(i) If $X_n \xrightarrow{d} X$, and $Y_n - X_n = o_p(1)$, then $Y_n \xrightarrow{d} X$. (ii) Show that $X_n \xrightarrow{p} c$ if and only if $X_n \xrightarrow{d} c$. (iii) Show that $o_p(1)O_p(1) = o_p(1)$.*

Theorem 2 (CLT) *Suppose $X_1, \cdots, X_n, \cdots$ are i.i.d. random variables, $E\left[X\right] = 0$, and $Var(X) = 1$; then $\sqrt{n}\overline{X}_n \xrightarrow{d} N(0,1)$.*

$\sqrt{n}\overline{X}_n \xrightarrow{d} N(0,1)$ implies $\overline{X}_n \xrightarrow{p} 0$, so the CLT is stronger than the WLLN. $\overline{X}_n \xrightarrow{p} 0$ means $\overline{X}_n = o_p(1)$, but does not provide any information about $\sqrt{n}\overline{X}_n$. The CLT tells that $\sqrt{n}\overline{X}_n = O_p(1)$ or $\overline{X}_n = O_p(n^{-1/2})$. But the WLLN does not require the second moment finite; that is, a stronger result is not free. This result can be extended to the multi-dimensional case with nonstandardized mean and variance: suppose $X_1, \cdots, X_n, \cdots$ are i.i.d. random k vectors, $E\left[X\right] = \boldsymbol{\mu}$, and $Var(X) = \boldsymbol{\Sigma}$; then $\sqrt{n}\left(\overline{X}_n - \boldsymbol{\mu}\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$.

Theorem 3 (CMT) *Suppose $X_1, \cdots, X_n, \cdots$ are random k vectors, and g is a continuous function on the support of X (to $\mathbb{R}^l$) a.s. P_X; then*

$$\begin{aligned} X_n \xrightarrow{p} X &\implies g(X_n) \xrightarrow{p} g(X); \\ X_n \xrightarrow{d} X &\implies g(X_n) \xrightarrow{d} g(X). \end{aligned}$$

The CMT was first proved by Mann and Wald (1943) and is therefore sometimes referred to as the **Mann-Wald Theorem**. Note that the CMT allows the function g to be discontinuous but the probability of being at a discontinuity point is zero. For example, the function $g(u) = u^{-1}$ is discontinuous at $u = 0$, but if $X_n \xrightarrow{d} X \sim N(0,1)$ then $P(X = 0) = 0$ so $X_n^{-1} \xrightarrow{d} X^{-1}$.

In the CMT, X_n converges to X *jointly* in various modes of convergence. For the convergence in probability ($\xrightarrow{p}$), marginal convergence implies joint convergence, so there is no problem if we substitute joint convergence by marginal convergence. But for the convergence in distribution ($\xrightarrow{d}$), $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} Y$ does not imply $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ Y \end{pmatrix}$. Nevertheless, there is a special case where this result holds, which is Slutsky's theorem.

Theorem 4 (Slutsky's Theorem) *If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{d} c \left(\Longleftrightarrow Y_n \xrightarrow{p} c\right)$, where c is a constant, then $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$. This implies $X_n+Y_n \xrightarrow{d} X+c$, $Y_nX_n \xrightarrow{d} cX$, $Y_n^{-1}X_n \xrightarrow{d} c^{-1}X$ when $c \neq 0$. Here X_n, Y_n, X, c can be understood as vectors or matrices as long as the operations are compatible.*

One important application of Slutsky's Theorem is as follows.

Example 1 *Suppose $X_n \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma})$, and $Y_n \xrightarrow{p} \mathbf{\Sigma}$; then $Y_n^{-1/2}X_n \xrightarrow{d} \mathbf{\Sigma}^{-1/2}N(0, \mathbf{\Sigma}) = N(0, \mathbf{I})$, where $\mathbf{I}$ is the identity matrix.*

Combining with the CMT, we have the following important application.

Example 2 *Suppose $X_n \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma})$, and $Y_n \xrightarrow{p} \mathbf{\Sigma}$; then $X_n' Y_n^{-1} X_n \xrightarrow{d} \chi_k^2$, where k is the dimension of X_n.*

Another important application of Slutsky's theorem is the Delta method.

Theorem 5 (Delta Method) *Suppose $\sqrt{n}\,(Z_n - \mathbf{c}) \xrightarrow{d} Z \sim N(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{c} \in \mathbb{R}^k$, and $g(z) : \mathbb{R}^k \to \mathbb{R}^l$. If $\frac{dg(z)}{dz'}$ is continuous at $\mathbf{c}$,[2] then $\sqrt{n}\,(g(Z_n) - g(\mathbf{c})) \xrightarrow{d} \frac{dg(\mathbf{c})}{dz'}Z$.*

Intuitively,

$$\sqrt{n}\,(g(Z_n) - g(\mathbf{c})) = \sqrt{n}\frac{dg(\overline{\mathbf{c}})}{dz'}\,(Z_n - \mathbf{c})\,,$$

where $\frac{dg(\cdot)}{dz'}$ is the Jacobian of g, and $\overline{\mathbf{c}}$ is between Z_n and $\mathbf{c}$. $\sqrt{n}\,(Z_n - \mathbf{c}) \xrightarrow{d} Z$ implies that $Z_n \xrightarrow{p} \mathbf{c}$, so by the CMT, $\frac{dg(\overline{\mathbf{c}})}{dz'} \xrightarrow{p} \frac{dg(\mathbf{c})}{dz'}$. By Slutsky's theorem, $\sqrt{n}\,(g(Z_n) - g(\mathbf{c}))$ has the asymptotic distribution $\frac{dg(\mathbf{c})}{dz'}Z$. The Delta method implies that asymptotically, the randomness in a transformation of Z_n is completely controlled by that in Z_n.

Exercise 2 (*) *Suppose $g(z) : \mathbb{R}^k \to \mathbb{R}$ and $g \in C^{(2)}$ in a neighborhood of $\mathbf{c}$, $\frac{dg(\mathbf{c})}{dz'} = \mathbf{0}$ and $\frac{d^2g(\mathbf{c})}{dz'dz} \neq \mathbf{0}$. What is the asymptotic distribution of $g(Z_n)$?*

[2] This assumption can be relaxed to that $g(z)$ is differentiable at c.

2 Asymptotics for the MoM Estimator

Recall that the MoM estimator is defined as the solution to

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i|\boldsymbol{\theta}) = \mathbf{0}.$$

We can prove the MoM estimator is consistent and asymptotically normal (CAN) under some regularity conditions.[3] Specifically, the asymptotic distribution of the MoM estimator is

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{M}^{-1}\boldsymbol{\Omega}\mathbf{M}'^{-1}\right),$$

where $\mathbf{M} = \frac{dE[m(X|\boldsymbol{\theta}_0)]}{d\boldsymbol{\theta}'}$ and $\boldsymbol{\Omega} = E\left[m(X|\boldsymbol{\theta}_0)m(X|\boldsymbol{\theta}_0)'\right]$.[4] The asymptotic variance takes a *sandwich* form and can be estimated by its sample analog.

Consider a simple example. Suppose $E[X] = g(\theta_0)$ with $g \in C^{(1)}$ in a neighborhood of θ_0; then $\theta_0 = g^{-1}(E[X]) \equiv h(E[X])$. The MoM estimator of θ is to set $\overline{X} = g(\theta)$, so $\widehat{\theta} = h(\overline{X})$. By the WLLN, $\overline{X} \xrightarrow{p} E[X]$; then by the CMT, $\widehat{\theta} \xrightarrow{p} h(E[X]) = \theta_0$ since $h(\cdot)$ is continuous. Now, we derive the asymptotic distribution of $\widehat{\theta}$.

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = \sqrt{n}\left(h(\overline{X}) - h(E[X])\right) = \sqrt{n}h'\left(\overline{X}^*\right)\left(\overline{X} - E[X]\right) = h'\left(\overline{X}^*\right)\sqrt{n}\left(\overline{X} - E[X]\right),$$

where the second equality is from the mean value theorem (MVT). Because $\overline{X}^*$ is between $\overline{X}$ and $E[X]$ and $\overline{X}$ converges to $E[X]$ in probability, $\overline{X}^* \xrightarrow{p} E[X]$. So by the CMT, $h'\left(\overline{X}^*\right) \xrightarrow{p} h'(E[X])$. By the CLT, $\sqrt{n}\left(\overline{X} - E[X]\right) \xrightarrow{d} N(0, Var(X))$. Then by Slutsky's theorem,

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} h'(E[X])\,N(0, Var(X)) = N\left(0, h'(E[X])^2\,Var(X)\right) = N\left(0, \frac{Var(X)}{g'(\theta_0)^2}\right).$$

The larger $g'(\theta_0)$ is, the smaller the asymptotic variance of $\widehat{\theta}$ is. Consider a more specific example. Suppose the density of X is $\frac{2}{\theta}x\exp\left\{-\frac{x^2}{\theta}\right\}$, $\theta > 0$, $x > 0$, that is, X follows the Weibull $(2, \theta)$ distribution;[5] then $E[X] = g(\theta) = \frac{\sqrt{\pi}}{2}\theta^{1/2}$, and $Var(X) = \theta\left(1 - \frac{\pi}{4}\right)$. So $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} N\left(0, \frac{\theta\left(1-\frac{\pi}{4}\right)}{\left(\frac{\sqrt{\pi}}{2}\frac{1}{2}\theta^{-1/2}\right)^2}\right) = N\left(0, 16\theta^2\left(\frac{1}{\pi} - \frac{1}{4}\right)\right)$. Figure 1 shows $E[X]$ and the asymptotic variance of $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ as a function of θ. Intuitively, the larger the derivative of $E[X]$ with respect to θ, the

[3] You may wonder whether any CAN estimator must be $\sqrt{n}$-consistent. This is not correct. There are CAN estimators which have different convergence rates from $\sqrt{n}$. Nevertheless, in this course all CAN estimators are $\sqrt{n}$-consistent.

[4] We use $\frac{dE[m(X|\boldsymbol{\theta}_0)]}{d\boldsymbol{\theta}'}$ instead of $E\left[\frac{dm(X|\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}'}\right]$ because $E[m(X|\boldsymbol{\theta})]$ is more smooth than $m(X|\boldsymbol{\theta})$ and can be applied to such situations as quantile estimation where $m(X|\boldsymbol{\theta})$ is not differentiable at $\boldsymbol{\theta}_0$. In this course, we will not meet such cases.

[5] (*) If the quantity X is a "time-to-failure", e.g., the time of ending the state of unemployment, the Weibull distribution gives a distribution for which the failure rate is proportional to a power of time. Since the power of x is 2 which is greater than 1, the failure rate increases with time.

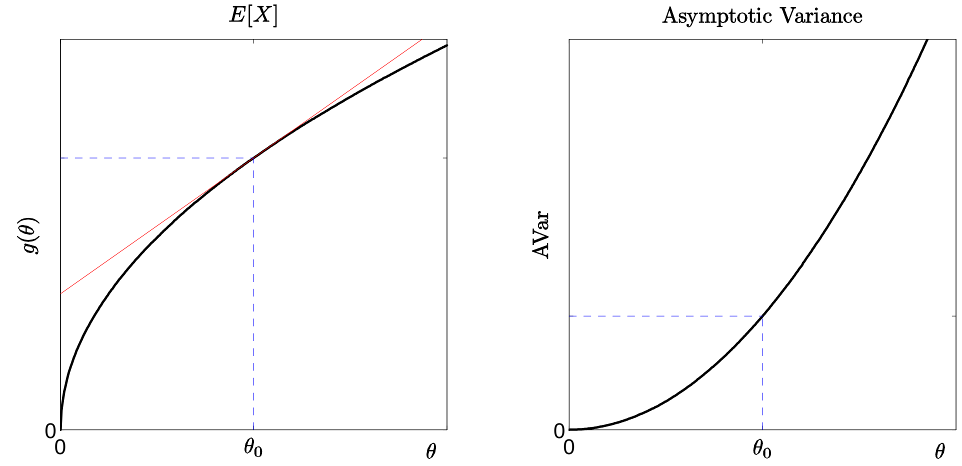

Figure 1: $E\left[X\right]$ and Asymptotic Variance as a Function of θ

easier to identify θ from $\overline{X}$, so the smaller the asymptotic variance.

Exercise 3 *(i) Show that $h'\left(E\left[X\right]\right) = 1/g'(\theta_0)$. (ii) Show that if X follows the Weibull $(2, \theta)$ distribution, then $E\left[X\right] = \frac{\sqrt{\pi}}{2}\theta^{1/2}$ and $Var\left(X\right) = \theta\left(1 - \frac{\pi}{4}\right)$.*

Generally, the asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ can be derived intuitively as follows:

$$\begin{aligned}
&\tfrac{1}{n}\sum_{i=1}^{n} m(X_i|\widehat{\boldsymbol{\theta}}) = \mathbf{0} \\
&\Longrightarrow \tfrac{1}{n}\sum_{i=1}^{n} m(X_i|\boldsymbol{\theta}_0) + \tfrac{1}{n}\sum_{i=1}^{n} \tfrac{dm(X_i|\overline{\boldsymbol{\theta}})}{d\boldsymbol{\theta}'}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \mathbf{0} \\
&\Longrightarrow \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\left(\tfrac{1}{n}\sum_{i=1}^{n} \tfrac{dm(X_i|\overline{\boldsymbol{\theta}})}{d\boldsymbol{\theta}'}\right)^{-1} \tfrac{1}{\sqrt{n}}\sum_{i=1}^{n} m(X_i|\boldsymbol{\theta}_0) \xrightarrow{d} -\mathbf{M}^{-1}N\left(\mathbf{0}, \boldsymbol{\Omega}\right),
\end{aligned}$$

where $\overline{\boldsymbol{\theta}}$ is between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, and the convergence in distribution is from Slutsky's theorem. Note that $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n} -\mathbf{M}^{-1}m(X_i|\boldsymbol{\theta}_0)$, so $-\mathbf{M}^{-1}m(X_i|\boldsymbol{\theta}_0)$ is called the **influence function**. Rigorous proof can be found in Chapter 8, where we also show that the sample analog of the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ is consistent.

Example 3 *Suppose the moment conditions are*

$$E\begin{bmatrix} X - \mu \\ \left(X - \mu\right)^2 - \sigma^2 \end{bmatrix} = 0.$$

Then the sample analog is

$$\frac{1}{n}\begin{pmatrix} \sum\limits_{i=1}^{n} X_i - n\mu \\ \sum\limits_{i=1}^{n}\left(X_i - \mu\right)^2 - n\sigma^2 \end{pmatrix} = 0,$$

so the solution is

$$\begin{aligned}\widehat{\mu} &= \overline{X}\\ \widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\overline{X^2}-\overline{X}^2.\end{aligned}$$

Now, we check the consistency and asymptotic normality of this MoM estimator.
Consistency: $\widehat{\mu}=\overline{X}\xrightarrow{p}\mu$, $\widehat{\sigma}^2=\overline{X^2}-\overline{X}^2\xrightarrow{p}\left(\mu^2+\sigma^2\right)-\mu^2=\sigma^2$.
Asymptotic Normality: $\mathbf{M}=E\left[\begin{pmatrix}-1 & 0\\ -2\left(X-\mu\right) & -1\end{pmatrix}\right]=\begin{pmatrix}-1 & 0\\ 0 & -1\end{pmatrix}$,

$$\begin{aligned}\mathbf{\Omega} &= E\left[\begin{pmatrix}\left(X-\mu\right)^2 & \left(X-\mu\right)^3-\sigma^2\left(X-\mu\right)\\ \left(X-\mu\right)^3-\sigma^2\left(X-\mu\right) & \left(X-\mu\right)^4-2\sigma^2\left(X-\mu\right)^2+\sigma^4\end{pmatrix}\right]\\ &= \begin{pmatrix}\sigma^2 & E\left[\left(X-\mu\right)^3\right]\\ E\left[\left(X-\mu\right)^3\right] & E\left[\left(X-\mu\right)^4\right]-\sigma^4\end{pmatrix},\end{aligned}$$

so

$$\sqrt{n}\begin{pmatrix}\widehat{\mu}-\mu\\ \widehat{\sigma}^2-\sigma^2\end{pmatrix}\xrightarrow{d}N\left(\mathbf{0},\mathbf{\Omega}\right).$$

Of course, we could get this asymptotic distribution in other ways. Here are two of these ways which are left as exercises. □

Exercise 4 *In Example 3,*

(i) *derive the asymptotic distribution of $(\widehat{\mu},\widehat{\sigma}^2)'$ using the Delta method. (Hint: $\left(\widehat{\mu},\widehat{\sigma}^2\right)$ is a function of $\left(\overline{X},\overline{X^2}\right)$).*

(ii) *derive the asymptotic distribution of $(\widehat{\mu},\widehat{\sigma}^2)'$ by Slutsky's theorem and noting that $\begin{pmatrix}\widehat{\mu}\\ \widehat{\sigma}^2\end{pmatrix}=\begin{pmatrix}\overline{X}\\ \frac{1}{n}\sum\limits_{i=1}^{n}\left(X_i-\mu\right)^2-\left(\overline{X}-\mu\right)^2\end{pmatrix}$. (Hint: $\sqrt{n}\left(\overline{X}-\mu\right)^2=o_p(1)$).*

Exercise 5 *In Example 3, if $X\sim N\left(\mu,\sigma^2\right)$, then what is $\mathbf{\Omega}$?*

Example 4 *Suppose we want to estimate $\theta=F(x)$ for a fixed x, where $F(\cdot)$ is the cdf of a random variable X. An intuitive estimator is the ratio of samples below x, $n^{-1}\sum_{i=1}^{n}1(X_i\leq x)$, which is called the* ***empirical distribution function*** *(EDF), while it is a MoM estimator. To see why, note that the moment condition for this problem is*

$$E\left[1(X\leq x)-F(x)\right]=0.$$

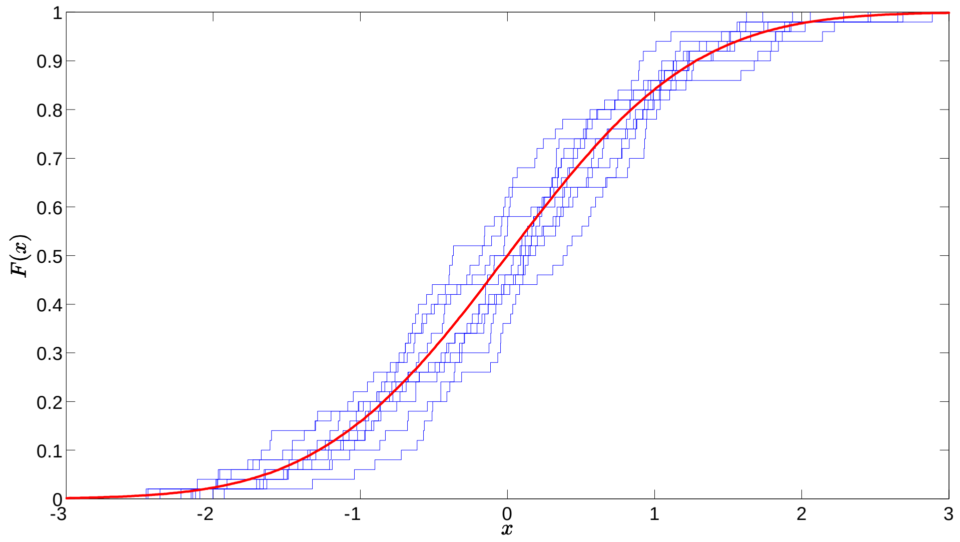

Figure 2: Empirical Distribution Functions

Its sample analog is

$$\frac{1}{n}\sum_{i=1}^{n}\left(1(X_i \le x) - F(x)\right) = 0,$$

so

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \le x).$$

By the WLLN, it is consistent. By the CLT,

$$\sqrt{n}\left(\widehat{F}(x) - F(x)\right) \xrightarrow{d} N\left(0, F(x)\left(1 - F(x)\right)\right). \textit{(why?)}$$

An interesting phenomenon is that the asymptotic variance reaches its maximum at the median of the distribution of X*. Figure 2 shows the EDF for 10 samples from* $N(0,1)$ *with sample size* $n = 50$*. Obviously, the variance of* $\widehat{F}(x)$ *is a decreasing function of* $|x|$*.*

(*) As a stochastic process indexed by $F(\cdot)$, $\sqrt{n}\left(\widehat{F}(\cdot) - F(\cdot)\right)$ *converges to a Brownian Bridge.*
□

Example 5 (*) *Recall from the Introduction that the Euler equation in macroeconomics is*

$$E_0\left[\beta\left(\frac{c_t}{c_{t+1}}\right)^{\alpha} R_{t+1}\right] = 1.$$

Suppose β is known, α is unknown; then we could use the sample analog

$$\frac{1}{T}\sum_{t=1}^{T}\left[\beta_0\left(\frac{c_t}{c_{t+1}}\right)^{\alpha}R_{t+1}\right]=1$$

to get $\widehat{\alpha}$. Although the data here are not iid, we expect the general results above about the MoM estimator holds, that is, $\widehat{\alpha}$ is CAN, and the asymptotic variance is

$$\frac{Var\left(\left(\left(\frac{c_t}{c_{t+1}}\right)^{\alpha_0}\right)R_{t+1}\right)}{\left(E\left[\left(\frac{c_t}{c_{t+1}}\right)^{\alpha_0}\ln\left(\frac{c_t}{c_{t+1}}\right)R_{t+1}\right]\right)^2}.$$

Here, note that the asymptotic variance does not depend on β_0 although α_0 depends on β_0. □

As shown in Chamberlain (1987), the MoM estimator is semiparametric efficient.

3 Asymptotics for the MLE

The section studies the asymptotic properties of the MLE. Related materials can be found in Chapter 7 of Hayashi (2000), Chapter 5 of Cameron and Trivedi (2005), Appendix D of Hansen (2007), and Chapter 13 of Wooldridge (2010).

Recall that

$$\widehat{\boldsymbol{\theta}}_{MLE}=\arg\max_{\boldsymbol{\theta}\in\Theta}\ell_n(\boldsymbol{\theta}),\tag{1}$$

where $\ell_n(\boldsymbol{\theta})=n^{-1}\sum_{i=1}^{n}\ln f\left(X_i|\boldsymbol{\theta}\right)$, and $f(\cdot|\boldsymbol{\theta})$ is the parametrized pdf (or pmf) of X.

3.1 Consistency (*)

Two main conditions for proving the consistency of $\widehat{\boldsymbol{\theta}}_{MLE}$ are as follows:

(i) $E\left[\log f(X|\boldsymbol{\theta})\right]$ is maximized uniquely at $\boldsymbol{\theta}_0$.

(ii) $\sup\limits_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum\limits_{i=1}^{n}\ln f\left(X_i|\boldsymbol{\theta}\right)-E\left[\log f(X|\boldsymbol{\theta})\right]\right|\xrightarrow{p}0.$

Intuitively, if $n^{-1}\sum_{i=1}^{n}\ln f\left(X_i|\boldsymbol{\theta}\right)$ is uniformly close to $E\left[\log f(X|\boldsymbol{\theta})\right]$, then the maximizer of $n^{-1}\sum_{i=1}^{n}\ln f\left(X_i|\boldsymbol{\theta}\right)$, $\widehat{\boldsymbol{\theta}}_{MLE}$, should be close to the maximizer of $E\left[\log f(X|\boldsymbol{\theta})\right]$, $\boldsymbol{\theta}_0$. Replacing $n^{-1}\sum_{i=1}^{n}\ln f\left(X_i|\boldsymbol{\theta}\right)$ by $-\left\|n^{-1}\sum_{i=1}^{n}m(X_i|\boldsymbol{\theta})\right\|$ and $E\left[\log f(X|\boldsymbol{\theta})\right]$ by $-\left\|E\left[m(X|\boldsymbol{\theta})\right]\right\|$, we get the proof for the consistency of the MoM estimator.

(i) means that $E\left[\log f(X|\boldsymbol{\theta})\right]\leq E\left[\log f(X)\right]$ with equality reached if and only if $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, where $f(x)=f(x|\boldsymbol{\theta}_0)$. Some literature calls this inequality as the **Kullback-Leibler information inequality**. This term comes from the fact that $\boldsymbol{\theta}_0$ minimizes the Kullback-Leibler information

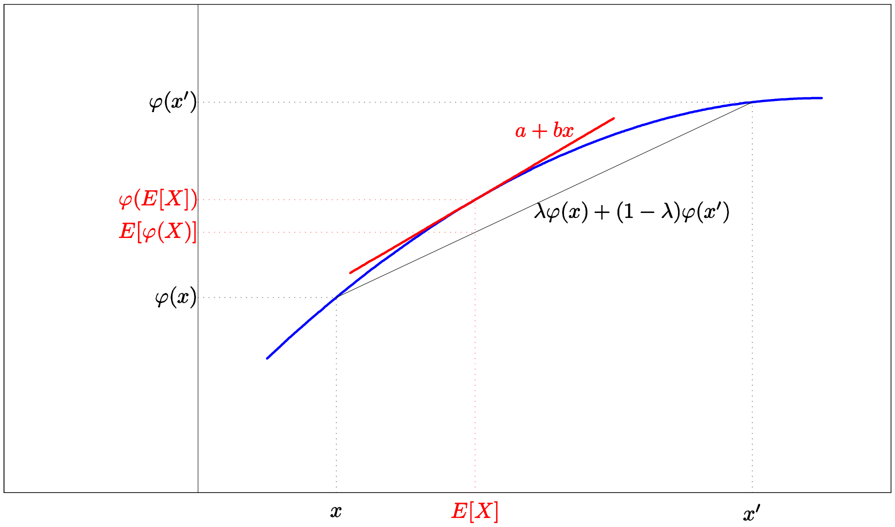

Figure 3: Intuition for Jensen's Inequality

distance between the true density $f(x)$ and the parametric density $f(x|\boldsymbol{\theta})$. Recall from the Introduction that

$$KLIC\left(\boldsymbol{\theta}\right)=\int f(x)\log\left(\frac{f(x)}{f(x|\boldsymbol{\theta})}\right)dx=\int f(x)\left(\log f(x)-\log f(x|\boldsymbol{\theta})\right)dx=E\left[\log f(X)\right]-E\left[\log f(X|\boldsymbol{\theta})\right].$$

Note that

$$-KLIC\left(\boldsymbol{\theta}\right)=E\left[\log f(X|\boldsymbol{\theta})\right]-E\left[\log f(X)\right]=\int\log\left(\frac{f(x|\boldsymbol{\theta})}{f(x)}\right)f(x)dx\leq\log\int\frac{f(x|\boldsymbol{\theta})}{f(x)}f(x)dx=\log 1=0,$$

where the equality is reached if and only if $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ by the strict Jensen's inequality, so $KLIC\left(\boldsymbol{\theta}\right)$ is minimized at $\boldsymbol{\theta}_0$, which implies $E\left[\log f(X|\boldsymbol{\theta})\right]\leq E\left[\log f(X)\right]$.

We briefly review Jensen's inequality here. It states that for a random variable X and a concave function φ, $\varphi(E[X])\geq E\left[\varphi(X)\right]$, where the equality holds if and only if, for every line $a+bx$ that is tangent to $\varphi(x)$ at $x=E[X]$, $P(\varphi(X)=a+bX)=1$. So if φ is strictly concave and X is not a constant almost surely, the strict inequality holds. Figure 3 illustrates the idea of Jensen's inequality, where X takes only two values x and x' with probability 2/3 and 1/3 respectively.

Exercise 6 *Suppose $\theta\in\mathbb{R}$, and $\widehat{\theta}$ is an estimator of θ. Show that $\sqrt{E\left[\left(\widehat{\theta}-\theta_0\right)^2\right]}\geq E\left[\left|\widehat{\theta}-\theta_0\right|\right]$, i.e., the root mean squared error (RMSE) cannot be less than the mean absolute error (MAE).*

(ii) is known as the **uniform law of large numbers**. A useful lemma to guarantee it is Mickey's Theorem (see, e.g., Theorem 2 of Jennrich (1969) or Lemma 1 of Tauchen (1985)).

Lemma 1 *If the data are i.i.d., Θ is compact, $a(w_i, \boldsymbol{\theta})$ is continuous at each $\boldsymbol{\theta} \in \Theta$ with probability one, and there is $d(w)$ with $|a(w_i, \boldsymbol{\theta})| \leq d(w)$ for all $\boldsymbol{\theta} \in \Theta$ and $E[d(w)] < \infty$, then $E[a(w, \boldsymbol{\theta})]$ is continuous and $\sup\limits_{\boldsymbol{\theta} \in \Theta} \left| n^{-1} \sum_{i=1}^{n} a(w_i, \boldsymbol{\theta}) - E\left[a(w, \boldsymbol{\theta})\right] \right| \xrightarrow{p} 0$.*

When these two conditions are satisfied, we can show the consistency of $\widehat{\boldsymbol{\theta}}_{MLE}$. Take $\delta > 0$, let $A = \{\boldsymbol{\theta} | \, \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \geq \delta\}$, we need show that $P\left(\widehat{\boldsymbol{\theta}}_{MLE} \in A\right) \to 0$ or $P\left(\widehat{\boldsymbol{\theta}}_{MLE} \in A^c\right) \to 1$. By the definition of $\widehat{\boldsymbol{\theta}}_{MLE}$,

$$\frac{1}{n} \sum_{i=1}^{n} \ln f\left(X_i | \widehat{\boldsymbol{\theta}}_{MLE}\right) \geq \frac{1}{n} \sum_{i=1}^{n} \ln f\left(X_i | \boldsymbol{\theta}_0\right).$$

From (ii), the left hand side is less than $E\left[\log f(X|\widehat{\boldsymbol{\theta}}_{MLE})\right] + \varepsilon/2$ and the right hand side is greater than $E\left[\log f(X|\boldsymbol{\theta}_0)\right] - \varepsilon/2$ for any $\varepsilon > 0$ with probability approaching 1 as n large enough. So we have

$$E\left[\log f(X|\widehat{\boldsymbol{\theta}}_{MLE})\right] + \varepsilon/2 > E\left[\log f(X|\boldsymbol{\theta}_0)\right] - \varepsilon/2$$

or

$$E\left[\log f(X|\widehat{\boldsymbol{\theta}}_{MLE})\right] > E\left[\log f(X|\boldsymbol{\theta}_0)\right] - \varepsilon.$$

From (i), choose ε as $E\left[\log f(X|\boldsymbol{\theta}_0)\right] - \sup\limits_{\boldsymbol{\theta} \in A} E\left[\log f(X|\boldsymbol{\theta})\right] > 0$,[6] and then we have $E\left[\log f(X|\widehat{\boldsymbol{\theta}}_{MLE})\right] > \sup\limits_{\boldsymbol{\theta} \in A} E\left[\log f(X|\boldsymbol{\theta})\right]$, so $\widehat{\boldsymbol{\theta}}_{MLE} \notin A$. The arguments above are intuitively illustrated in Figure 4.

In summary, to prove $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Theta} Q_n(\boldsymbol{\theta})$ is a consistent estimator of $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta} \in \Theta} Q(\boldsymbol{\theta})$, we need four conditions: (I) $Q(\boldsymbol{\theta})$ is uniquely maximized at $\boldsymbol{\theta}_0$; (II) Θ is compact; (III) $Q(\boldsymbol{\theta})$ is continuous; (IV) $Q_n(\boldsymbol{\theta})$ converges uniformly in probability to $Q(\boldsymbol{\theta})$. This is Theorem 2.1 of Newey and McFadden (1994).

3.2 Asymptotic Normality

When $f(x|\cdot)$ is smooth, the FOCs of (1) are

$$S_n(\boldsymbol{\theta}) \equiv \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{n} \sum_{i=1}^{n} s\left(X_i | \boldsymbol{\theta}\right) = 0, \tag{2}$$

where $\sum_{i=1}^{n} s\left(X_i | \boldsymbol{\theta}\right) = n S_n(\boldsymbol{\theta}) = \mathcal{S}_n(\boldsymbol{\theta}) \equiv \partial \mathcal{L}_n(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}$ is called as the **score function** or **normal equations**,[7] and $s\left(X_i | \boldsymbol{\theta}\right) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln f\left(X_i | \boldsymbol{\theta}\right) \equiv s_i\left(\boldsymbol{\theta}\right)$ is the contribution of the ith observation to the score function. So the MLE is a MoM estimator, and the moment conditions are $E\left[s\left(X | \boldsymbol{\theta}_0\right)\right] = 0$.

[6] A condition to guarantee $E\left[\log f(X|\boldsymbol{\theta}_0)\right] - \sup\limits_{\boldsymbol{\theta} \in A} E\left[\log f(X|\boldsymbol{\theta})\right] > 0$ is that $E\left[\log f(X|\boldsymbol{\theta})\right]$ is continuous and Θ is compact.

[7] The score function for $\boldsymbol{\theta}$ is usually defined as the gradient of the log-likelihood $\mathcal{L}_n(\boldsymbol{\theta}) = n l_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Some literature also loosely calls $s(x|\boldsymbol{\theta})$ as the score function.

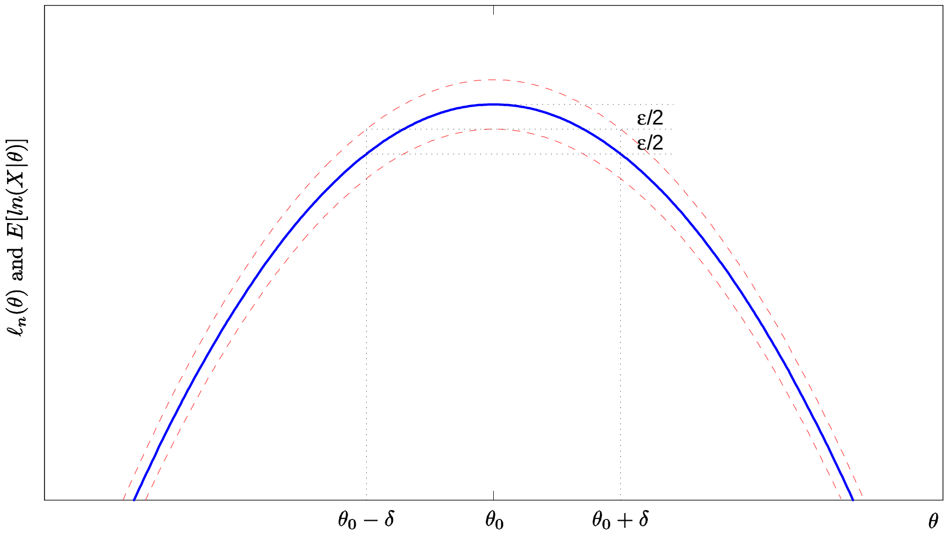

Figure 4: Illustration of the Consistency Proof in Maximum Likelihood Estimation

It is easy to check $E\left[s\left(X|\boldsymbol{\theta}_0\right)\right]=0$ by noting that

$$E\left[s\left(X|\boldsymbol{\theta}_0\right)\right]=\int\frac{\frac{\partial}{\partial\boldsymbol{\theta}}f\left(x|\boldsymbol{\theta}\right)\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{f\left(x|\boldsymbol{\theta}_0\right)}f(x|\boldsymbol{\theta}_0)dx=\int\left.\frac{\partial}{\partial\boldsymbol{\theta}}f\left(x|\boldsymbol{\theta}\right)\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}dx=\left.\frac{\partial}{\partial\boldsymbol{\theta}}\int f\left(x|\boldsymbol{\theta}\right)dx\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}=0.$$

Under some regularity conditions, $\widehat{\boldsymbol{\theta}}_{MLE}$ is $\sqrt{n}$ consistent and asymptotically normal. Specifically, the asymptotic distribution of the MLE is

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{MLE}-\boldsymbol{\theta}_0\right)\xrightarrow{d}N\left(\mathbf{0},\mathbf{H}^{-1}\mathbf{J}\mathbf{H}'^{-1}\right)=N\left(\mathbf{0},\mathbf{J}^{-1}\right),$$

where $\mathbf{H}\equiv E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ln f\left(X|\boldsymbol{\theta}_0\right)\right]$ is the Hessian, $\mathbf{J}\equiv E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln f\left(X|\boldsymbol{\theta}_0\right)\frac{\partial}{\partial\boldsymbol{\theta}'}\ln f\left(X|\boldsymbol{\theta}_0\right)\right]$ is an outer product matrix called the **information** (matrix)[8], and the equality is from the **information equality $\mathbf{J}=-\mathbf{H}$**. The information equality can be proved as follows.

$$\begin{aligned}\mathbf{H} &= E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ln f\left(X|\boldsymbol{\theta}\right)\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}=E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\frac{\partial}{\partial\boldsymbol{\theta}'}f\left(X|\boldsymbol{\theta}\right)}{f\left(X|\boldsymbol{\theta}\right)}\right)\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\\ &= E\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(X|\boldsymbol{\theta}\right)}{f\left(X|\boldsymbol{\theta}\right)}-\frac{\frac{\partial}{\partial\boldsymbol{\theta}}f\left(X|\boldsymbol{\theta}\right)\frac{\partial}{\partial\boldsymbol{\theta}'}f\left(X|\boldsymbol{\theta}\right)}{f^2\left(X|\boldsymbol{\theta}\right)}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}=E\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(X|\boldsymbol{\theta}\right)}{f\left(X|\boldsymbol{\theta}\right)}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}-\mathbf{J},\end{aligned}$$

[8] The role of the information in the asymptotic theory of maximum likelihood estimation was emphasized by R.A. Fisher (following some initial results by F.Y. Edgeworth), so sometimes the information is also called the Fisher information.

so we need only prove $E\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ln f(X|\boldsymbol{\theta})}{f(X|\boldsymbol{\theta})}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}$, which follows from

$$E\left[\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(X|\boldsymbol{\theta}\right)}{f\left(X|\boldsymbol{\theta}\right)}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \int\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(x|\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{f\left(x|\boldsymbol{\theta}_0\right)}f(x|\boldsymbol{\theta}_0)dx = \int\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(x|\boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}dx$$

$$= \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\int f\left(x|\boldsymbol{\theta}_0\right)dx\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}.$$

So the speciality of the MLE as a MoM estimator is that $\mathbf{M} = \mathbf{M}' = \mathbf{H} = -\mathbf{J} = -\boldsymbol{\Omega}$.

As mentioned in the Introduction, the MoM estimator is a semiparametric estimator and does not impose parametric restrictions on the density function, so it is more robust than the MLE. When the model is misspecified, $\boldsymbol{\theta}_0 = \arg\max_{\boldsymbol{\theta}} E\left[\log f(x|\boldsymbol{\theta})\right]$ may not satisfy $f(x|\boldsymbol{\theta}_0) = f(x)$. In this case, $\boldsymbol{\theta}_0$ is called the **quasi-true value**, and the corresponding $\widehat{\boldsymbol{\theta}}_{MLE}$ is called the **quasi-MLE** (QMLE). As shown in White (1982a), the information equality fails, but the form of the asymptotic variance $\mathbf{H}^{-1}\mathbf{J}\mathbf{H}'^{-1}$ is still valid.

We provide some intuition for the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_{MLE}$ here. To simplify our discussion, we restrict $\boldsymbol{\theta}$ to be one-dimensional.

(i) The asymptotic variance is $1/J$; that is, the larger J is, the smaller the asymptotic variance is. We know J represents the information - variation of the score function, so

More Information, Less Variance.

(ii) Since $J = -H$, J represents the curvature at θ_0. In other words, the larger the curvature of $E\left[\ln f\left(X|\theta\right)\right]$ at θ_0 is, the more accurately we can identify $\boldsymbol{\theta}_0$ by the MLE. In Figure 5, we can identify θ_0 better in (2) than (1). For (3), we cannot identify θ_0, so the asymptotic variance is infinity. (Why can we draw $E\left[\ln f\left(X|\theta\right)\right]$ like a mountain at least in the neighborhood of θ_0?)

(iii) Treat the MLE as a MoM estimator. We can see that H $(= -J)$ plays the role of M in the MoM estimator. From the discussion in Section 2, we know that the larger M is, the smaller the asymptotic variance is.

There are four methods to estimate $\mathbf{J}$:

$$\widehat{\mathbf{J}}^{(1)} = \mathbf{J}_n\left(\widehat{\boldsymbol{\theta}}_{MLE}\right) = \frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\boldsymbol{\theta}}\ln f\left(X_i|\widehat{\boldsymbol{\theta}}_{MLE}\right)\frac{\partial}{\partial\boldsymbol{\theta}'}\ln f\left(X_i|\widehat{\boldsymbol{\theta}}_{MLE}\right) = \frac{1}{n}\sum_{i=1}^n s_i\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)s_i\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)',$$

$$\widehat{\mathbf{J}}^{(2)} = -\widehat{\mathbf{H}}_n \equiv -\mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}_{MLE}\right) = -\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ln f\left(X_i|\widehat{\boldsymbol{\theta}}_{MLE}\right) = -\frac{1}{n}\mathcal{H}_n\left(\widehat{\boldsymbol{\theta}}_{MLE}\right) = -\frac{1}{n}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\mathcal{L}_n\left(\widehat{\boldsymbol{\theta}}_{MLE}\right),$$

$$\widehat{\mathbf{J}}^{(3)} = \widehat{\mathbf{J}}^{(2)}\widehat{\mathbf{J}}^{(1)-1}\widehat{\mathbf{J}}^{(2)},$$

$$\widehat{\mathbf{J}}^{(4)} = \mathbf{J}\left(\widehat{\boldsymbol{\theta}}_{MLE}\right),$$

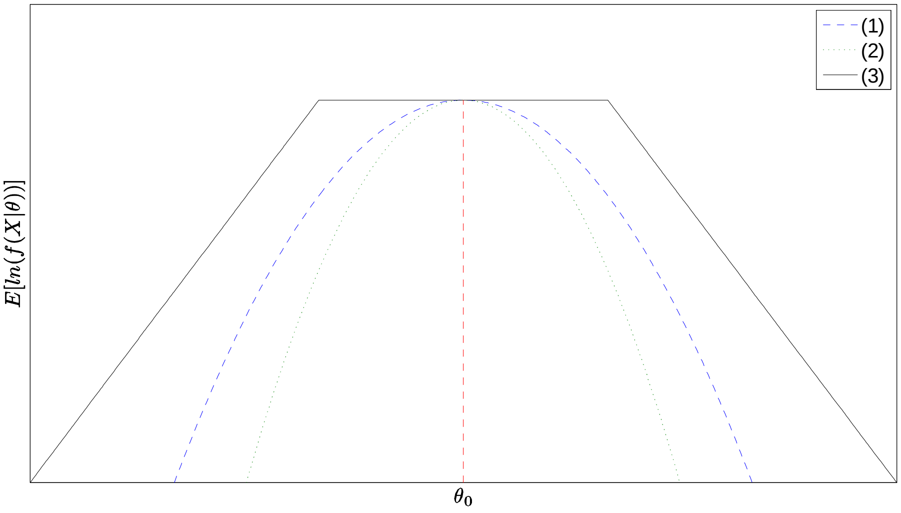

Figure 5: Identification Interpretation of J

where $\mathbf{J}(\boldsymbol{\theta}) \equiv E\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ln f(X|\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}'}\ln f(X|\boldsymbol{\theta})\right] = -E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ln f(X|\boldsymbol{\theta})\right]$. $\widehat{\mathbf{J}}^{(3)}$ is robust even in a misspecified model as shown in White (1982a); see also Gourieroux et al. (1984). It is a good simulation exercise to check which one is the best among $\widehat{\mathbf{J}}^{(i)}$, $i = 1, 2, 3, 4$, in finite samples.

Example 6 *Suppose X_i follows the the Bernoulli distribution, i.e.,*

$$L_n(\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}, \text{ and } \ell_n(\theta) = \ln\theta\frac{1}{n}\sum_{i=1}^{n} X_i + \ln(1-\theta)\frac{1}{n}\sum_{i=1}^{n}(1-X_i),$$

where X_i takes only 0 and 1. The FOC implies $\widehat{\theta}_{MLE} = \overline{X}$, which is the ratio of 1's among all samples. From the CLT,

$$\sqrt{n}\left(\overline{X}-\theta\right) \xrightarrow{d} N\left(0, J^{-1}\right) = N\left(0, \theta(1-\theta)\right),$$

so a natural estimator of J is $\frac{1}{\overline{X}\left(1-\overline{X}\right)}$. Now, we calculate $\widehat{J}^{(i)}$, $i = 1, 2, 3{,}4$, in this simple example.

(i) $\frac{\partial}{\partial\theta}\ln f(x|\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta}$, *so* $\left(\frac{\partial}{\partial\theta}\ln f(x|\theta)\right)^2 = \frac{x^2}{\theta^2} + \frac{(1-x)^2}{(1-\theta)^2}$ *(why?) and*

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

(ii) $\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$, *so*

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

(iii) $\widehat{J}^{(3)} = \widehat{J}^{(2)}\widehat{J}^{(1)-1}\widehat{J}^{(2)} = \frac{1}{\overline{X}} + \frac{1}{1-\overline{X}}$.

(iv) $J(\theta) = \frac{1}{\theta(1-\theta)}$, *so*

$$\widehat{J}^{(4)} = \frac{1}{\overline{X}\left(1-\overline{X}\right)} = \frac{1}{\overline{X}} + \frac{1}{1-\overline{X}}.$$

So in the Bernoulli case, $\widehat{J}^{(1)} = \widehat{J}^{(2)} = \widehat{J}^{(3)} = \widehat{J}^{(4)}$. *Generally, this result will not hold, and the performance of* $\widehat{J}^{(i)}$, $i = 1, 2, 3$, *depends on the specific problem in hand.* □

3.3 One-step Estimator and Algorithms (*)

When the score function (2) is nonlinear but smooth, the Newton-Raphson algorithm is usually used to find $\widehat{\boldsymbol{\theta}}_{MLE}$: iterate

$$\widehat{\boldsymbol{\theta}}_2 = \widehat{\boldsymbol{\theta}}_1 - \mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}_1\right)^{-1} S_n(\widehat{\boldsymbol{\theta}}_1)$$

until convergence.[9] The idea of the Newton-Raphson algorithm is to first linearly approximate $S_n(\boldsymbol{\theta}) = 0$ at $\widehat{\boldsymbol{\theta}}_1$ by $S_n(\boldsymbol{\theta}) = S_n(\widehat{\boldsymbol{\theta}}_1) + \mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}_1\right)\left(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_1\right) = 0$ and then solve out $\boldsymbol{\theta}$. Figure 6 illustrates the basic idea of this algorithm. To start the algorithm, we need an initial estimator. When it is consistent and $O_p(n^{-1/2})$ (e.g., the OLS estimator in the parametric linear regression), the first iteration will generate an efficient estimator which has the same asymptotic distribution as $\widehat{\boldsymbol{\theta}}_{MLE}$. To see why, Taylor expand $S_n(\widehat{\boldsymbol{\theta}}_1)$ at $\boldsymbol{\theta}_0$, we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0\right) = \left\{\mathbf{I} - \mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}_1\right)^{-1}\mathbf{H}_n(\boldsymbol{\theta}^*)\right\}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\right) - \mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}_1\right)^{-1}\sqrt{n}S_n(\boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ lies on the line segment between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_1$. Because both $\mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}_1\right)$ and $\mathbf{H}_n(\boldsymbol{\theta}^*)$ converges to $\mathbf{H}$ and $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_0\right) = O_p(1)$, the first term is $o_p(1)$. As a result, the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_0\right)$ is the same as that of $\mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}_1\right)^{-1}\sqrt{n}S_n(\boldsymbol{\theta}_0)$, so $\widehat{\boldsymbol{\theta}}_2$ is efficient.

The Newton-Raphson algorithm is based on the so-called *Gradient Theorem*: any vector, $\mathbf{d}$, in the same halfspace as $S_n(\boldsymbol{\theta})$ (that is, with $\mathbf{d}'S_n(\boldsymbol{\theta}) > 0$) is a direction of increase of $\ell_n(\boldsymbol{\theta})$, in the sense that $\ell_n(\boldsymbol{\theta} + \lambda\mathbf{d})$ is an increasing function of the scalar λ, at least for small enough λ. An algorithm that chooses a suitable direction $\mathbf{d}$, and then finds a value of λ which maximizes $\ell_n(\boldsymbol{\theta} + \lambda\mathbf{d})$ can ensure convergence. The set of directions, $\mathbf{d}$, are those with the form $\mathbf{Q}S_n(\boldsymbol{\theta})$, where $\mathbf{Q}$ is a positive definite matrix. $\mathbf{H}_n(\boldsymbol{\theta})^{-1}$ is a good choice of $\mathbf{Q}$ when $\ell_n(\boldsymbol{\theta})$ is concave, especially in

[9] A related method is the **Fisher scoring algorithm**, where $\mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}_1\right)$ is replaced by $\mathbf{H}(\widehat{\boldsymbol{\theta}}_1)$, where $\mathbf{H}(\boldsymbol{\theta}) = E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ln f(X|\boldsymbol{\theta})\right]$. In practice, these two algorithms are used with various modifications designed for speedier convergence. See Goldfeld and Quandt (1972, Ch.1) for further discussion.

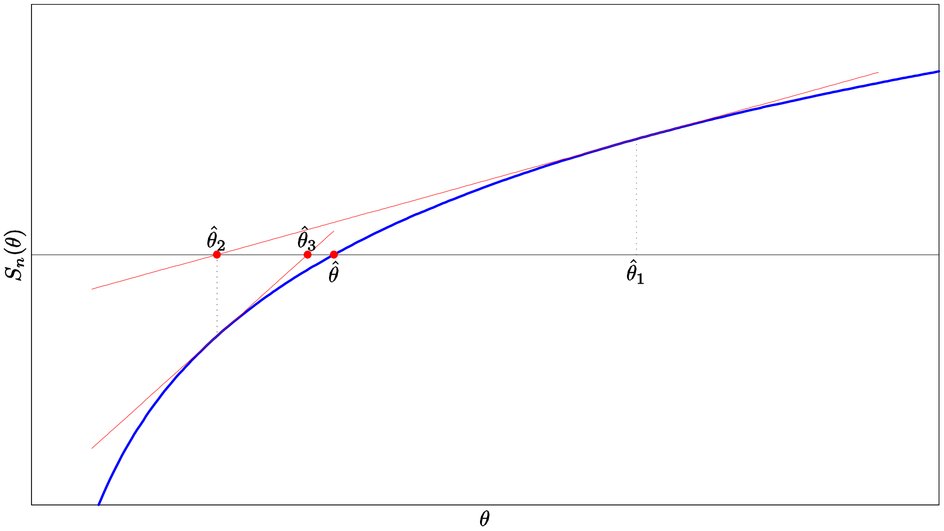

Figure 6: Intuition for the Newton-Raphson Algorithm

the neighborhood of $\boldsymbol{\theta}_0$ where the use of the Hessian makes final convergence quadratic. Even in this case, a suitable λ choice is important to guarantee convergence. Trying out decreasing λ until a higher $\ell_n(\boldsymbol{\theta}+\lambda\mathbf{d})$ is found can generate an infinite sequence of iterations that do not converge to a critical point. Choosing λ by maximizing $\ell_n(\boldsymbol{\theta}+\lambda\mathbf{d})$ can impose an unacceptable computational burden. The BHHH algorithm in Berndt et al. (1974) suggests to choose λ in the following way.[10] Let δ be a prescribed constant in the interval $(0,1/2)$. Define

$$\gamma\left(\boldsymbol{\theta},\lambda\right)=\frac{\ell_n(\boldsymbol{\theta}+\lambda\mathbf{d})-\ell_n(\boldsymbol{\theta})}{\lambda\mathbf{d}'S_n(\boldsymbol{\theta})}.$$

If $\gamma\left(\boldsymbol{\theta},\lambda\right)\geq\delta$, take $\lambda=1$. Otherwise, choose λ to satisfy $\delta\leq\gamma\left(\boldsymbol{\theta},\lambda\right)\leq1-\delta$. Such a λ always exists if $\ell_n(\boldsymbol{\theta})$ is twice continuously differentiable and has compact upper contour sets. If further assume the direction $\mathbf{d}$ satisfy $\frac{\mathbf{d}'S_n(\boldsymbol{\theta})}{S_n(\boldsymbol{\theta})'S_n(\boldsymbol{\theta})}>\alpha\in(0,1)$, then such an algorithm always converges. Berndt et al. (1974) suggest to choose $\mathbf{d}$ as $\mathbf{J}_n\left(\boldsymbol{\theta}\right)^{-1}S_n(\boldsymbol{\theta})$, where $\mathbf{J}_n\left(\boldsymbol{\theta}\right)$ is always positive definite but $\mathbf{H}_n\left(\boldsymbol{\theta}\right)$ may not be (e.g., near a saddle point). This choice of $\mathbf{Q}$ is closely related to the modified Gauss-Newton method in Hartley (1961). In summary, in the BHHH algorithm, iterate

$$\widehat{\boldsymbol{\theta}}^{(j+1)}=\widehat{\boldsymbol{\theta}}^{(j)}-\lambda^{(j)}\mathbf{J}_n\left(\widehat{\boldsymbol{\theta}}^{(j)}\right)^{-1}S_n(\widehat{\boldsymbol{\theta}}^{(j)})$$

until

$$\max_k\frac{d_k}{\max\left(1,\widehat{\theta}_k\right)}<\text{prescribed tolerance,}$$

[10] The idea of the BHHH algorithm originates from Anderson (1959).

where d_k and $\widehat{\theta}_k$ are kth component of d and $\widehat{\boldsymbol{\theta}}$, $k = 1, \cdots, d$. A corollary of this algorithm is an estimator of the asymptotic variance of the MLE, $\mathbf{J}_n\left(\widehat{\boldsymbol{\theta}}\right)$.

Extension of the Newton's method is the so-called Quasi-Newton methods. Two popular methods are due to Davidon-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS). When there are multiple local maxima or the objective function is non-differentiable, the derivative-free method such as the simplex method of Nelder-Mead (1965) is suggested. A more recent innovation is the method of simulated annealing (SA). For a review see Goffe et al. (1994).

4 Three Asymptotically Equivalent Tests

When more and more big datasets are available, asymptotic tests get popular. So we concentrate on asymptotic tests throughout this course. There are three asymptotically equivalent tests in the likelihood framework: the likelihood ratio (LR) test of Neyman and Pearson (1928), Wald (1943) test and Rao (1948)'s score test (or Lagrange Multiplier test), which are the topic of this section. Our discussion emphasizes the intuition behind these three tests and is based on Buse (1982). More related discussions can be found in Engle (1984), Section 7.4 of Hayashi (2000), Section 12.6 of Wooldridge (2010), and Section 7.3 of Cameron and Trivedi (2005).

To aid intuition, we first consider the simplest case:

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0,$$

where θ is a scalar, and then extend to the multi-parameter case with general nonlinear constraints $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$, where $\mathbf{r}(\boldsymbol{\theta})$ is a $q \times 1$ vector of constraints.

Intuitively, if $\ell_n(\widehat{\theta}_{MLE}) - \ell_n(\theta_0)$ is large, we should reject H_0. In Figure 7, we first measure the height of the mountain $\ell_n(\cdot)$ at two points, $\widehat{\theta}_{MLE}$ and θ_0, and then calculate the difference. If this difference is large, we reject H_0. This observation induces the LR test.

(i) $LR = 2n\left(\ell_n(\widehat{\theta}_{MLE}) - \ell_n(\theta_0)\right)$

The LR statistic is called *deviance* in the literature. Here, "2" is to offset the "$\frac{1}{2}$" in the second term of the Taylor expansion. To be specific, under H_0,

$$\begin{aligned} LR &= -2n\left(\ell_n(\theta_0) - \ell_n(\widehat{\theta}_{MLE})\right) \\ &= -2n\left(\frac{\partial \ell_n(\widehat{\theta}_{MLE})}{\partial \theta}\left(\theta_0 - \widehat{\theta}_{MLE}\right) + \frac{1}{2}\frac{\partial^2 \ell_n(\widehat{\theta}_{MLE})}{\partial \theta^2}\left(\theta_0 - \widehat{\theta}_{MLE}\right)^2 + o_p\left(\frac{1}{n}\right)\right) \\ &= \sqrt{n}\left(\widehat{\theta}_{MLE} - \theta_0\right)'\left(-\frac{\partial^2 \ell_n(\widehat{\theta}_{MLE})}{\partial \theta \partial \theta'}\right)\sqrt{n}\left(\widehat{\theta}_{MLE} - \theta_0\right) + o_p(1) \xrightarrow{d} \chi^2(1) \end{aligned}$$

where the third equality is from $\frac{\partial \ell_n(\widehat{\theta}_{MLE})}{\partial \theta} = 0$ which is due to the FOC of $\widehat{\theta}_{MLE}$.[11] Note that

[11] This is why we Taylor expand $l_n(\cdot)$ at $\widehat{\theta}_{MLE}$ instead of θ_0.

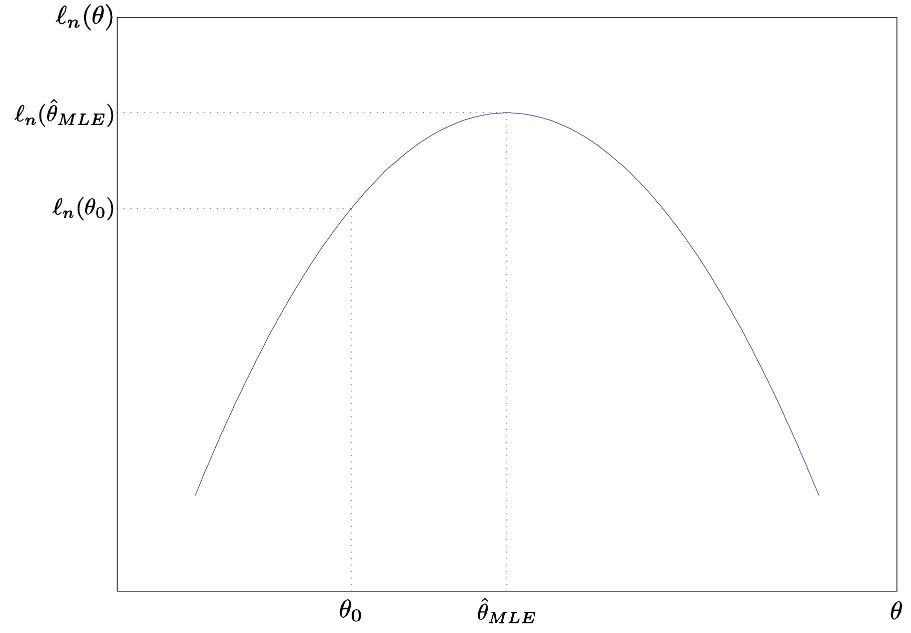

Figure 7: LR Test

the asymptotic null distribution of the LR statistic is independent of nuisance parameters. This property is called the "Wilks phenomenon" in the literature as it was first observed by Wilks (1938).

To measure the height difference in Figure 7, we have two other methods, depending on where we stand on the mountain. If we stand at $\widehat{\theta}_{MLE}$, then we get the Wald test; if we stand at θ_0, then we get the LM test.

First, suppose we stand at $\widehat{\theta}_{MLE}$, the top of the mountain. Intuitively, if the horizontal difference $\left|\widehat{\theta}_{MLE} - \theta_0\right|$ is large, the height difference between $\ell_n(\widehat{\theta}_{MLE})$ and $\ell_n(\theta_0)$ should be large, but we need take account of one more thing. Comparing the mountain A and B in Figure 8, we see that although $\left|\widehat{\theta}_{MLE} - \theta_0\right|$ is the same, $\ell_n(\widehat{\theta}_{MLE}) - \ell_n(\theta_0)$ is very different. The greater is the curvature of the mountain at $\widehat{\theta}_{MLE}$, the larger is the height difference. So we should weight $\left|\widehat{\theta}_{MLE} - \theta_0\right|$ by the curvature of $\ell_n(\theta)$ at $\widehat{\theta}_{MLE}$. This intuition induces the Wald test.

(ii) $W = n\left(\widehat{\theta}_{MLE} - \theta_0\right)^2 \left|\frac{\partial^2 \ell_n(\widehat{\theta}_{MLE})}{\partial \theta^2}\right| = \sqrt{n}\left(\widehat{\theta}_{MLE} - \theta_0\right)' \left(-H_n\left(\widehat{\theta}_{MLE}\right)\right) \sqrt{n}\left(\widehat{\theta}_{MLE} - \theta_0\right) \xrightarrow{d} \chi^2(1)$ under H_0

Second, suppose we stand at θ_0, the half-way up the mountain. Intuitively, the steeper is the mountain at θ_0, the larger is the height difference, but we still need take account of one more thing. Comparing the mountain A and B in Figure 9, we see that although the slope $S_n(\theta_0)$ is the same for these two mountains, the height difference is very different. The greater is the curvature of the mountain at θ_0, the closer is θ_0 to $\widehat{\theta}_{MLE}$ and the smaller is the height difference. So we should weight $|S_n(\theta_0)|$ by the inverse of the curvature at θ_0. This intuition induces the LM test.

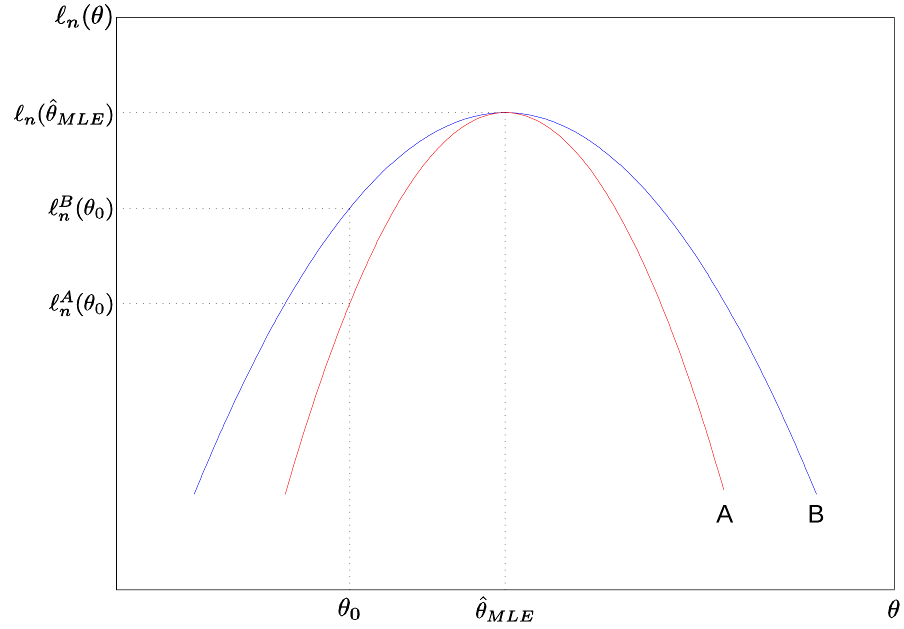

Figure 8: Wald Test

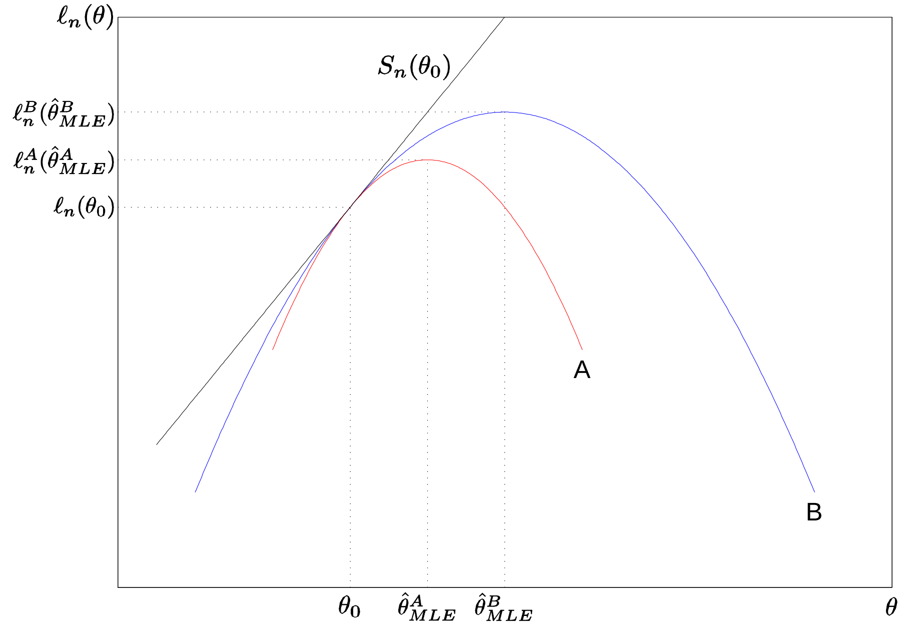

Figure 9: LM Test

(iii) $LM = \frac{nS_n(\theta_0)^2}{\left|\frac{\partial^2 \ell_n(\theta_0)}{\partial\theta^2}\right|} = \sqrt{n}S_n(\theta_0)'(-H_n(\theta_0))^{-1}\sqrt{n}S_n(\theta_0) \xrightarrow{d} \chi^2(1)$ under H_0

The name of the LM test (or the score test) is from the following fact: testing $\theta = \theta_0$ is equivalent to testing $S_n(\theta_0) = 0$, which is equivalent to test $\lambda = 0$. Here, λ is the Lagrange multiplier in the following Lagrangian problem:

$$\max_{\theta\in\Theta} \ell_n(\theta) \text{ s.t. } \theta = \theta_0.$$

If the true value of θ is θ_0, then we expect $S_n(\theta_0) = 0$ and $\lambda = 0$, since the FOC of this problem is $S_n(\theta) + \lambda = 0$.

For the general q constraints $\mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$, suppose the constrained estimator

$$\widetilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell_n(\boldsymbol{\theta}) \text{ s.t. } \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}.$$

$\widetilde{\boldsymbol{\theta}}$ can be obtained by solving the Lagrangian problem with the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \ell_n(\boldsymbol{\theta}) + \boldsymbol{\lambda}'\mathbf{r}(\boldsymbol{\theta}),$$

where $\boldsymbol{\lambda}$ is a $q \times 1$ vector of Lagrange multipliers. The FOCs for this problem are

$$\widetilde{S}_n + \widetilde{\mathbf{R}}\widetilde{\boldsymbol{\lambda}} = \mathbf{0}, \tag{3}$$

where $\widetilde{S}_n = S_n\left(\widetilde{\boldsymbol{\theta}}\right)$ and $\widetilde{\mathbf{R}}' = \partial\mathbf{r}(\widetilde{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}'$. In this general testing problem,

$$\begin{aligned} LR &= 2n\left(\ell_n\left(\widehat{\boldsymbol{\theta}}\right) - \ell_n\left(\widetilde{\boldsymbol{\theta}}\right)\right), \\ W &= n\mathbf{r}\left(\widehat{\boldsymbol{\theta}}\right)'\left[\widehat{\mathbf{R}}'\left(-\widehat{\mathbf{H}}_n\right)^{-1}\widehat{\mathbf{R}}\right]^{-1}\mathbf{r}\left(\widehat{\boldsymbol{\theta}}\right), \\ LM &= n\widetilde{S}_n'\left(-\widetilde{\mathbf{H}}_n\right)^{-1}\widetilde{S}_n = n\widetilde{\boldsymbol{\lambda}}'\widetilde{\mathbf{R}}'\left(-\widetilde{\mathbf{H}}_n\right)^{-1}\widetilde{\mathbf{R}}\widetilde{\boldsymbol{\lambda}}, \end{aligned}$$

where $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_{MLE}$, $\widehat{\mathbf{R}}' = \partial\mathbf{r}(\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}'$, $\widehat{\mathbf{H}}_n = \mathbf{H}_n\left(\widehat{\boldsymbol{\theta}}\right)$ and $\widetilde{\mathbf{H}}_n = \mathbf{H}_n\left(\widetilde{\boldsymbol{\theta}}\right)$. These test statistics can also be expressed using $\mathcal{L}_n(\cdot)$, $\mathcal{S}_n(\cdot)$, $\mathcal{H}_n(\cdot)$ and $\boldsymbol{\Lambda} = n\boldsymbol{\lambda}$ in a parallel way. For example,

$$\begin{aligned} LR &= 2\left(\mathcal{L}_n\left(\widehat{\boldsymbol{\theta}}\right) - \mathcal{L}_n\left(\widetilde{\boldsymbol{\theta}}\right)\right), \\ W &= \mathbf{r}\left(\widehat{\boldsymbol{\theta}}\right)'\left[\widehat{\mathbf{R}}'\left(-\widehat{\mathcal{H}}_n\right)^{-1}\widehat{\mathbf{R}}\right]^{-1}\mathbf{r}\left(\widehat{\boldsymbol{\theta}}\right), \\ LM &= \widetilde{\mathcal{S}}_n'\left(-\widetilde{\mathcal{H}}_n\right)^{-1}\widetilde{\mathcal{S}}_n = \widetilde{\boldsymbol{\Lambda}}'\widetilde{\mathbf{R}}'\left(-\widetilde{\mathcal{H}}_n\right)^{-1}\widetilde{\mathbf{R}}\widetilde{\boldsymbol{\Lambda}}. \end{aligned}$$

It can be shown that under H_0, all three test statistics converges in distribution to $\chi^2(q)$. We will provide more discussion on these three tests in the next chapter in the linear regression framework and in Chapter 8 in the GMM framework.

We discuss more about the LM test. This test is widely used in specification testing when the model under H_0 is much simplified;[12] see Breusch and Pagan (1980) for some of such examples. A special case is that $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$ and

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10} \text{ or } H_0 : \mathbf{r}(\boldsymbol{\theta}) = (\mathbf{I}_q, \mathbf{0}) \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} - \boldsymbol{\theta}_{10} = \mathbf{0}.$$

Partitioning S_n, $\mathbf{J}$ and $-\widetilde{\mathbf{H}}_n$ conformably,

$$\widetilde{S}_n = \begin{pmatrix} \widetilde{S}_{n1} \\ \widetilde{S}_{n2} \end{pmatrix} = \begin{pmatrix} \widetilde{S}_{n1} \\ \mathbf{0} \end{pmatrix}$$

and

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_{11} & \mathbf{J}_{12} \\ \mathbf{J}_{21} & \mathbf{J}_{22} \end{pmatrix}, -\widetilde{\mathbf{H}}_n = \begin{pmatrix} \widetilde{\mathbf{H}}_{11} & \widetilde{\mathbf{H}}_{12} \\ \widetilde{\mathbf{H}}_{21} & \widetilde{\mathbf{H}}_{22} \end{pmatrix}$$

where $\widetilde{S}_{n2} = \mathbf{0}$ is from the FOCs (3). In this case, the LM statistic will be

$$LM = \widetilde{S}_{n1}' \left(\widetilde{\mathbf{H}}_{11} - \widetilde{\mathbf{H}}_{12} \widetilde{\mathbf{H}}_{22}^{-1} \widetilde{\mathbf{H}}_{21} \right)^{-1} \widetilde{S}_{n1} \equiv \widetilde{S}_{n1}' \widetilde{\mathbf{H}}^{11} \widetilde{S}_{n1}.$$

If $\mathbf{J}_{12} = \mathbf{J}_{21}' = \mathbf{0}$, i.e., $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are "informationally" independent, then

$$LM = \widetilde{S}_{n1}' \widetilde{\mathbf{H}}_{11}^{-1} \widetilde{S}_{n1}.$$

The LM test can be conducted by running an auxiliary regression

$$1 = \widetilde{s}_i' \gamma + u_i,$$

where $\widetilde{s}_i = s_i \left(\widetilde{\boldsymbol{\theta}} \right)$, and computing

$$LM^* = nR_u^2 = \mathbf{1}'\mathbf{S}(\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{1} = nS_n \left(\widetilde{\boldsymbol{\theta}} \right)' \left(\textstyle\sum_{i=1}^n \widetilde{s}_i \widetilde{s}_i' \right)^{-1} nS_n \left(\widetilde{\boldsymbol{\theta}} \right) = \sqrt{n} S_n \left(\widetilde{\boldsymbol{\theta}} \right)' \mathbf{J}_n \left(\widetilde{\boldsymbol{\theta}} \right)^{-1} \sqrt{n} S_n \left(\widetilde{\boldsymbol{\theta}} \right), \tag{4}$$

where R_u^2 is the uncentered R^2, $\mathbf{S}$ is a $n \times k$ matrix by stacking $\widetilde{s}_i'$, and $\mathbf{1}$ is a column of ones. This version of the LM test statistic is called the **outer-product-of-the-gradient** (OPG) version since $-\mathbf{H}_n \left(\widetilde{\boldsymbol{\theta}} \right)$ in LM is substituted by $\mathbf{J}_n \left(\widetilde{\boldsymbol{\theta}} \right)$.

Although these three tests have the same asymptotic distribution under H_0 and the same asymptotic power against local alternatives, there are some differences among them.

(1) From the formulas of these test statistics, the LR test involves both $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$, the Wald test involves only $\widehat{\boldsymbol{\theta}}$, and the LM test involves only $\widetilde{\boldsymbol{\theta}}$. This provides some guidance on choosing

[12] (*) Of course, if $\widetilde{\boldsymbol{\theta}}$ is not easy to obtain due to the nonlinear constraint, the Wald test may be preferable in computation. Nevertheless, Neyman (1959) proposed what is called a $C(\alpha)$ test which involves the construction of a pseudo-LM test, with identical asymptotic properties to the true one, but requiring only $\sqrt{n}$ consistent estimators of the unknown parameters rather than ML ones. See also Buhler and Puri (1966). For a survey of literature, see Moran (1970).

among the three tests based on computational burdens.

(2) If the information matrix equality does not hold, e.g., $f(x|\boldsymbol{\theta})$ is misspecified, we can modify the LM and Wald test statistics to make them converge to $\chi^2(q)$ under H_0, but the LR test statistic does not converge to $\chi^2(q)$ any more in this case.

(3) The Wald test is not invariant to reparametrization of the null hypothesis, whereas the LR test is invariant. Some, e.g., LM^*, but not all versions of the LM test are invariant.

Also, these three test statistics may lead to conflicting inferences in small samples. Gallant (1975) conducted Monte Carlo studies in a nonlinear model to suggest that "use the likelihood ratio test when the hypothesis is an important aspect of the study".

Exercise 7 *If $f(x|\boldsymbol{\theta})$ is misspecified, how to modify the LM and Wald test statistics to make them converge to $\chi^2(q)$ under H_0? Why does the LR test statistic not converge to $\chi^2(q)$ any more in this case?*

Example 7 *To illustrate the three tests, consider an iid example with $y_i \sim N(\theta, 1)$ and test H_0: $\theta = \theta_0$ vs H_1: $\theta \neq \theta_0$. Then $\widehat{\theta} = \overline{y}$ and $\widetilde{\theta} = \theta_0$.*

For the LR test, $\ell_n(\theta) = -\frac{1}{2}\ln 2\pi - \frac{1}{2n}\sum_{i=1}^n (y_i - \theta)^2$ and some algebra yields

$$LR = 2n\left(\ell_n(\overline{y}) - \ell_n(\theta_0)\right) = n(\overline{y} - \theta_0)^2.$$

The Wald test is based on whether $\overline{y} - \theta_0 \approx 0$. Here it is easy to show that $\overline{y} - \theta_0 \sim N(0, n^{-1})$ under H_0 and $\frac{\partial^2 \ell_n(\theta_0)}{\partial \theta^2} = -1$, leading to the quadratic form

$$W = n(\overline{y} - \theta_0)[-(-1)](\overline{y} - \theta_0) = (\overline{y} - \theta_0)(n^{-1})^{-1}(\overline{y} - \theta_0).$$

This simplifies to $n(\overline{y} - \theta_0)^2$ and so $W = LR$.

The LM test is based on closeness to zero of $\partial \ell_n(\theta_0)/\partial \theta = \frac{1}{n}\sum_{i=1}^n (y_i - \theta_0) = \overline{y} - \theta_0$. It is not hard to see that $LM = n(\overline{y} - \theta_0)[-(-1)]^{-1}(\overline{y} - \theta_0) = n(\overline{y} - \theta_0)^2 = W = LR$.

Despite their quite different motivations, the three test statistics are equivalent here. This exact equivalence is special to this example with constant curvature owing to a log-likelihood quadratic in θ. Also, the three test statistics follows exact χ_1^2 under H_0. More generally the three test statistics differ in finite samples but are equivalent asymptotically. □

Exercise 8 *Suppose $y_i \overset{iid}{\sim} Bernoulli(\theta)$. Construct the three test statistics for testing H_0: $\theta = \theta_0$ vs H_1: $\theta \neq \theta_0$. Are they numerically equivalent?*

Each of the three tests has its own strength. A rule of thumb is: the Wald test would be the first to try due to its simplicity. If the Wald test is not suitable because of, e.g., nonlinear null constraints as will be discussed in Section 9 of the next chapter, the LR test is often tried. The LR test is invariant to constraint formulation and can be applied to very general cases, e.g., when the null constraints are nonlinear or even when the likelihood function is not differentiable. The LM test requires the smoothness of the likelihood function, but when the model under the null is simple, it sometimes can provide a surprisingly simplified test.

5 Normal Regression Model

In the last chapter, we studied the t test and the F test in the normal regression model. Recall that $t = \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)} \sim t_{n-k}$. It can be shown that $t_{n-k} \xrightarrow{d} N(0,1)$ (why?). We now state the three asymptotically equivalent tests and study their relationships with the F test.

The three asymptotically equivalent tests of H_0: $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$ are summarized as follows (See Hayashi (2000) p503 Ex3.):

$$
\begin{aligned}
W &= n \cdot \frac{\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)' \left[\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1} \left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right)}{SSR_U} = n \cdot \frac{SSR_R - SSR_U}{SSR_U} \xrightarrow{d} \chi^2_q, \\
LM &= n \cdot \frac{\left(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)' \mathbf{P}_{\mathbf{X}} \left(\mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}\right)}{SSR_R} = n \cdot \frac{SSR_R - SSR_U}{SSR_R} \xrightarrow{d} \chi^2_q, \\
LR &= n \cdot \left[\ln\left(\frac{SSR_R}{n}\right) - \ln\left(\frac{SSR_U}{n}\right)\right] = n \cdot \ln\left(\frac{SSR_R}{SSR_U}\right) \xrightarrow{d} \chi^2_q,
\end{aligned}
$$

where

$$
-\widehat{\mathbf{H}}_n = \begin{pmatrix} \frac{1}{\widehat{\sigma}^2}\frac{1}{n}\sum\limits_{i=}^{n}\mathbf{x}_i'\mathbf{x}_i & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\left(\widehat{\sigma}^2\right)^2} \end{pmatrix} \text{ and } -\widetilde{\mathbf{H}}_n = \begin{pmatrix} \frac{1}{\widetilde{\sigma}^2}\frac{1}{n}\sum\limits_{i=}^{n}\mathbf{x}_i'\mathbf{x}_i & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\left(\widetilde{\sigma}^2\right)^2} \end{pmatrix}
$$

are used to get neat function forms of W and LM, although other estimators of the information matrix could be used with the same asymptotic distribution. As in the F test, we can express W, LM and LR in the form of R_U^2 and R_R^2.

Exercise 9 *Verify that $-\widehat{\mathbf{H}}_n$ and $-\widetilde{\mathbf{H}}_n$, although not the same as $-\mathbf{H}_n(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)$ and $-\mathbf{H}_n(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)$, are consistent for the information matrix under the null.*

Exercise 10 *(i) Show why the second equality in (4) holds. (ii) Show that LM can be interpreted as nR_u^2, where R_u^2 is the uncentered R^2 in the regression of $\widetilde{u}_i$ on $\mathbf{x}_i$. What is the intuition for this test?*

The result in the above exercise is due to the facts that we are testing only $\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ is "informationally" independent of σ^2; see Section 3 of Breusch and Pagan (1980) for related discussions.

The relationship between F, W, LM and LR is as follows:

$$\begin{aligned} F &= \frac{n-k}{n}\frac{W}{q} < W, \\ LM &= \frac{W}{1+\frac{W}{n}}, \\ LR &= n\ln\left(1+\frac{W}{n}\right), \end{aligned}$$

so $W \geq LR \geq LM$ and

$$\begin{array}{rcc} F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k)} & \sim & F_{q,n-k} \\ & n \to \infty & \downarrow \\ = \frac{n-k}{n}\frac{W}{q} & \longrightarrow & \frac{\chi_q^2}{q} \end{array}$$

As a result, W, LR and LM have the same asymptotic null distribution, but the decision based on them may not be the same in practice. See Savin (1976), Berndt and Savin (1977), Breusch (1979), Geweke (1981) and Evans and Savin (1982) for discussions on the conflict among these three test procedures.

Because of the relationship between F and W, W is called the F **form of the Wald statistic**. Since $W \xrightarrow{d} \chi_q^2$ only under the homoskedasticity assumption, W is also called a **homoskedastic form of the Wald statistic**. Actually, all of t, W, LM and LR are asymptotically valid only under homoskedasticity. We will develop heteroskedasticity-robust versions of these test statistics after studying the asymptotics for the LSE and the RLS estimator in the next chapter. To distinguish from the homoskedasticity-only test statistics, the heteroskedasticity-robust test statistics are indexed by a subscript n.

Technical Appendix: Proofs for the Five Weapons

Proof of WLLN. Without loss of generality, assume $X \in \mathbb{R}$ and $E[X] = 0$ by considering the convergence element by element[13] and by recentering X_i at its expectation.

We need to show that $\forall \delta > 0$ and $\eta > 0$, $\exists N < \infty$ such that $\forall n < N$, $P\left(\left|\overline{X}_n\right| > \delta\right) \leq \eta$. Fix δ and η. Set $\varepsilon = \delta\eta/3$. Pick $C < \infty$ large enough so that

$$E\left[|X_i| 1\left(|X_i| > C\right)\right] \leq \varepsilon,$$

which is possible since $E\left[|X_i|\right] < \infty$. Define the random vectors

$$\begin{aligned} W_i &= X_i 1\left(|X_i| \leq C\right) - E\left[X_i 1\left(|X_i| \leq C\right)\right], \\ Z_i &= X_i 1\left(|X_i| > C\right) - E\left[X_i 1\left(|X_i| > C\right)\right]. \end{aligned}$$

[13] This is possible because $\|(x_1, y_1) - (x_2, y_2)\| \leq \|x_1 - x_2\| + \|y_1 - y_2\|$.

By the triangle inequality,

$$\begin{aligned} E\left[\left|\overline{Z}_n\right|\right] &= E\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right|\right] \leq \frac{1}{n}\sum_{i=1}^{n} E\left[\left|Z_i\right|\right] = E\left[\left|Z_i\right|\right] \\ &\leq E\left[\left|X_i\right| 1\left(\left|X_i\right| > C\right)\right] + \left|E\left[X_i 1\left(\left|X_i\right| > C\right)\right]\right| \\ &\leq 2E\left[\left|X_i\right| 1\left(\left|X_i\right| > C\right)\right] \leq 2\varepsilon. \end{aligned}$$

By Jensen's inequality,

$$\left(E\left[\left|\overline{W}_n\right|\right]\right)^2 \leq E\left[\left|\overline{W}_n\right|^2\right] = \frac{E\left[W_i^2\right]}{n} \leq \frac{4C^2}{n} \leq \varepsilon^2$$

for $n \geq N$ with $N = 4C^2/\varepsilon^2$.

Finally, by Markov's inequality,

$$P\left(\left|\overline{X}_n\right| > \delta\right) \leq \frac{E\left[\left|\overline{X}_n\right|\right]}{\delta} \leq \frac{E\left[\left|\overline{W}_n\right|\right] + E\left[\left|\overline{Z}_n\right|\right]}{\delta} \leq \frac{3\varepsilon}{\delta} = \eta,$$

as $n \geq N$. ■

Proof of CLT. We first state Lévy's continuity theorem: Let X_n and X be random vectors in $\mathbb{R}^k$. Then $X_n \xrightarrow{d} X$ if and only if $E\left[e^{it'X_n}\right] \to E\left[e^{it'X}\right]$ for every $t \in \mathbb{R}^k$, i.e., the characteristic function of X_n converges to that of X. Moreover, if $E\left[e^{it'X_n}\right]$ converges pointwise to a function $\phi(t)$ that is continuous at zero, then ϕ is the characteristic function of a random vector X and $X_n \xrightarrow{d} X$.

From Lévy's continuity theorem, we need only to show that $E\left[e^{itX_n}\right] \to e^{-\frac{1}{2}t^2}$ for every $t \in \mathbb{R}$, where $e^{-\frac{1}{2}t^2}$ is the characteristic function of $N(0,1)$. Taylor expanding $E\left[e^{itX}\right]$ at $t = 0$, we have $\phi(t) \equiv E\left[e^{itX}\right] = 1 + itE[X] - \frac{t^2}{2}E[X^2] + o(t^2) = 1 - \frac{t^2}{2} + o(t^2)$. As a result, for any fixed t,

$$\begin{aligned} E\left[e^{it\sqrt{n}\overline{X}_n}\right] &= E\left[\exp\left(i\frac{t}{\sqrt{n}}\sum_{i=1}^{n} X_i\right)\right] \\ &= \prod_{i=1}^{n} E\left[\exp\left(i\frac{t}{\sqrt{n}}X_i\right)\right] = \phi^n\left(\frac{t}{\sqrt{n}}\right) \\ &= \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \to e^{-\frac{1}{2}t^2}. \end{aligned}$$

This theorem is stated for the scalar X case. By the Cramér-Wold device, we can easily extend the result to the vector X case. The Cramér-Wold device states that for $X_n \in \mathbb{R}^k$, $X_n \xrightarrow{d} X$ if and only if $t'X_n \xrightarrow{d} t'X$ for all $t \in \mathbb{R}^k$. That is, we can reduce higher-dimensional problems tot he one-dimensional case. This device is valid because the characteristic function $t \mapsto E\left[e^{it'X}\right]$ of a vector X is determined by the set of all characteristic functions $u \mapsto E\left[e^{iu(t'X)}\right]$ of linear combinations $t'X$. ■

Proof of CMT. Denote the continuity point of X as C.

(i) Fix arbitrary $\varepsilon > 0$. For each $\delta > 0$ let B_δ be the set of x for which there exists y with $\|y - x\| < \delta$, but $\|g(y) - g(x)\| > \varepsilon$. If $X \notin B_\delta$ and $\|g(X_n) - g(X)\| > \varepsilon$, then $\|X_n - X\| \geq \delta$. Consequently,

$$P\left(\|g(X_n) - g(X)\| > \varepsilon\right) \leq P\left(X \in B_\delta\right) + P\left(\|X_n - X\| \geq \delta\right).$$

The second term on the right converges to zero as $n \to \infty$ for every fixed $\delta > 0$. Because $B_\delta \cap C \downarrow \emptyset$ by continuity of g, the first term converges to zero as $\delta \downarrow 0$.

(ii) To prove the CMT for convergence in distribution, we first state part of the Portmanteau lemma: $X_n \xrightarrow{d} X$ if and only if $\overline{\lim} P\left(X_n \in F\right) \leq P\left(X \in F\right)$ for every closed set F.

The event $\{g(X_n) \in F\}$ is identical to the event $\left\{X_n \in g^{-1}\left(F\right)\right\}$. For every closed set F,

$$g^{-1}(F) \subset \overline{g^{-1}(F)} \subset g^{-1}(F) \cup C^c.$$

To see the second inclusion, take x in the closure of $g^{-1}(F)$. Thus, there exists a sequence x_m with $x_m \to x$ and $g(x_m) \in F$. If $x \in C$, then $g(x_m) \to g(x)$, which is in F because F is closed; otherwise $x \in C^c$. By the Portmanteau lemma,

$$\overline{\lim} P\left(g\left(X_n\right) \in F\right) \leq \overline{\lim} P\left(X_n \in \overline{g^{-1}(F)}\right) \leq P\left(X \in \overline{g^{-1}(F)}\right).$$

Because $P\left(X \in C^c\right) = 0$, the probability on the right is $P\left(X \in \overline{g^{-1}(F)}\right) = P\left(g(X) \in F\right)$. Apply the Portmanteau lemma again, in the opposite direction, to conclude that $g(X_n) \xrightarrow{d} X$. ■

Proof of Slutsky's Theorem. We first state another part of the Portmanteau lemma: $X_n \xrightarrow{d} X$ if and only if $E\left[f(X_n)\right] \to E\left[f(X)\right]$ for all bunded, continuous function f.

First note that $\|(X_n, Y_n) - (X_n, c)\| = \|Y_n - c\| \xrightarrow{p} 0$. Thus by Exercise 1(i), it suffices to show that $(X_n, c) \xrightarrow{d} (X, c)$. For every continuous, bounded function $(x, y) \mapsto f(x, y)$, the function $x \mapsto f(x, c)$ is continuous and bounded. By the Portmanteau lemma, it suffices to show that $E[f(X_n, c)] \to E[f(X, c)]$ which holds if $X_n \xrightarrow{d} X$ by applying the Portmanteau lemma again. ■