Chapter 4

Efficient Likelihood Estimation and Related Tests

1 Maximum likelihood and efficient likelihood estimation

We begin with a brief discussion of *Kullback - Leibler information*.

Definition 1.1 Let P be a probability measure, and let Q be a sub-probability measure on $(\mathbb{X}, \mathcal{A})$ with densities p and q with respect to a sigma-finite measure μ ($\mu = P + Q$ always works). Thus $P(\mathbb{X}) = 1$ and $Q(\mathbb{X}) \le 1$. Then the *Kullback - Leibler information* $K(P, Q)$ is

$$(1) \qquad K(P,Q) \equiv E_P \left\{ \log \frac{p(X)}{q(X)} \right\} .$$

Lemma 1.1 For a probability measure Q and a (sub-)probability measure Q, the Kullback-Leibler information $K(P, Q)$ is always well-defined, and

$$K(P,Q) \begin{cases} \in [0, \infty] & \text{always} \\ = 0 & \text{if and only if } Q = P . \end{cases}$$

Proof. Now

$$K(P,Q) = \begin{cases} \log 1 = 0 & \text{if } P = Q , \\ \log M > 0 & \text{if } P = MQ, \quad M > 1 . \end{cases}$$

If $P \ne MQ$, then Jensen's inequality is strict and yields

$$\begin{aligned} K(P,Q) &= E_P \left(- \log \frac{q(X)}{p(X)} \right) \\ &> -\log E_P \left(\frac{q(X)}{p(X)} \right) = -\log E_Q 1_{[p(X)>0]} \\ &\ge -\log 1 = 0 . \end{aligned}$$

□

Now we need some assumptions and notation. Suppose that the model $\mathcal{P}$ is given by

$$\mathcal{P} = \{ P_\theta : \ \theta \in \Theta \} .$$

We will impose the following hypotheses about $\mathcal{P}$:

Assumptions:

A0. $\theta \neq \theta^*$ implies $P_\theta \neq P_{\theta^*}$.

A1. $A \equiv \{x : \ p_\theta(x) > 0\}$ does not depend on θ.

A2. P_θ has density p_θ with respect to the $\sigma-$finite measure μ and $X_1, \ldots, X_n$ are i.i.d. $P_{\theta_0} \equiv P_0$.

Notation:

$$
\begin{aligned}
L(\theta) &\equiv L_n(\theta) \equiv L(\theta|\underline{X}) \equiv \prod_{i=1}^n p_\theta(X_i)\,, \\
l(\theta) &= l(\theta|\underline{X}) \equiv l_n(\theta) \equiv \log L_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)\,, \\
l(B) &\equiv l(B|\underline{X}) \equiv l_n(B) = \sup_{\theta \in B} l(\theta|\underline{X})\,.
\end{aligned}
$$

Here is a preliminary result which motivates our definition of the maximum likelihood estimator.

Theorem 1.1 If A0 - A2 hold, then for $\theta \neq \theta_0$

$$
\frac{1}{n} \log \left(\frac{L_n(\theta_0)}{L_n(\theta)} \right) = \frac{1}{n} \sum_{i=1}^n \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \to_{a.s.} K(P_{\theta_0}, P_\theta) > 0\,,
$$

and hence

$$
P_{\theta_0}(L_n(\theta_0|\underline{X}) > L_n(\theta|\underline{X})) \to 1 \qquad \text{as} \qquad n \to \infty\,.
$$

Proof. The first assertion is just the strong law of large numbers; note that

$$
E_{\theta_0} \log \frac{p_{\theta_0}(X)}{p_\theta(X)} = K(P_{\theta_0}, P_\theta) > 0
$$

by lemma 1.1 and A0. The second assertion is an immediate consequence of the first. $\square$

Theorem 1.1 motivates the following definition.

Definition 1.2 The value $\widehat{\theta} = \widehat{\theta}_n$ of θ which maximizes the likelihood $L(\theta|\underline{X})$, if it exists and is unique, is the *maximum likelihood estimator* (MLE) of θ. Thus $L(\widehat{\theta}) = L(\Theta)$ or $\mathbf{l}(\widehat{\theta}_n) = \mathbf{l}(\Theta)$.

Cautions:

- $\widehat{\theta}_n$ may not exist.
- $\widehat{\theta}_n$ may exist, but may not be unique.
- Note that the definition depends on the version of the density p_θ which is selected; since this is not unique, different versions of p_θ lead to different MLE's

When $\Theta \subset R^d$, the usual approach to finding $\widehat{\theta}_n$ is to solve the *likelihood* (or *score*) equations

(2) $\quad \underline{\dot{\mathbf{l}}}(\theta|\underline{X}) \equiv \underline{\dot{\mathbf{l}}}_n(\theta) = \underline{0}\,;$

i.e. $\dot{\mathbf{l}}_{\theta_i}(\theta|\underline{X}) = 0$, $i = 1, \ldots, d$. The solution $\widetilde{\theta}_n$ say, may not be the MLE, but may yield simply a local maximum of $l(\theta)$.

The *likelihood ratio statistic* for testing $H : \theta = \theta_0$ versus $K : \theta \neq \theta_0$ is

$$\begin{aligned} \lambda_n &= \frac{L(\Theta)}{L(\theta_0)} = \frac{\sup_{\theta\in\Theta} L(\theta|\underline{X})}{L(\theta_0|\underline{X})} = \frac{L(\widehat{\theta}_n)}{L(\theta_0)}\,, \\ \widetilde{\lambda}_n &= \frac{L(\widetilde{\theta}_n)}{L(\theta_0)}\,. \end{aligned}$$

Write P_0, E_0 for P_{θ_0}, E_{θ_0}. Here are some more assumptions about the model $\mathcal{P}$ which we will use to treat these estimators and test statistics.

Assumptions, continued:

A3. Θ contains an open neighborhood $\Theta_0 \subset R^d$ of θ_0 for which:

- **(i)** For μ a.e. x, $l(\theta|x) \equiv \log p_\theta(x)$ is twice continuously differentiable in θ.
- **(ii)** For a.e. x, the third order derivatives exist and $\dddot{\mathbf{l}}_{jkl}(\theta|x)$ satisfy $|\dddot{\mathbf{l}}_{jkl}(\theta|x)| \leq M_{jkl}(x)$ for $\theta \in \Theta_0$ for all $1 \leq j, k, l \leq d$ with $E_0 M_{jkl}(X) < \infty$.

A4. **(i)** $E_0\{\dot{\mathbf{l}}_j(\theta_0|X)\} = 0$ for $j = 1, \ldots, d$.

- **(ii)** $E_0\{\dot{\mathbf{l}}_j^2(\theta_0|X)\} < \infty$ for $j = 1, \ldots, d$.
- **(iii)** $I(\theta_0) = (-\,E_0\{\ddot{\mathbf{l}}_{jk}(\theta_0|X)\})$ is positive definite.

Let

$$Z_n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\mathbf{l}}(\theta_0|X_i) \qquad \text{and} \qquad \widetilde{\mathbf{l}}(\theta_0|X) = I^{-1}(\theta_0)\dot{\mathbf{l}}(\theta_0|X)\,,$$

so that

$$I^{-1}(\theta_0) Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{\mathbf{l}}(\theta_0|X_i)\,.$$

Theorem 1.2 Suppose that $X_1, \ldots, X_n$ are i.i.d. $P_{\theta_0} \in \mathcal{P}$ with density p_{θ_0} where $\mathcal{P}$ satisfies A0 - A4. Then:

(i) With probability converging to 1 there exist solutions $\widetilde{\theta}_n$ of the likelihood equations such that $\widetilde{\theta}_n \to_p \theta_0$ when $P_0 = P_{\theta_0}$ is true.

(ii) $\widetilde{\theta}_n$ is asymptotically linear with influence function $\widetilde{\mathbf{l}}(\theta_0|x)$. That is,

$$\begin{aligned} \sqrt{n}(\widetilde{\theta}_n - \theta_0) &= I^{-1}(\theta_0) Z_n + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{\mathbf{l}}(\theta_0|X_i) + o_p(1) \\ &\to_d I^{-1}(\theta_0) Z \equiv D \sim N_d(0, I^{-1}(\theta_0))\,. \end{aligned}$$

(iii)

$$2\log\widetilde{\lambda}_n \to_d Z^T I^{-1}(\theta_0) Z = D^T I(\theta_0) D \sim \chi^2_d \,.$$

(iv)

$$W_n \equiv \sqrt{n}(\widetilde{\theta}_n - \theta_0)^T \widehat{I}_n(\widetilde{\theta}_n)\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to_d D^T I(\theta_0) D = Z^T I^{-1}(\theta_0) Z \sim \chi^2_d \,,$$

where

$$\widehat{I}_n(\widetilde{\theta}_n) = \begin{cases} I(\widetilde{\theta}_n)\,, & \text{or} \\ n^{-1}\sum_{i=1}^n \dot{\mathbf{l}}(\widetilde{\theta}_n|X_i)\dot{\mathbf{l}}(\widetilde{\theta}_n|X_i)^T\,, & \text{or} \\ -n^{-1}\sum_{i=1}^n \ddot{\mathbf{l}}(\widetilde{\theta}_n|X_i)\,. & \end{cases}$$

(v)

$$R_n \equiv Z_n^T I^{-1}(\theta_0) Z_n \to Z^T I^{-1}(\theta_0) Z \sim \chi^2_d \,.$$

Here we could replace $I(\theta_0)$ by any of the possibilities for $\widehat{I}_n(\widetilde{\theta}_n)$ given in (iv) and the conclusion continues to hold.

(vi) The model $\mathcal{P}$ satisfies the LAN condition at θ_0:

$$\begin{aligned} l(\theta_0 + n^{-1/2}t) - l(\theta_0) &= t^T Z_n - \frac{1}{2} t^T I(\theta_0) t + o_{P_0}(1) \\ &\to_d t^T Z - \frac{1}{2} t^T I(\theta_0) t \sim N(-(1/2)\sigma_0^2, \sigma_0^2) \end{aligned}$$

where $\sigma_0^2 \equiv t^T I(\theta_0) t$. Note that

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{t}_n &= \text{argmax}\{l_n(\theta_0 + n^{-1/2}t) - l_n(\theta_0)\} \\ &\to_d \text{argmax}\{t^T Z - (1/2) t^T I(\theta_0) t\} = I^{-1}(\theta_0) Z \\ &\sim N_d(0, I^{-1}(\theta_0)). \end{aligned}$$

Remark 1.1 Note that the asymptotic form of the log-likelihood given in part (vi) of theorem 1.2 is exactly the log-likelihood ratio for a normal mean model $N_d(I(\theta_0)t, I(\theta_0))$. Also note that

$$t^T Z - \frac{1}{2} t^T I(\theta_0) t = \frac{1}{2} Z^T I^{-1}(\theta_0) Z - \frac{1}{2}(t - I^{-1}(\theta_0)Z)^T I(\theta_0)(t - I^{-1}(\theta_0)Z)\,,$$

which is maximized as a function of t by $\widehat{t} = I^{-1}(\theta_0)Z$ with maximum value $Z^T I^{-1}(\theta_0) Z/2$.

Corollary 1 Suppose that A0-A4 hold and that $\nu \equiv \nu(P_\theta) = q(\theta)$ is differentiable at $\theta_0 \in \Theta$. Then $\widetilde{\nu}_n \equiv q(\widetilde{\theta}_n)$ satisfies

$$\sqrt{n}(\widetilde{\nu}_n - \nu_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\mathbf{l}}_\nu(\theta_0|X_i) + o_p(1) \to_d N(0,\, \dot{q}^T(\theta_0) I^{-1}(\theta_0)\, \dot{q}(\theta_0))\,.$$

where $\widetilde{\mathbf{l}}_\nu(\theta_0|X_i) = \dot{q}^T(\theta_0) I^{-1}(\theta_0) \dot{\mathbf{l}}(\theta_0|X_i)$ and $\nu_0 \equiv q(\theta_0)$.

If the likelihood equations (2) are difficult to solve or have multiple roots, then it is possible to use a one-step approximation. Suppose that $\overline{\theta}_n$ is a preliminary estimator of θ and set

(3) $$\check{\theta}_n \equiv \overline{\theta}_n + \widehat{I}_n^{-1}(\overline{\theta}_n)(n^{-1}\dot{\mathbf{l}}(\overline{\theta}_n|\underline{X}))\,.$$

The estimator $\check{\theta}_n$ is sometimes called a *one-step* estimator.

Theorem 1.3 Suppose that A0-A4 hold, and that $\overline{\theta}_n$ satisfies $n^{1/4}(\overline{\theta}_n - \theta_0) = o_p(1)$; note that the latter holds if $\sqrt{n}(\overline{\theta}_n - \theta_0) = O_p(1)$. Then

$$\sqrt{n}(\check{\theta}_n - \theta_0) = I^{-1}(\theta_0)Z_n + o_p(1) \to_d N_d(0, I^{-1}(\theta_0))$$

where $Z_n \equiv n^{-1/2}\sum_{i=1}^n \dot{\mathbf{l}}(\theta_0|X_i)$.

Proof. **Theorem 1.2**. (i) Existence and consistency. For $a > 0$, let

$$Q_a \equiv \{\theta \in \Theta : \ |\theta - \theta_0| = a\}\,.$$

We will show that

(a) $$P_0\{l(\theta) < l(\theta_0) \text{ for all } \theta \in Q_a\} \to 1 \qquad \text{as } n \to \infty\,.$$

This implies that L has a local maximum inside Q_a. Since the likelihood equations must be satisfied at a local maximum, it will follow that for any $a > 0$ with probability converging to 1 that the likelihood equations have a solution $\widetilde{\theta}_n(a)$ within Q_a; taking the root closest to θ_0 completes the proof.

To prove (a), write

$$\begin{aligned}\frac{1}{n}(l(\theta) - l(\theta_0)) &= \frac{1}{n}(\theta-\theta_0)^T\underline{\dot{\mathbf{l}}}(\theta_0) - \frac{1}{2}(\theta-\theta_0)^T\left(-\frac{1}{n}\ddot{\mathbf{l}}(\theta_0)\right)(\theta-\theta_0)\\ &\quad + \frac{1}{6n}\sum_{j=1}^d\sum_{k=1}^d\sum_{l=1}^d(\theta_j-\theta_{j0})(\theta_k-\theta_{k0})(\theta_l-\theta_{l0})\sum_{i=1}^n\gamma_{jkl}(X_i)M_{jkl}(X_i)\\ \text{(b)} \qquad &= S_1 + S_2 + S_3\end{aligned}$$

where, by A3(ii), $0 \le |\gamma_{jkl}(x)| \le 1$. Furthermore, by A3(ii) and A4,

(c) $$S_1 \to_p 0\,,$$

(d) $$S_2 \to_p -\frac{1}{2}(\theta-\theta_0)^T I(\theta_0)(\theta-\theta_0)\,,$$

where

(e) $$(\theta-\theta_0)^T I(\theta_0)(\theta-\theta_0) \ge \lambda_d|\theta-\theta_0|^2 = \lambda_d a^2$$

and λ_d is the smallest eigenvalue of $I(\theta_0)$ (recall that $\sup_x(x^TAx)/(x^Tx) = \lambda_1$, $\inf_x(x^TAx)/(x^Tx) = \lambda_d$ where $\lambda_1 \ge \ldots \ge \lambda_d > 0$ are the eigenvalues of A symmetric and positive definite), and

(f) $$S_3 \to_p \frac{1}{6}\sum_j\sum_k\sum_l(\theta_j-\theta_{j0})(\theta_k-\theta_{k0})(\theta_l-\theta_{l0})E\gamma_{jkl}(X_1)M_{jkl}(X_1)\,.$$

Thus for any given $\epsilon, a > 0$, for n sufficiently large with probability larger than $1-\epsilon$, for all $\theta \in Q_a$,

(g) $|S_1| < da^3\,,$

(h) $S_2 < -\lambda_d a^2/4\,,$

and

(i) $$|S_3| \le \frac{1}{3}(da)^3 \sum_{j,k,l} m_{jkl} \equiv Ba^3$$

where $m_{jkl} \equiv EM_{jkl}(X)$. Hence, combining (g), (h), and (i) yields

(j) $$\begin{aligned} \sup_{\theta\in Q_a}(S_1+S_2+S_3) &\le \sup_{\theta\in Q_a}|S_1+S_3| + \sup_{\theta\in Q_a} S_2 \\ &\le da^3 + Ba^3 - \frac{\lambda_d}{4}a^2 \\ &\le (B+d)a^3 - \frac{\lambda_d}{4}a^2 = \left\{(B+d)a - \frac{\lambda_d}{4}\right\}a^2\,. \end{aligned}$$

The right side of (j) is < 0 if $a < \lambda_d/\{4(B+d)\}$, and hence (a) holds.

On the set

(k) $$G_n \equiv \{\widetilde{\theta}_n \ \text{ solves } \ \dot{\mathbf{l}}_n(\widetilde{\theta}_n) = 0 \ \text{ and } \ |\widetilde{\theta}_n - \theta_0| < \epsilon\}$$

with $P_0(G_n) \to 1$ as $n \to \infty$, we have

(l) $$0 = \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\widetilde{\theta}_n) = \frac{1}{\sqrt{n}}\dot{\mathbf{l}}(\theta_0) - (-n^{-1}\ddot{\mathbf{l}}_n(\theta_n^*))\sqrt{n}(\widetilde{\theta}_n - \theta_0)$$

where $|\theta_n^* - \theta_0| \le |\widetilde{\theta}_n - \theta_0|$. Now from A4(i), (ii)

(m) $$Z_n \equiv \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\mathbf{l}}(\theta_0|X_i) \to_d N_d(0, I(\theta_0))\,.$$

Furthermore

(n) $$-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_n^*) = -\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0) + o_p(1) \to_p I(\theta_0)$$

by using $\widetilde{\theta}_n \to_p \theta_0$ and A3(ii) together with Taylor's theorem. Since matrix inversion is continuous (at nonsingular matrices), it follows that the inverse

(o) $$\left(-\frac{1}{n}\ddot{\mathbf{l}}(\theta_n^*)\right)^{-1}$$

exists with high probability, and satisfies

(p) $$\left(-\frac{1}{n}\ddot{\mathbf{l}}(\theta_n^*)\right)^{-1} \to_p I(\theta_0)^{-1}\,.$$

Hence we can use (l) to write, on G_n,

(q) $$\begin{aligned} \sqrt{n}(\widetilde{\theta}_n - \theta_0) &= I^{-1}(\theta_0)Z_n + o_p(1) \\ &\to_d I^{-1}(\theta_0)Z \sim N_d(0, I^{-1}(\theta_0))\,. \end{aligned}$$

This proves (ii).

It also follows from (n) that

(r) $$\sqrt{n}(\widetilde{\theta}_n - \theta_0)^T \left(-\frac{1}{n}\ddot{\mathbf{l}}(\widetilde{\theta}_n)\right) \sqrt{n}(\widetilde{\theta}_n - \theta_0) \rightarrow_d Z^T I^{-1}(\theta_0) Z \sim \chi^2_d\,,$$

and that, since $I(\theta)$ is continuous at θ_0,

(s) $$\sqrt{n}(\widetilde{\theta}_n - \theta_0)^T I(\widetilde{\theta}_n)\sqrt{n}(\widetilde{\theta}_n - \theta_0) \rightarrow_d Z^T I^{-1}(\theta_0) Z \sim \chi^2_d\,.$$

To prove (iii), we write, on the set G_n,

(t) $$l(\theta_0) = l(\widetilde{\theta}_n) + \dot{\mathbf{l}}^T(\widetilde{\theta}_n)(\theta_0 - \widetilde{\theta}_n) - \frac{1}{2}\sqrt{n}(\theta_0 - \widetilde{\theta}_n)^T \left(-\frac{1}{n}\ddot{\mathbf{l}}(\theta_n^*)\right)\sqrt{n}(\theta_0 - \widetilde{\theta}_n)$$

where $|\theta_n^* - \theta_0| \le |\widetilde{\theta}_n - \theta_0|$. Thus

$$\begin{aligned} 2\log\widetilde{\lambda}_n &= 2\{l(\widetilde{\theta}_n) - l(\theta_0)\} \\ &= 0 + 2\frac{1}{2}\sqrt{n}(\widetilde{\theta}_n - \theta_0)^T \left(-\frac{1}{n}\ddot{\mathbf{l}}(\theta_n^*)\right)\sqrt{n}(\widetilde{\theta}_n - \theta_0) \\ &= D_n^T I(\theta_0) D_n + o_p(1)\,, \qquad \text{with } \ D_n \equiv \sqrt{n}(\widetilde{\theta}_n - \theta_0) \\ &\rightarrow_d D^T I(\theta_0) D \qquad \text{where } \ D \sim N_d(0, I^{-1}(\theta_0)) \\ &\sim \chi^2_d\,. \end{aligned}$$

Finally, (v) is trivial since everything is evaluated at the fixed point θ_0. □

Proof. **Theorem 1.3**. First note that

$$\begin{aligned} \frac{1}{n}\ddot{\mathbf{l}}_n(\overline{\theta}_n) &= \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0) + \frac{1}{n}\dddot{\mathbf{l}}_n(\theta_n^*)(\overline{\theta}_n - \theta_0) \\ &= \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0) + O_p(1)|\overline{\theta}_n - \theta_0| \end{aligned}$$

so that

(a) $$\left(-\frac{1}{n}\ddot{\mathbf{l}}_n(\overline{\theta}_n)\right)^{-1} = \left(-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\right)^{-1} + O_p(1)|\overline{\theta}_n - \theta_0|$$

and

(b) $$\begin{aligned} \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\overline{\theta}_n) &= \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\theta_0) + \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\sqrt{n}(\overline{\theta}_n - \theta_0) \\ &\quad + \frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T \left(\frac{1}{n}\dddot{\mathbf{l}}_n(\theta_n^*)\right)(\overline{\theta}_n - \theta_0)\,. \end{aligned}$$

Therefore it follows that

$$\begin{aligned} \sqrt{n}(\check{\theta}_n - \theta_0) &= \sqrt{n}(\overline{\theta}_n - \theta_0) + \left(-\frac{1}{n}\ddot{\mathbf{l}}_n(\overline{\theta}_n)\right)^{-1}\frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\overline{\theta}_n) \\ &= \sqrt{n}(\overline{\theta}_n - \theta_0) \\ &\quad + \left\{\left(-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\right)^{-1} + O_p(1)|\overline{\theta}_n - \theta_0|\right\} \end{aligned}$$

$$
\begin{aligned}
& \cdot \left\{ Z_n + \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\sqrt{n}(\overline{\theta}_n - \theta_0) + \frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T \left(\frac{1}{n} \dddot{\mathbf{l}}_n (\theta_n^*) \right) (\overline{\theta}_n - \theta_0) \right\} \\
= & \left(-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0) \right)^{-1} Z_n + O_p(1)|\overline{\theta}_n - \theta_0| Z_n \\
& + O_p(1)\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\sqrt{n}|\overline{\theta}_n - \theta_0|^2 \\
& + O_p(1)\frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T \left(\frac{1}{n} \dddot{\mathbf{l}}_n (\theta_n^*) \right) (\overline{\theta}_n - \theta_0) \\
= & \ I^{-1}(\theta_0) Z_n + o_p(1) + O_p(1)\sqrt{n}|\overline{\theta}_n - \theta_0|^2 \\
= & \ I^{-1}(\theta_0) Z_n + o_p(1)\,.
\end{aligned}
$$

Here we used

$$
\begin{aligned}
& \left| \frac{1}{\sqrt{n}} \dddot{\mathbf{l}}_n (\theta_n^*)(\overline{\theta}_n - \theta_0)(\overline{\theta}_n - \theta_0) \right| \\
= & \left| \sum_{k=1}^d \sum_{l=1}^d \sqrt{n}(\overline{\theta}_{nk} - \theta_{0k})(\overline{\theta}_{nl} - \theta_{0l}) \frac{1}{n} \dddot{\mathbf{l}}_{jkl} (\theta_n^* | \underline{X}) \right| \\
\le & \ d^3 \sqrt{n} |\overline{\theta}_n - \theta_0|^2 \sum_{j=1}^d \frac{1}{n} \sum_{i=1}^n | \dddot{\mathbf{l}}_{jkl} (\theta_n^* | X_i)| \\
= & \ O_p(1)\sqrt{n}|\overline{\theta}_n - \theta_0|^2
\end{aligned}
$$

since $|\overline{\theta}_{nk} - \theta_{0k}| \le |\overline{\theta}_n - \theta_0|$ for $k = 1, \ldots, d$ and $|\underline{x}| \le d \max_{1 \le k \le d} |x_k| \le d \sum_{k=1}^d |x_k|$. $\square$

Exercise 1.1 Show that $K(P, Q) \ge 2H^2(P, Q)$.