Contents

1 M-estimation ............................................ 2
  1.1 From model to estimator ............................. 2
  1.2 Probability facts .................................... 4
    1.2.1 Law of Large Numbers ....................... 5
    1.2.2 Central Limit Theorem ....................... 5
    1.2.3 Cauchy-Schwarz inequality ................. 5
    1.2.4 Jensen’s inequality ......................... 6
  1.3 Consistency ......................................... 6
  1.4 Asymptotic normality .............................. 6
  1.5 Asymptotic efficiency .............................. 9
    1.5.1 Special case: $\tau(\theta) = \theta$ .......... 11
    1.5.2 General case .................................. 11
Chapter 1
M-estimation

1.1 From model to estimator

This handout describes a general way of constructing pretty good estimators for models where \( X = \mathbb{R}^n \) with generic point \( x = (x_1, \ldots, x_n) \) and where, under each \( P_\theta \), the \( x_1, \ldots, x_n \) are independent, each distributed according to some probability distribution \( P_\theta \) on \( \mathbb{R} \).

Note well that \( P_\theta \) is a probability distribution on the real line and \( P_\theta \) is a joint distribution. For example, if \( P_\theta \) is specified by a density \( p(z, \theta) \) on the real line then \( P_\theta \) is specified by the joint density

\[
p(x_1, \theta)p(x_2, \theta) \ldots p(x_n, \theta).
\]

I assume that the index set \( \Theta \) for the model is a subset of some Euclidean space, \( \mathbb{R}^L \), and that \( \tau : \Theta \to T \subseteq \mathbb{R}^k \) is the quantity that is to be estimated. (In class I will talk mostly about the case \( k = L = 1 \).)

**Remark.** Each \( x_i \) takes values in the real line, \( \mathbb{R} \). I could replace this \( \mathbb{R} \) by some other Euclidean space, or something even fancier, without much effect on the argument that follows.

Some authors write \( P_{\theta,n} \) instead of just \( P_\theta \), to stress that the theory concerns a sequence of models based on increasing sample sizes.

Consider a function \( g : \mathbb{R} \times T \to \mathbb{R} \). Define

\[
G(\theta, t) = \mathbb{E}_{x_1} g(x_1, t).
\]

In this handout I consider the case where \( g \) can be chosen so that \( t \mapsto G(\theta, t) \) is minimized at \( t = \tau(\theta) \), for each \( \theta \). That is,

\[
\tau(\theta) = \arg\min_{t \in T} G(\theta, t) \quad \text{for each } \theta.
\]

If you prefer, you could actually define \( \tau \) by the last equality.
To estimate $\tau(\theta)$ first define

$$G_n(t) = G_n(t, x_1, \ldots, x_n) := n^{-1} \sum_{i \leq n} g(x_i, t).$$

Suppose $G_n$ is minimized at some point $\hat{\tau}_n$ of $T$, that is,

$$\hat{\tau}_n(x_1, \ldots, x_n) = \arg\min_{t \in T} G_n(t, x_1, \ldots, x_n).$$

The random variable $\hat{\tau}_n(x_1, \ldots, x_n)$ is called an M-estimator because it is defined by a minimization operation. Some authors prefer to deal with maximization operations. The two approaches are equivalent: just replace $g$ by $-g$.

In this handout I will show, under broad assumptions, when sampling from the $P_{\theta}$ distribution the $\hat{\tau}_n$ is close to $\tau(\theta)$ with high $P_{\theta}$ probability if $n$ is large enough.

The idea is that, in the limit as the sample size goes to infinity, we learn the whole function $G(\theta, \cdot)$ exactly if we are sampling from $P_{\theta}$. A simple minimization would then give us the value $\tau(\theta)$. If we know $G(\theta, \cdot)$ only approximately (based on a finite sample size) then we can determine $\tau(\theta)$ only approximately.

**Example.** For $z$ and $t$ in $\mathbb{R}$ define $g(z, t) = |z - t| - |z|$. If $G(t) = \mathbb{E}_{P_{\theta}} g(z, t) = \int g(z, t) dP$, show that $G$ is minimized at a median of $P$. See Homework 1.

**Remark.** Here and elsewhere you should interpret $\int g(z, t) dP$ as $\sum_j g(z_j, t) P\{z_j\}$ if $P$ is a discrete distribution with atoms at $z_1, z_2, \ldots$ and as $\int g(z, t) p(z) dz$ if $P$ is a continuous distribution with density $p$.

That is, the $\tau(\theta)$ defined by equality <1> is equal to the median of the $P_{\theta}$ distribution. Implicitly, this definition assumes that the median is unique for the “population” $P_{\theta}$.

For the sample the median is not unique when $n$ is even. I intend to ignore that small detail.

**Example.** Consider the case where $T = \Theta$ for some subset $\Theta$ of $\mathbb{R}^L$. Suppose $p(z, \theta)$ is a probability density for each $\theta$ in $\Theta$. Define $g(z, t) =$
\[-\ell(z, t) \text{ where } \ell(z, \theta) = \log p(z, \theta). \text{ Write } \dot{p}(z, \theta) \text{ for } \partial p(z, \theta) / \partial \theta \text{ and } \ell(z, \theta) \text{ for } \partial \ell(z, \theta) / \partial \theta = \dot{p}(z, \theta) / p(z, \theta). \text{ Then}

\[
\frac{\partial G(\theta, t)}{\partial t} = - \int p(z, \theta) \frac{\partial}{\partial t} \ell(z, t) = \int p(z, \theta) \frac{\dot{p}(z, t)}{p(z, t)}
\]

If we put \( t \) equal to \( \theta \) then factors cancel, leaving

\[
\frac{\partial G(\theta, t)}{\partial t} \bigg|_{t=\theta} = \int \frac{\partial}{\partial \ell} p(z, t) \bigg|_{t=\theta} = \left( \frac{\partial}{\partial t} \int p(z, t) \right) \bigg|_{t=\theta} = 0.
\]

The last equality comes from the fact that \( \int p(z, t) = 1 \) for every \( t \).

That is, when \( g = -\ell \) we have \( \tau(\theta) = \theta \). The estimator \( \hat{\theta}_n \) minimizes

\[
n^{-1} \sum_{i \leq n} \log p(x_i, t) \quad \text{over all } t \in \Theta.
\]

Equivalently, it maximizes the joint density

\[
p(x_1, t) \ldots p(x_n, t) \quad \text{over all } t \in \Theta.
\]

That is, \( \hat{\theta}_n \) is the \textbf{maximum likelihood estimator} (MLE)

□

Remark. The quantity

\[
G(\theta, t) - G(\theta, \theta) = \int p(z, \theta) \log \left( \frac{p(z, \theta)}{p(z, t)} \right) dz
\]

is often called the Kullback-Leibler distance (or relative entropy) between \( P_\theta \) and \( P_t \), often denoted by \( K(P_\theta, P_t) \). It can be shown using Jensen’s inequality that \( K(P, Q) \), for probability distributions \( P \) and \( Q \), is always nonnegative. If \( P \) and \( Q \) are given by densities \( p \) and \( q \), then it can also be shown that

\[
K(P, Q) \geq \frac{1}{2} \left( \int |p - q| \right)^2,
\]

a result known as Pinsker’s inequality.

1.2 Probability facts

The asymptotic theory for M-estimators depends mostly on three important probability tools. The fourth result (Jensen’s inequality) in this Section is there just for future reference.
1.2.1 Law of Large Numbers
For independent random variables $Y_1, Y_2, \ldots$ each with the same distribution, the Law of Large Numbers (LLN) asserts that, in various probabilistic senses
\[ n^{-1} \sum_{i \leq n} Y_i \rightarrow EY_1 \quad \text{as } n \rightarrow \infty. \]

1.2.2 Central Limit Theorem
If $EY_1 = 0$ and $\sigma^2 = \text{var}(Y) < \infty$ then the Central Limit Theorem (CLT) asserts that
\[ n^{-1/2} \sum_{i \leq n} Y_i \sim N(0, \sigma^2), \]
in the sense that the distribution approaches normality as $n$ gets larger. Similar assertions hold for random vectors. For example, if the $Y_i$’s are identically distributed random vectors with zero expected value and variance matrix $V = E(Y_1Y_1')$ then
\[ n^{-1/2} \sum_{i \leq n} Y_i \sim N(0, V), \]
with $N(0, V)$ denoting a multivariate normal with zero mean and variance matrix $V$.

1.2.3 Cauchy-Schwarz inequality
If $X$ and $Y$ are random variables then
\[ |E(XY)| \leq \sqrt{E(X^2)} \sqrt{E(Y^2)} \]
Here is a quick proof, just in case you haven’t seen the inequality before. Write $a$ for $\sqrt{E(X^2)}$ and $b$ for $\sqrt{E(Y^2)}$. Then
\[ 0 \leq E \left( \frac{X}{a} \pm \frac{Y}{b} \right)^2 = 1 \pm 2E(XY)/(ab) + 1, \]
which rearranges to $2ab \geq \pm 2E(XY)$.
Notice that the Cauchy-Schwarz inequality becomes an equality if and only if either $bX = aY$ or $bX = -aY$. 

Draft: 2Sept2013  Statistics 610 ©David Pollard
1.2.4 Jensen’s inequality

If $\Psi$ is a convex function then

$$E[\Psi(X)] \geq \Psi(E[X]).$$

For the special case where $\psi(z) = z^2$ the inequality becomes $E(X^2) \geq (E[X])^2$, which is equivalent to the fact that $\text{var}(X) \geq 0$.

1.3 Consistency

The Law of Large Numbers for each fixed $t$ gives $G_n(t) \to G(\theta, t)$ as $n \to \infty$ under $P_{\theta}$. Something a little more than pointwise convergence gives

$$\tau_n = \arg\min_{t \in T} G_n(t) \to \arg\min_{t \in T} G(\theta, t) = \tau(\theta)$$

in some probabilistic sense. For example, $\{\hat{\tau}_n\}$ is said to be (weakly) consistent for $\tau(\theta)$ if

$$P_{\theta}\{|\hat{\tau}_n - \tau(\theta)| > \epsilon\} \to 0 \quad \text{for each } \epsilon > 0 \text{ and each } \theta \in \Theta.$$

1.4 Asymptotic normality

Once we know that $\hat{\theta}$ concentrates near $\tau(\theta)$ it makes sense to look at approximations to $G_n(t)$ for $t$ in a neighborhood of $\tau(\theta)$.

The analysis in this subsection is carried out for a fixed $\theta$. To avoid the temptation to differentiate with respect to a fixed $\theta$, I will temporarily omit it from the notation, writing $P$ instead of $P_{\theta}$ and $G(t)$ instead of $G(\theta, t)$ and $\tau$ instead of $\tau(\theta)$.

Remark. In fact the analysis works even if $P$ is not one of the distributions specified by the model. It matters only that the function $G(t) = E g(x_1, t) = \int g(x_1, t) \, dP$ has a unique maximum at $\tau$.

Assuming that $g$ is a smooth function of $t$, make a Taylor expansion of $g$ around $\tau$.

$$g(x_1, t) \approx g(x_1, \tau) + (t-\tau)^i g_i(x_1, \tau) + \frac{1}{2} (t-\tau)^i j g_{ij}(x_1, \tau)(t-\tau)$$

for $t$ near $\tau$.

Here $g_i(x_1, \tau)$ denotes the $k \times 1$ vector of functions $\partial g(x_1, t)/\partial t_i$, for $1 \leq i \leq k$, and $g_{ij}(x_1, \tau)$ denotes the $k \times k$ matrix of functions $\partial^2 g(x_1, t)/\partial t_i \partial t_j$, for $1 \leq i, j \leq k$, and $'$ denotes transpose.
Remark. I am using the $\bullet$ to denote differentiation instead of $'$, which gets confused with transpose. If you are not used to Taylor expansions of vector-valued functions you could just assume $k = 1$.

If we replace $x_1$ by $x_i$ then average over $1 \leq i \leq n$ approximation <4> gives

\[
G_n(t) \approx G_n(\tau) + (t - \tau)n^{-1/2}Z_n + \frac{1}{2}(t - \tau)'J_n(t - \tau)
\]

where

\[
Z_n = n^{-1/2}\sum_{i\leq n}\bullet g(x_i, \tau) \quad \text{AND} \quad J_n = n^{-1}\sum_{i\leq n}\bullet\bullet g(x_i, \tau).
\]

You’ll see soon why $Z_n$ only needs the $n^{-1/2}$ rescaling.

If we take expectations of both sides of <4>, ignoring the remainder terms we get a Taylor expansion of $G$ around $\tau$,

\[
G(t) \approx G(\tau) + (t - \tau)'\bullet\E g(x_1, \tau) + \frac{1}{2}(t - \tau)'\bullet\bullet\E g(x_1, \tau)(t - \tau)
\]

Compare with

\[
G(t) \approx G(\tau) + \frac{1}{2}(t - \tau)'\bullet\bullet G(\tau)(t - \tau).
\]

The linear term, $(t - \tau)'\bullet G(\tau)$, vanishes because $\tau$ is the minimizing value for $G$.

Remark. Here I am assuming that $\tau$ lies in the interior of $T$, so that minimization implies a zero derivative. Very strange things can happen when $\tau$ lies on the boundary of $T$.

Deduce that

\[
\E\bullet g(x_1, \tau) = \bullet G(\tau) = 0.
\]

Similarly, the $k \times k$ matrix

\[
J = \E\bullet\bullet g(x_1, \tau) = \bullet\bullet G(\tau)
\]

must be at least nonnegative definite, that is

\[
\delta' J \delta \geq 0 \quad \text{for all } \delta \in \mathbb{R}^k,
\]

otherwise there would be some direction along which $G$ decreased below its minimum.
Remark. If there is some \( \delta \neq 0 \) for which \( \delta'J\delta = 0 \) the theory gets more complicated. (The matrix \( J \) is then singular and does not have an inverse.) Textbooks seldom mention that possibility. I’ll ignore it too. That is, I’ll assume \( J \) is actually positive definite with \( \delta'J\delta > 0 \) for all nonzero \( \delta \).

The LLN (see subsection 1.2.1) applied to each of the \( k^2 \) entries of the \( k \times k \) matrix \( J_n \) from \( \text{<5> \, 1.2.1} \) gives \( J_n \approx J \), with an error of approximation that goes to zero in some suitable probabilistic sense. The \( Z_n \) is an average of random vectors with zero expected values, by \( \text{<6> \, 1.2.1} \), so the CLT gives

\[
Z_n \sim N(0, V)
\]

with \( V = \mathbb{E} g(x_1, \tau)g(x_1, \tau)' \).

Approximation \( \text{<5> \, 1.2.1} \) can be rewritten as

\[
G_n(t) \approx G_n(\tau) + (t - \tau)'n^{-1/2}Z_n + \frac{1}{2} (t - \tau)'J(t - \tau)
\]

for \( t \) near \( \tau \).

The quadratic on the right-hand side can be simplified by writing \( s \) for \( t - \tau \) and \( W \) for \( n^{-1/2}Z_n \), leaving

\[
\frac{1}{2} s'Js + s'W = \frac{1}{2} (s + J^{-1}W)'J(s + J^{-1}W) - \frac{1}{2} W'J^{-1}W
\]

on the right-hand side. Positive definiteness of \( J \) ensures that the quadratic is minimized when \( s = -J^{-1}W \). Assuming that the approximation of \( G_n \) translates into approximation of its minimizer, conclude that \( \hat{\tau}_n - \tau \approx -n^{-1/2}J^{-1}Z_n \), so that

\[
\text{<8> \, 1.2.1} \quad n^{1/2} (\hat{\tau}_n - \tau) \approx -J^{-1}Z_n \sim N(0, J^{-1}VJ^{-1})
\]

with

\[
J = \mathbb{E} g(x_1, \tau)g(x_1, \tau)' \AND V = \mathbb{E} g(x_1, \tau)g(x_1, \tau)'.
\]

Example. Consider once more the maximum likelihood estimator from Example \( \text{<3> \, 1.2.1} \), where \( T = \Theta \subseteq \mathbb{R}^L \) and \( g(z, t) = -\ell(z, t) = -\log p(z, t) \). Remember that \( \tau(\theta) = \theta \) and that I was writing \( \hat{\theta}_n \) instead of \( \hat{\tau}_n \). Remember also (cf. equality \( \text{<6> \, 1.2.1} \)) that \( \mathbb{E} g(\theta) = 0 \).
For this special case, the $L \times L$ variance matrix $V(\theta)$ equals

$$E g(x_1, \tau) g(x_1, \tau)' = \text{var}(\ell(x_1, \tau)) =: \mathbb{I}(\theta).$$

The matrix $\mathbb{I}(\theta)$ is called the **Fisher information matrix**.

The $J(\theta) = E_{\theta} \ell'(x_1, \theta)$ also takes a special form. First note that

$$\dddot{\ell}(z, t) = \frac{\partial \ddot{\ell}(z, t)}{\partial t} = \frac{\partial \dot{p}_t(z, t)}{\partial t} = \dot{p}_t(z, t) \dddot{p}_t(z, t)' / p(z, t)^2.$$

Take expectations.

$$E_{\theta} \dddot{\ell}(x_1, \theta) = \int \frac{\partial^2 p(z, t)}{\partial t^2} |_{t=\theta} - \int p(z, \theta) \ddot{\ell}(z, \theta) \ell(z, \theta)' .$$

The first term vanishes—takes the second derivative outside the integral sign then use the fact that $\int p(z, t) = 1$ for all $t$. The second term is just the information matrix again. In summary, $V(\theta) = -J(\theta) = \mathbb{I}(\theta)$. Approximation $\langle 8 \rangle$ becomes

\begin{align*}
\text{MLE:} \quad & Z_n = -n^{-1/2} \sum_{i \leq n} \ell(x_i, \theta) \sim N(0, \mathbb{I}(\theta)) \text{ and} \\
& n^{1/2} \left( \hat{\theta}_n - \theta \right) \approx -\mathbb{I}(\theta)^{-1} Z_n \sim N(0, \mathbb{I}(\theta)^{-1}) \quad \text{under } P_\theta, \\
\end{align*}

a very famous approximation.

### 1.5 Asymptotic efficiency

Classical (Fisherian) statistical theory assigns the MLE a very special role. According to Fisher, the MLE minimizes the limiting variance amongst all estimators. He called this property (asymptotic) **efficiency**. Modern theory has shown Fisher’s assertion to be wrong unless hedged with further restrictions. One such restriction is to consider only M-estimators as competitors.

**Remark.** Actually Fisher usually didn’t include the word “asymptotic”.

I find that this omission leads to a lot of confusion between optimality properties for fixed, finite sample size and optimality properties for the limiting distribution.
For this section suppose $P_\theta$ is specified by some density function $p(z, \theta)$ on the real line. (The argument for discrete distributions is similar.)

Once again consider the problem of estimating $\tau(\theta) = (\tau_1(\theta), \ldots, \tau_k(\theta))$, based on independent samples $x_1, x_2, \ldots$ from $P_\theta$. Assume $\tau$ is a smooth function of $\theta$, so that the $k \times L$ matrix $\tau'(\theta)$ with $\tau_{ij}(\theta) = \partial \tau_i / \partial \theta_j$ is well defined.

For each fixed $\delta$ in $\mathbb{R}^k$, approximation $<8>$ implies

$$n^{1/2} \delta' (\hat{\tau}_n - \tau(\theta)) \sim N(0, \sigma^2(\theta))$$

under $P_\theta$ where

$$\sigma^2_\delta(\theta) := \delta' J(\theta)^{-1} \nu(\theta) J(\theta)^{-1} \delta$$

$$J(\theta) = \mathbb{E}_\theta \ell(x_1, \tau(\theta))$$

$$V(\theta) = \mathbb{E}_\theta \ell(x_1, \tau(\theta)) g(x_1, \tau(\theta))'$$

The aim is to find a $g$ to minimize the asymptotic variance $\sigma^2_\delta(\theta)$.

The first step is to find a lower bound for $\sigma^2(\theta)$ by considering the behaviour of $\hat{\tau}_n$ for values of $\theta$ of the form $\theta_s = \theta + s \gamma$ for $s \in \mathbb{R}$.

Here $\theta$ denotes some point of $\Theta$ that is fixed throughout the argument and $\gamma$ is a nonzero vector in $\mathbb{R}^L$. Notice that

$$\frac{\partial}{\partial s} \tau(\theta_s) = \tau'(\theta_s) \gamma,$$

the product of a $k \times L$ matrix with an $L \times 1$ vector.

Along the path defined by $\theta_s$ inequality $<6>$ becomes

$$0 = \mathbb{E}_{\theta_s} \ell(x_1, \tau(\theta_s)) = \int \ell(x_1, \tau(\theta_s)) p(z, \theta_s) \quad \text{for all } s.$$

Notice the way $\theta_s$ appears in two places. Differentiate with respect to $s$.

$$0 = \int \ell(x_1, \tau(\theta_s)) \tau'(\theta_s) \gamma p(z, \theta_s) + \int \ell(x_1, \tau(\theta_s)) p(z, \theta_s) \gamma$$

$$= \mathbb{E}_{\theta_s} \ell(x_1, \tau(\theta_s)) \tau'(\theta_s) \gamma + \mathbb{E}_{\theta_s} \ell(x_1, \tau(\theta_s)) \ell(z, \theta_s)' \gamma$$

Put $s = 0$ then multiply both sides by $\delta' J(\theta)^{-1}$ to deduce

$$\delta' \tau'(\theta) \gamma = -\mathbb{E}_{\theta} \left( \delta' J(\theta)^{-1} \ell(x_1, \tau(\theta)) \ell(z, \theta)' \gamma \right).$$
The right-hand side is of the form \( E_\theta(XY) \) for random variables

\[
X = \delta'J(\theta)^{-1}g(x, \tau(\theta)) \quad \text{AND} \quad Y = \ell(z, \theta)'\gamma.
\]

Notice that

\[
E_\theta X^2 = \sigma^2(\theta) \quad \text{AND} \quad E_\theta Y^2 = \gamma'\Pi(\theta)\gamma.
\]

Invoke the Cauchy-Schwarz inequality (subsection 1.2.3) to deduce that

\[
(\delta'\tau(\theta)\gamma)^2 \leq \sigma^2(\theta) \gamma'\Pi(\theta)\gamma
\]

That is,

\[
\sigma^2(\theta) \geq \frac{(\delta'\tau(\theta)\gamma)^2}{\gamma'\Pi(\theta)\gamma} \quad \text{for all } \gamma \neq 0.
\]

Replace \( \gamma \) by \( \Pi(\theta)^{-1/2}\beta \) then choose \( \beta \) as the unit vector in the direction \( \delta'\tau(\theta)\Pi(\theta)^{-1/2} \) to maximize the right-hand side, leaving the lower bound

\[
\delta'J(\theta)^{-1}V(\theta)J(\theta)^{-1}\delta = \sigma^2(\theta) \geq \lambda_\delta(\theta) := \delta'\tau(\theta)\Pi^{-1}(\theta)\tau(\theta)'\delta.
\]

### 1.5.1 Special case: \( \tau(\theta) = \theta \)

When \( \tau(\theta) = \theta \) the matrix \( \tau(\theta) \) of derivatives becomes the identity matrix \( I_L \) and the right-hand side of \( <12> \) becomes \( \delta'\Pi(\theta)^{-1}\delta \). For the special case of the MLE \( \hat{\theta}_n \) the left-hand side of \( <12> \) also equals \( \delta'\Pi(\theta)^{-1}\delta \). That is, the MLE achieves the lower bound, for every \( \delta \).

### 1.5.2 General case

When \( \tau(\theta) \) is not the identity matrix it is not obvious to me how to find an M-estimator that achieves the lower bound in \( <12> \). But there is another way to estimate \( \tau(\theta) \) that does give an asymptotic variance equal to \( \lambda_\delta(\theta) \). By Taylor’s Theorem,

\[
\tau(\theta + h) \approx \tau(\theta) + \tau(\theta)h \quad \text{for small } h \in \mathbb{R}^L.
\]
In particular, for \( h = \hat{\theta}_n - \theta \) the approximation \(<10>\) for the MLE \( \hat{\theta}_n \) gives
\[
\tau(\hat{\theta}_n) - \tau(\theta) \approx -\tau(\theta)n^{-1/2}\mathbb{I}(\theta)^{-1}Z_n
\]
where \( Z_n = -n^{-1/2} \sum_{i \leq n} \ell'(x_i, \theta) \sim N(0, \mathbb{I}(\theta)) \). Thus
\[
n^{1/2} \left( \tau(\hat{\theta}_n) - \tau(\theta) \right) \approx \tau(\theta)\mathbb{I}(\theta)^{-1}Z_n \sim N(0, W(\theta)) \quad \text{under } \mathbb{P}_{\theta},
\]
where
\[
W(\theta) = \tau(\theta)\mathbb{I}(\theta)^{-1}\mathbb{I}(\theta)\mathbb{I}(\theta)^{-1}\tau(\theta)'.
\]
In consequence, \( n^{1/2} \delta' \left( \tau(\hat{\theta}_n) - \tau(\theta) \right) \sim N(0, \lambda_\delta(\theta)) \).