THE LOWER BOUND FOR MSE IN STATISTICAL PREDICTION THEORY

Yoichi Miyata*

This paper is concerned with an inequality for MSE in statistical prediction theory. Takeuchi (1975) provided the inequality for a risk of unbiased predictor under certain regularity conditions. We shall provide an inequality for MSE of an unbiased predictor from $L^2$-differentiability of densities point of view. In addition, this inequality is simplified and corresponded to the above under slightly stronger conditions. We shall also state the criterion for $L^2$-differentiability in the case that an observable random vector and a predictive random variable are not independent.

Key words and phrases: unbiased predictor, Cramér-Rao inequality, differentiability in quadratic mean.

1. Introduction

Suppose that $\{X, \mathcal{A}, \mu\}$ and $\{Y, \mathcal{B}, \nu\}$ are two measure spaces. Let $\mathcal{A}$ and $\mathcal{B}$ be $\sigma$-algebras of subsets of $X$ and $Y$, and $\mu$ and $\nu$ be $\sigma$-finite measures on $X$ and $Y$, respectively. $\{X \times Y, \mathcal{A} \times \mathcal{B}, \mu \times \nu\}$ is the cartesian product space of $\{X, \mathcal{A}, \mu\}$ and $\{Y, \mathcal{B}, \nu\}$. Let $(P^\theta_{XY})$ be probability distributions on $X \times Y$ with densities $p^\theta(x, y)$ relative to a $\sigma$-finite measure $\mu \times \nu$, and $(P^\theta_{X})$ a marginal distribution for a random vector $X$. The parameter space $\Theta$ is an open subset of $\mathbb{R}^d$. Let $X = (X_1, \ldots, X_n)$ be an observable random vector taking values in $X$, and $Y$ an unknown random variable taking values in $Y$. Then the problem of predicting the value of $Y$ based on $X$ is called a statistical prediction problem, as described in Takeuchi (1975). A predictor $T(X)$ is said to be unbiased if $E^\theta(T(X) - Y) = 0$ for every $\theta \in \Theta$. We denote the conditional expectation of $Y$ given $X$ by $g(X : \theta) := E^\theta(Y | X)$, the transpose of a matrix $M$ by $M'$.

The following inequality gives a lower bound for MSE of an unbiased predictor. This is a version of the Cramér-Rao inequality to the prediction problem. For the proof, see Takeuchi (1975) p. 16 and Ishii (1978a) pp. 72–75.

Theorem 1.1. Under suitable regularity conditions

$$E^\theta(Y - T(X))^2 \geq E^\theta(Y - g(X : \theta))^2 + G(\theta)'I^{-1}(\theta)G(\theta),$$

for any unbiased predictor $T$.


*Waseda University Honjo Senior High School, 1136 Okuboyama, Nishitomita, Honjo-shi Saitama, 367-0035, Japan.
where $G(\theta) = E_\theta\left(\frac{\partial}{\partial \theta} g(X; \theta)\right)$ and $I(\theta) = \int \left(\frac{\partial}{\partial \theta} \log p_\theta(x)\right)\left(\frac{\partial}{\partial \theta} \log p_\theta(x)\right)' dP^X_\theta$.

On the other hand, some probability distributions do not satisfy such regularity conditions since the conditions of Theorem 1.1 need differentiability of densities and $g(X; \theta)$ with respect to $\theta$, having common support of densities with respect to $\theta$ and so on. For example, Double Exponential Distribution $X_1, \ldots, X_n, Y \overset{i.i.d.}{\sim} q_\theta(\cdot) = \frac{1}{2}\exp[-|x-\theta|]$ does not satisfy usual regularity condition. But it is well known that the family $(\sqrt{q_\theta})$ is continuous $L^2$-differentiable. The purpose of this paper is to provide an inequality for MSE of unbiased predictor from $L^2$-differentiability point of view. The main results we will prove in Section 3 are as follows. (Theorem 3.1 and 3.2):

Suppose that Fisher’s information matrix is nonsingular. Then, a lower bound for MSE of any unbiased predictor is provided when the family $(\sqrt{p_\theta(x)})$ is $L^2$-differentiable and $E_\theta(Y)$ is ordinary partial differentiable with respect to $\theta$. Furthermore, under more stringent conditions, this bound could be simplified.

In section 4, we will prove that $L^2$-differentiability of $(\sqrt{p_\theta(x, y)})$ is equivalent to that of $(\sqrt{p_\theta(x)})$ and $(\sqrt{p_\theta(y|x)})$ (Theorem 4.1 and 4.2). Section 5 deals with the proofs of Section 3 and 4.

2. Definition

First, we shall introduce some concepts. We denote the square root of densities by $s_\theta(x) := \sqrt{p_\theta(x)}$, an inner product by $(\cdot, \cdot)$, the Euclidean norm by $|\cdot|$ and the space of functions with $\int |f|^k d\mu < \infty$ by $L^k(\mu)$.

**Definition 2.1.** A parametrization $\theta \to P^X_\theta$ is $L^2$-differentiable if for every $\theta \in \Theta$, there exists a function $\hat{s}_\theta(x) \in L^2(\mu)$ such that

$$\int |s_{\theta+h}(x) - s_\theta(x) - (\hat{s}_\theta(x), h)|^2 d\mu = o(|h|^2).$$

If $\theta$ is a multivariate parameter point, we consider $\hat{s}_\theta(x)$ as a row vector. We sometimes denote that the family $(\sqrt{p_\theta(x)})$ is $L^2$-differentiable when a parametrization $\theta \to P^X_\theta$ is $L^2$-differentiable. We shall use the same notation $\hat{s}_\theta(\cdot)$ for the $L^2$-derivative as well as for the $L^2$-gradient; i.e. the row vector of $L^2$-partial derivatives $\hat{s}_{\theta_i}(\cdot) (i = 1, \ldots, n)$. Define Fisher’s information matrix of $\theta$ with respect to $\mu$ by

$$I(\theta) = 4 \int \hat{s}_\theta(x)\hat{s}_\theta(x)' d\mu \quad (2.1)$$

and the score function $\hat{l}_\theta(x)$ by
\[ \dot{l}_\theta(x) = 2 \frac{s_\theta(x)}{s_\theta(x)} 1_{[s_\theta(x) > 0]} = \frac{\dot{p}_\theta(x)}{p_\theta(x)} 1_{[p_\theta(x) > 0]} . \]

Putting \( I(\theta) = \int \dot{l}_\theta(x) \dot{l}_\theta(x)' dP^X_\theta \), it coincides with (2.1) if \( \sqrt{p_\theta(x)} \) is \( L^2 \)-differentiable. See Bickel, Klaassen, Ritov and Wellner (1993) p. 13. Likewise, we shall define \( L^2 \)-differentibility and Fisher’s information matrix for probability measures \( (P^X_{\theta}) \). Secondly, we consider \( L^2 \)-differentiability for the conditional probability density of \( Y \) given \( X \). This basic concept is stated by Kuboki (1987). Define

\[ p_\theta(y \mid x) := \begin{cases} \frac{p_\theta(x,y)}{p_\theta(x)} & \text{where } p_\theta(x) > 0, \\ 0 & \text{elsewhere} \end{cases} \]

and the square root of conditional density of \( Y \) by \( s_\theta(y \mid x) := \sqrt{p_\theta(y \mid x)} \).

**Definition 2.2.** A parametrization \( \theta \rightarrow P_\theta(Y \mid X) \) is \( L^2 \)-differentiable in measure \( P^X_\theta \times \nu \) if for every \( \theta \in \Theta \), there exists a function \( \dot{s}_\theta(y \mid x) \in L^2(P^X_\theta \times \nu) \) such that

\[
\int \int |s_{\theta + h}(y \mid x) - s_\theta(y \mid x) - (\dot{s}_\theta(y \mid x), h)|^2 dP^X_\theta d\nu = o(|h|^2).
\]

To provide the inequality for MSE, we consider the following conditions.

**Definition 2.3.** The family \( (P^X_{\theta}) \) satisfies conditions (i) if the family \( (\sqrt{p_\theta(x)}) \) is partial \( L^2 \)-differentiable, \( E_\theta(Y) \) is ordinary partially differentiable with respect to \( \theta \) and Fisher’s information matrix is nonsingular.

**Definition 2.4.** The family \( (P^X_{\theta}) \) satisfies conditions (ii) if the family \( (\sqrt{p_\theta(x,y)}) \) is partial \( L^2 \)-differentiable, Fisher’s information matrix is nonsingular, and \( E_{\theta_0}g(X : \theta)^2 \) is bounded for every \( \theta \) in some neighborhood of every fixed point \( \theta_0 \in \Theta \).

**Remark 2.5.** Conditions (ii) imply condition (i).

**Proof.** \( L^2 \)-differentiability of \( (\sqrt{p_\theta(x)}) \) is proved from that of induced probability measure (Apply Bickel, Klaassen, Ritov and Wellner (1993), Appendix A.5, Proposition 5). Partial differentiability of \( E_\theta(Y) \) is verified by using Lemma 5.3. □
3. Main results

**Theorem 3.1.** Suppose that $T(X)$ is any unbiased predictor with $E_\theta T(X)^2 < \infty$ and $E_\theta(Y^2) < \infty$. Then, under conditions (i), the following inequality for MSE of $T(X)$ holds:

$$E_\theta(Y - T(X))^2 \geq E_\theta(Y - g(X : \theta))^2 + C(\theta)'I(\theta)^{-1}C(\theta),$$

where $C(\theta) = E_\theta[g(X : \theta)\hat{l}_\theta(X)] - \frac{\partial}{\partial \theta}E_\theta(Y)$.  

**Theorem 3.2.** Suppose that $T(X)$ is any unbiased predictor with $E_\theta T(X)^2 < \infty$ and $E_\theta(Y^2) < \infty$. Under conditions (ii),

$$E_\theta(Y - T(X))^2 \geq E_\theta(Y - g(X : \theta))^2 + E_\theta\hat{g}(X : \theta)'I^{-1}(\theta)E_\theta\hat{g}(X : \theta),$$

where $\hat{g}(X : \theta) = (\hat{g}_1(X : \theta), \ldots, \hat{g}_d(X : \theta))'$ and $\hat{g}_i(X : \theta)$ is the $L^1$-derivative of $g(X : \theta)$ with respect to each $\theta_i$. Of course, if $g(X : \theta)$ is ordinary partially differentiable, then the derivative coincides with $\hat{g}(X : \theta)$.

**Remark 3.3.** If $X = (X_1, \ldots, X_n)$ and $Y$ are independent then $C(\theta) = -\frac{\partial}{\partial \theta}E_\theta(Y)$.

In fact, using Remark 5.4, we have $E_\theta[g(X : \theta)\hat{l}_\theta(X)] = E_\theta(Y)E_\theta[\hat{l}_\theta(X)] = 0$. Hence $C(\theta) = -\frac{\partial}{\partial \theta}E_\theta(Y)$.

**Theorem 3.4.** The equality in (3.1) or (3.2) holds if and only if

$$T_\theta(X) = g(X : \theta) - C(\theta)'I^{-1}(\theta)\hat{l}_\theta(X)$$

for $P_\theta^X$ - almost all $x$.

**Proof.** This assertion is proved from the same argument given in the proof in Ishii (1978b) pp. 69–70.

Note that the function $T_\theta$ specified in (3.3) may depend on $\theta$. In such a case, $T_\theta$ is not a predictor and there is no predictor that attains the lower bound. Conversely, if $T_\theta$ is an unbiased predictor and does not depend on $\theta$ then it attains the lower bound.

**Example 1.** (Double Exponential Distribution)

$$X_1, \ldots, X_n, Y \overset{i.i.d.}{\sim} p(\cdot; \theta) = \frac{1}{2}\exp[-|x - \theta|] \quad \theta \in \mathbb{R},$$

where $\theta$ is unknown. It is well-known that this model is continuous $L^2$-differentiable from Hájek conditions (see Bickel, Klaassen, Ritov, Wellner (1993), p. 460, Proposition 4, Corollary 1). By the $L^1$-derivative $\hat{p}_\theta(x) = \text{sgn}(x - \theta)\frac{1}{2}\exp[-|x - \theta|]$, Fisher’s information of $X_i$ is $I_{X_i}(\theta) = 1$. Since
random variables $X_i$ ($i = 1, \ldots, n$) are mutually independent, we have $I_X(\theta) = n$, $C(\theta) = -1$. Hence we obtain the inequality

$$E_\theta(Y - T(X))^2 \geq 2 + \frac{1}{n}.$$  

For an unbiased predictor $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, we have

$$E_\theta(Y - \bar{X})^2 = 2 + \frac{2}{n}.$$  

Therefore, MSE (mean square error) between $\bar{X}$ and $Y$ attains the lower bound asymptotically. Note that the sample median $\hat{Y}_0$ has asymptotically smaller MSE than that of the sample mean:

$$E_\theta(Y - \hat{Y}_0)^2 = 2 + \frac{1}{n} \left(1 + \frac{2\sqrt{2}}{\sqrt{\pi} n} + o \left(\frac{1}{\sqrt{n}}\right)\right).$$


**Example 2.** (Gamma Distribution)

$$X_1, \ldots, X_n, Y \overset{i.i.d.}{\sim} p(\cdot; \gamma, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\cdot - \gamma)^{\alpha-1} e^{-\beta(\cdot - \gamma)} \quad (\cdot > \gamma),$$

where $\beta > 0$, $-\infty < \gamma < +\infty$ are unknown and $\alpha$ is known, i.e., $\theta = (\beta, \gamma) \in \Theta$. In view of Lemma 5.10, restricting $\alpha > 2$, this model is continuous $L^2$-differentiable. Let $X = (X_1, \ldots, X_n)$. Since $I_{X_i}(\theta) = \left(\frac{\frac{\alpha}{\beta^2}}{-1} \frac{-1}{\beta^2} \right)$ and $X_i$ ($i = 1, \ldots, n$) are independent, Fisher’s information matrix of $X$ is $I_X(\theta) = n \left(\frac{\frac{\alpha}{\beta^2}}{-1} \frac{-1}{\beta^2} \right)$. Hence, we have $I_X^{-1}(\theta) = \left(\frac{\frac{\alpha}{\beta^2}}{2n} \frac{\frac{\alpha}{\beta^2}}{1} \frac{1}{\beta^2}\right)$ and $C(\theta) = \left(\frac{\frac{\alpha}{\beta^2}}{-1}\right)$. Finally, we obtain the following inequality for MSE.

$$E_\theta(T(X) - Y)^2 \geq E_\theta(Y - g(X : \theta))^2 + C(\theta)' I_X^{-1}(\theta) C(\theta)$$

$$= \left(1 + \frac{1}{n}\right) \frac{\alpha}{\beta^2}.$$  

Subsequently, we consider the statistics of Theorem 3.4. Since $g(X : \theta) = \gamma + \frac{\alpha}{\beta}$ and $C(\theta)' I_X^{-1}/\theta(X) = \gamma + \frac{\alpha}{\beta} - \bar{X}$, we have the predictor $T_1(X) = \bar{X}$. This predictor $T_1$ is unbiased and does not depend on $\theta$. Hence MSE of $T_1(X)$ and $Y$ coincides with this lower bound. Actually,

$$E_\theta(Y - \bar{X})^2 = E_\theta(Y - E_\theta(Y))^2 + E_\theta(\bar{X} - E_\theta(Y))^2$$

$$= \left(1 + \frac{1}{n}\right) \frac{\alpha}{\beta^2}.$$
Example 3. (AR(1)) Suppose that
\[ U_{i+1} = \rho U_i + \epsilon_i, \]
where \( U_1, \epsilon_i \quad (i = 1, \ldots, n) \) are i.i.d. according to Double Exponential Distribution with the expected value 0 and the known standard deviation \( \sigma \), and \( |\rho| < 1 \) is unknown. Let \( U_1, \ldots, U_n \) be observations and \( U_{n+1} \) be an unknown random variable. It holds from Example 5 that the family \( (\sqrt{p_\rho(u_1, \ldots, u_{n+1}))} \) is \( L^2 \)-differentiable.

First, we consider the inequality for this model. Since \( \frac{\partial}{\partial \rho} E_\rho(U_{n+1} \mid U_n) = U_n \), we have \( C(\theta) = E\{ \frac{\partial}{\partial \rho} E_\rho(U_{n+1} \mid U_n) \} = 0 \). Hence it follows from Theorem 3.2 that for any unbiased predictor \( \hat{U} \),
\[ E_\rho(U_{n+1} - \hat{U})^2 \geq \sigma^2. \]

Second, we consider the following estimator for \( \rho \):
\[ \hat{\rho} = \frac{\sum_{i=1}^{n-1} U_{i+1}U_i}{\sum_{i=1}^{n} U_i^2}, \]
which is called Yule Waker estimator. It is known that this estimator is bounded; \( |\hat{\rho}| \leq 1 \) (see Nakatuka (1978) p. 247).

It follows that \( \hat{\rho} - \rho = -\rho \frac{U_n^2}{\sum_{i=1}^{n} U_i^2} + \frac{1}{n} \sum_{i=1}^{n} U_i^2 \to \sigma^2 \frac{1}{1-\rho^2} \) and \( \frac{1}{n} \sum_{i=1}^{n} U_i \epsilon_i \to 0 \), we have \( \hat{\rho} \overset{P}{\to} \rho \). From the boundedness of \( \hat{\rho} \), we have
\[ E_\rho|\hat{\rho} - \rho|^m \to 0 \quad (n \to \infty) \quad \text{for every} \quad m \geq 1. \]

See T. Nakatuka (1978) p. 247. Now we consider a predictor \( \hat{U}_1 = \hat{\rho}U_n \).
It is verified that \( \hat{U}_1 \) is an unbiased predictor since \( \hat{U}_1 \) is an odd function. Applying the Cauchy-Schwarz inequality and (3.4), we have the following
\[ E_\rho[U_{n+1} - \hat{\rho}U_n]^2 = E_\rho[U_{n+1} - \rho U_n]^2 + E_\rho(\hat{\rho} - \rho)^2 U_n^2 \]
\[ \leq \sigma^2 + \sqrt{E_\rho(\hat{\rho} - \rho)^4} \sqrt{E_\rho U_n^4} \]
\[ = \sigma^2 + o(1). \]

Example 4. (Linear model)
\[ Y_i = \alpha + \beta X_i + \epsilon_i \quad (i = 1, \ldots, n, n+1), \]
where \( \epsilon_i(i = 1, \ldots, n+1) \) i.i.d. Double Exponential Distribution with the mean 0 and the known standard deviation \( \sigma \), i.e., \( f(\cdot) = \frac{1}{\sqrt{2\sigma}} \exp[-\frac{\sqrt{2}}{\sigma} \cdot \|] \), and \( X_i \) i.i.d. \( N(\mu, 1) \). Suppose that \( \epsilon_i \) and \( X_i(i = 1, \ldots, n) \) are mutually
independent, \( \sigma \) is known, and \( \alpha, \beta, \mu \) are unknown. Now we consider the problem predicting \( Y_{n+1} \) based on \( (X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1} \).

Let \( X := \{(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}\} \), \( Y := Y_{n+1} \) and the joint density of \((X, Y)\) be \( p_\theta(x, y) \). It follows from Theorem 4.2 that the family \((\sqrt{p_\theta(x_i, y_i)})\) is \( L^2 \)-differentiable since \((\sqrt{p_\theta(y_i \mid x_i)})\) is \( L^2 \)-differentiable in measure \( P_\theta^X \times \nu \) and \((\sqrt{p_\theta(x_i)})\) is \( L^2 \)-differentiable. Hence this model satisfies the conditions of Theorem 3.2. First, we establish the lower bound for \( H \). Since \( F \) satisfies the conditions of Theorem 3.2. First, we establish the lower bound for \( H \).

Similarly, we have

\[
E_\theta(Y - T(X))^2 \geq \left(1 + \frac{1}{2n}\right) \sigma^2.
\]

Here we consider an unbiased predictor of \( Y_{n+1} \). Since the conditional expectation of \( Y \) given \( X \) is \( g(X : \theta) = \alpha + \beta X_{n+1} \), we have a predictor \( \hat{Y} = \hat{\alpha} + \hat{\beta} X_{n+1} \), substituting LSE \( \hat{\alpha}, \hat{\beta} \) for \( \alpha, \beta \) based on \((X_i, Y_i)(i = 1, \ldots, n)\). Note that \( \hat{Y} \) is unbiased since \( \hat{\alpha} \) and \( \hat{\beta} \) are unbiased estimators of \( \alpha \) and \( \beta \) respectively. Let \( \bar{X} := (X_1, \ldots, X_n, X_{n+1}) \) and \( S_{XX} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 \).

\[
E_\theta(Y - \hat{Y})^2 = E_\theta(Y - g(X : \theta))^2 + E_\theta(g(X : \theta) - \hat{Y})^2 = \sigma^2 + V_\theta(\hat{\alpha}) + 2E_\theta(X_{n+1})Cov(\hat{\alpha}, \hat{\beta}) + E_\theta(X_{n+1}^2) V_\theta(\hat{\beta}).
\]

Applying the conditional variance of \( \hat{\alpha} \) given \( \bar{X} \), we have

\[
V_\theta(\hat{\alpha}) = E_\theta^\bar{X}\{E_\theta(\hat{\alpha} - \alpha)^2 \mid \bar{X}\} = E_\theta^\bar{X}\left\{\sigma^2 \left(\frac{1}{n} + \frac{\bar{X}}{S_{XX}}\right)\right\} = \frac{\sigma^2}{n-3}\left(1 - \frac{2}{n} + \mu^2\right).
\]

Similarly, we have

\[
V_\theta(\hat{\beta}) = \frac{\sigma^2}{n-3}
\]

\[
Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \mu}{n-3}.
\]

Finally, MSE of \( \hat{Y} \) is as follows. Hence, it attains the lower bound asymptotically.

\[
E_\theta(Y - \hat{Y})^2 = \sigma^2 + \frac{2\sigma^2(n-1)}{n(n-3)} \left(1 + \frac{1}{2n}\right) \sigma^2 + O\left(\frac{1}{n}\right).
\]
4. Properties of $L^2$-differentiability

In this section, we describe some properties of $L^2$-differentiability. We use the same notations as in the previous section. It is sometimes not an easy matter to check $L^2$-differentiability for the model where an unknown random variable $Y$ and an observation $X$ are not independent. Here we shall state and prove the useful Theorems.

**Theorem 4.1.** Suppose that the family $(\sqrt{p_{\theta}(x, y)})$ is $L^2$-differentiable. Then the family $(\sqrt{p_{\theta}(y \mid x)})$ is $L^2$-differentiable in measure $P_{\theta}^X \times \nu$ with the $L^2$-derivative

$$
\dot{s}_\theta(y \mid x) = \left( \frac{\dot{s}_\theta(x, y)}{s_\theta(x)} - \frac{s_\theta(x, y)}{s_\theta(x)^2} \dot{s}_\theta(x) \right) 1_{[p_\theta(x) > 0]}.
$$

**Theorem 4.2.** Suppose that the family $(\sqrt{p_{\theta}(y \mid x)})$ is $L^2$-differentiable in measure $P_{\theta}^X \times \nu$ and $(\sqrt{p_{\theta}(x)})$ is $L^2$-differentiable. Then the family $(\sqrt{p_{\theta}(x, y)})$ is $L^2$-differentiable with the $L^2$-derivative

$$
(4.1) \quad \dot{s}_\theta(x, y) = \dot{s}_\theta(y \mid x) \sqrt{p_\theta(x)} + \sqrt{p_\theta(y \mid x)} \dot{s}_\theta(x).
$$

In some cases, it is inconvenient to check $L^2$-differentiability of $(\sqrt{p_{\theta}(y \mid x)})$ in measure $P_{\theta}^X \times \nu$ even if $(\sqrt{p_{\theta}(y \mid x)})$ is $L^2$-differentiable in measure $\nu$. The following theorem gives the sufficient condition of $L^2$-differentiability in measure $P_{\theta}^X \times \nu$. This result is stated by Strasser (1998) Theorem 3.4, p. 120. Suppose that for every $x \in \mathcal{X}$, the family $(\sqrt{p_{\theta}(y \mid x)})$ is $L^2$-differentiable in measure $\nu$, i.e.,

$$
\int |s_{\theta+h}(y \mid x) - s_\theta(y \mid x) - (\dot{s}_\theta(y \mid x), h)|^2 d\nu = o(|h|^2) \quad \text{for every } \theta.
$$

Note that we shall use the same notation $\dot{s}_\theta(y \mid x)$ for the $L^2$-derivative in measure $P_{\theta}^X \times \nu$ as well as for the $L^2$-derivative in measure $\nu$. For the following theorem, we consider continuous $L^2$-differentiability instead of mere $L^2$-differentiability. We denote Fisher’s information matrix with respect to $\nu$ by $I_Y(x, \theta) := 4 \int \dot{s}_\theta(y \mid x) \dot{s}_\theta(y \mid x) d\nu$. Let $I_Y(x, h, \theta) := I_Y(x, \theta + h)$. $I_Y(x, h, \theta)$ is called to be uniformly $P_{\theta}^X$-integrable if $\lim_{M \to \infty} \sup_h \int_{I_Y(x, h, \theta) > M} |I_Y(x, h, \theta)| dP_{\theta}^X = 0$ for every $\theta \in \Theta$. Of course, $I_Y(x, h, \theta)$ is uniformly $P_{\theta}^X$-integrable if there exists a function $H_h(x : \theta)$ such that $I_Y(x, h, \theta) \leq H_h(x : \theta)$ a.e. $P_{\theta}^X$ and $\int H_h(x : \theta) dP_{\theta}^X \to \int H(x : \theta) dP_{\theta}^X$ as $|h| \to 0$.

**Theorem 4.3.** For every $x \in \mathcal{X}$, let the family $(\sqrt{p_{\theta}(y \mid x)})$ is continuous $L^2$-differentiable in measure $\nu$ with Fisher’s Information matrix.
\( I_{Y|x}(\theta) \) with respect to \( \nu \). If the family of functions \( I_{Y|x,h}(\theta) \) is uniformly \( P_\theta^X \)-integrable then the family \( (\sqrt{p_\theta(y|x)}) \) is continuous \( L^2 \)-differentiable in measure \( \nu \times P_\theta^X \).

**Example 5.** (AR(1))

\[
U_{i+1} = \rho U_i + \varepsilon_i \quad i = 1, \ldots, n,
\]

where \(|\rho| < 1\) is unknown and \( U_1, \varepsilon_i \) \( i = 1, \ldots, n \) are i.i.d. with common known density \( f(\cdot) \). \( U_{i+1} | U_i \) denotes the conditional random variable of \( U_{i+1} \) given \( U_i \). Suppose that the family \( (\sqrt{f(u_{i+1} - \rho u_i)}) \) is continuous \( L^2 \)-differentiable in Lebesgue measure \( \mu_{i+1} \) and Fisher’s information matrix \( I_{U_{i+1}U_i}(\theta) \) with respect to \( \mu_{i+1} \) is uniformly \( P_{U_{i+1}}^\rho \)-integrable. Since \( U_2 | U_1 \) has densities \( f(\cdot - \rho u_1) \) which is \( L^2 \)-differentiable in measure \( \mu_2 \times P_\rho^U \). The densities of \( U_1 \) are also \( L^2 \)-differentiable. By applying Theorem 4.2, the joint densities of \( (U_2, U_1) \) are \( L^2 \)-differentiable in measure \( \mu_2 \times \mu_1 \). Repeating this process inductively, we have that the joint densities of \( (U_1, \ldots, U_{n+1}) \) are \( L^2 \)-differentiable.

5. **Proof.**

**Lemma 5.1.** Suppose that \( M \) is a \((p + q) \times (p + q)\) nonnegative symmetric matrix

\[
M = \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix},
\]

where \( M_{11} \) is \( p \times p \), \( M_{22} \) is \( q \times q \), and \( M_{22} \) has an inverse matrix. Then the following inequality holds.

\[
M_{11} \geq M_{12}M_{22}^{-1}M_{21}.
\]

**Proof of Lemma 5.1.** Putting \( P = \begin{pmatrix} E_p & -M_{12}M_{22}^{-1} \\ 0 & M_{22}^{-1} \end{pmatrix} \), where \( E_p \) is a \( p \times p \) unit matrix, we have

\[
PMPP' = \begin{pmatrix} M_{11} - M_{12}M_{22}^{-1}M_{21} & 0 \\ 0 & E_q \end{pmatrix} \geq 0.
\]

Therefore, \( M_{11} - M_{12}M_{22}^{-1}M_{21} \geq 0 \). \( \square \)

**Lemma 5.2.** Let the family \( (\sqrt{p_\theta(x)}) \) be \( L^2 \)-differentiable and \( \phi_\theta(X) \) be a function such that \( E_{\theta_0}\phi_\theta(X)^2 \) is bounded for every \( \theta \) in some neighbourhood of every fixed \( \theta_0 \in \Theta \). Then, it follows

\[
\int \phi_{\theta+h}(x)|p_{\theta+h}(x) - p_\theta(x) - (\hat{p}_\theta(x), h)|d\mu = o(|h|).
\]
Proof of Lemma 5.2. The above assertion is proved by the same argument as Ibragimov and Has’minskii (1981), Theorem 7.2. It is sufficient to verify in one dimension parameter case.

\[ \int \phi_{\theta+h}(x) \left| \frac{p_{\theta+h}(x) - p_{\theta}(x)}{h} - \tilde{\phi}'(x) \right| d\mu \]

\[ = \int \phi_{\theta+h}(x) \left| \left( \sqrt{p_{\theta+h}(x)} + \sqrt{p_{\theta}(x)} \right) \frac{\sqrt{p_{\theta+h}(x)} - \sqrt{p_{\theta}(x)}}{h} - 2 \sqrt{p_{\theta}(x)} \tilde{s}_{\theta}(x) \right| d\mu \]

\[ (5.1) \leq \int \phi_{\theta+h}(x) \left| \left( \sqrt{p_{\theta+h}(x)} + \sqrt{p_{\theta}(x)} \right) \frac{\sqrt{p_{\theta+h}(x)} - \sqrt{p_{\theta}(x)}}{h} - \tilde{s}_{\theta}(x) \right| d\mu \]

\[ + \int \phi_{\theta+h}(x) |\tilde{s}_{\theta}(x)| \left| \sqrt{p_{\theta+h}(x)} - \sqrt{p_{\theta}(x)} \right| d\mu. \]

\[ (5.1) \text{ does not exceed} \]

\[ (E_{\theta+h}^{1/2} \phi_{\theta+h}(X))^2 + E_{\theta}^{1/2} \phi_{\theta+h}(X)^2 \left( \int \left| \frac{\sqrt{p_{\theta+h}(x)} - \sqrt{p_{\theta}(x)}}{h} - \tilde{s}_{\theta}(x) \right|^2 d\mu \right)^{1/2} \]

\[ (5.2) \text{ is bounded above by} \]

\[ \left( \int \phi_{\theta+h}(x)^2 (p_{\theta+h}(x) + p_{\theta}(x)) d\mu \right)^{1/2} \]

\[ \times \left( \int_{|\phi_{\theta+h}(x)| > 1/(\sqrt{|h|})} |\tilde{i}_{\theta}(x)|^2 dP_{X}^{\theta} \right)^{1/2} \]

\[ + \frac{1}{2} \int_{\theta}^{1/2} \left( \frac{1}{|h|} \int \left( \sqrt{p_{\theta+h}(x)} - \sqrt{p_{\theta}(x)} \right)^2 d\mu \right)^{1/2}. \]

Since the probability measure \( P_{\theta}(|\phi_{\theta+h}(x)| > 1/|h|) \leq \sqrt{|h|} E_{\theta}|\phi_{\theta+h}(X)| \to 0 \)

as \( |h| \to 0), (5.3) \text{ tends to } 0 \text{ as } |h| \to 0. \text{ Hence } (5.2) \to 0 \text{ as } |h| \to 0. \text{ Thus we obtain the assertion.} \]

Lemma 5.3. Suppose that the family \( \left( \sqrt{p_{\theta}(x)} \right) \text{ is partial } L^{2}- \text{differentiable and } T(X) \text{ is a statistics with } E_{\theta}T(X)^2 < \infty \text{ for every } \theta. \text{ Then,} \)

\[ \frac{\partial}{\partial \theta} E_{\theta}T = \frac{\partial}{\partial \theta} \int T(x)p_{\theta}(x)d\mu = \int T(x)p_{\theta}(x)d\mu, \]

where \( \tilde{p}_{\theta} \) is the row vector of partial \( L^{1} \)-derivatives.
Proof of Lemma 5.3. Putting \( T(X) = \phi_{\theta+h}(X) \), the assertion is proved from Lemma 5.2. □

Remark 5.4 Setting on \( T(X) = 1 \), we have
\[
E_{\theta}[\dot{l}_{\theta}(X)] = \int \dot{p}_{\theta}(x) d\mu = 0.
\]

Proof of Theorem 3.1. It follows that
\[
E_{\theta}(Y - T(X))^2 = E_{\theta}(Y - g(X : \theta))^2 + E_{\theta}(g(X : \theta) - T(X))^2.
\]

We shall consider the lower bound in the second term of (5.4) because the first term does not include a predictor \( T \). Let \( S(X; \theta) = (g(X : \theta) - T(X)) \). By Remark 5.4 and unbiasedness of \( T \), we have
\[
E_{\theta}(S(X; \theta)) = 0.
\]

Let the covariance matrix of \( S(X; \theta) \) be \( Z(\theta) \), we have
\[
Z(\theta) = E_{\theta}[S(X; \theta)S(X; \theta)']
\]
\[
= E_{\theta} \left[ g(X : \theta) - T(X) \right] \left[ g(X : \theta) - T(X) \dot{l}_{\theta}(X)' \right]
\]
\[
= \begin{pmatrix}
E_{\theta}(g(X : \theta) - T(X))^2 & E_{\theta}(g(X : \theta) - T(X))\dot{l}_{\theta}(X)'
\end{pmatrix}.
\]

In view of Lemma 5.3 and unbiasedness of \( T(X) \), it follows that
\[
E_{\theta}T(X)\dot{l}_{\theta}(X)'
\]
\[
= \int T(x)\dot{p}_{\theta}(x)'d\mu
\]
\[
= \left( \frac{\partial}{\partial\theta} \int T(x)p_{\theta}(x)d\mu \right)'
\]
\[
= \left( \frac{\partial}{\partial\theta} E_{\theta}(Y) \right)'.
\]

Hence, we have
\[
E_{\theta}(g(X : \theta) - T(X))\dot{l}_{\theta}(X)'
\]
\[
= E_{\theta}[g(X : \theta)\dot{l}_{\theta}(X)'] - \left( \frac{\partial}{\partial\theta} E_{\theta}(Y) \right)'
\]
\[
= C(\theta)'
\]
and
\[
E_{\theta}\dot{l}_{\theta}(X)\dot{l}_{\theta}(X)' = I(\theta).
\]
Therefore, we obtain

\[ Z(\theta) = \begin{pmatrix} E_\theta(g(X : \theta) - T(X))^2 C(\theta)' \\ C(\theta) \end{pmatrix} C(\theta)^{-1} I(\theta). \]

Since \( Z(\theta) \) is a nonnegative symmetric matrix, it follows from Lemma 5.1 that

\[ E_\theta(g(X : \theta) - T(X))^2 \geq C(\theta)' I^{-1}(\theta) C(\theta). \]

Hence,

\[ E_\theta(Y - T(X))^2 \geq E_\theta(Y - g(X : \theta))^2 + C(\theta)' I^{-1}(\theta) C(\theta). \]

Thus we have the assertion. \( \square \)

Secondly, we shall state \( L^2 \)-differentiability of the conditional densities of \( Y \) given \( X \). For the proof of Theorem 4.1, we need the following lemmas. From now on, we denote the set \( \{ x; \rho_\theta(x) > 0 \} \) by \( A_\theta \).

**Lemma 5.5.** Suppose that the family \( (\sqrt{\rho_\theta(x,y)}) \) is \( L^2 \)-differentiable, then

\[ \int \int |\sqrt{\rho_{\theta+h}(y \mid x)} - \sqrt{\rho_\theta(y \mid x)}|^2 dP_\theta^X d\nu = O(|h|^2). \]

**Lemma 5.6.** Suppose that the family \( (\sqrt{\rho_\theta(x,y)}) \) is \( L^2 \)-differentiable. Let \( T(X) \) be a statistics with \( E_\theta|T(X)| < \infty \). Then,

\[ \lim_{|h| \to 0} \int \int T(X)|\sqrt{\rho_{\theta+h}(y \mid x)} - \sqrt{\rho_\theta(y \mid x)}|^2 dP_\theta^X d\nu = 0. \]

**Proof of Lemma 5.5.**

\[
\begin{align*}
\int_Y \int_{A_{\theta+h} \cap A_\theta} |\sqrt{\rho_{\theta+h}(y \mid x)} - \sqrt{\rho_\theta(y \mid x)}|^2 dP_\theta^X d\nu &= \int_Y \int_{A_{\theta+h} \cap A_\theta} |\sqrt{\rho_{\theta+h}(x,y)} \frac{\rho_\theta(x)}{\rho_{\theta+h}(x)} - \sqrt{\rho_\theta(x,y)}|^2 d\mu d\nu \\
&\leq \int \int |\sqrt{\rho_{\theta+h}(x,y)} - \sqrt{\rho_\theta(x,y)}|^2 d\mu d\nu \\
&\quad + \int_{A_{\theta+h}} |\sqrt{\rho_{\theta+h}(x)} - \sqrt{\rho_\theta(x)}|^2 \left( \int_Y \frac{\rho_{\theta+h}(x,y)}{\rho_{\theta+h}(x)} d\nu \right) d\mu \\
&= O(|h|^2).
\end{align*}
\]

It is clear that \( \int \int_{X-A_\theta} |\sqrt{\rho_{\theta+h}(y \mid x)} - \sqrt{\rho_\theta(y \mid x)}|^2 dP_\theta^X d\nu = 0 \) and \( \int \int_{X-A_{\theta+h}} |\sqrt{\rho_{\theta+h}(y \mid x)} - \sqrt{\rho_\theta(y \mid x)}|^2 dP_\theta^X d\nu = O(|h|^2) \). Hence, we obtain
the assertion. □

**Proof of Lemma 5.6.**

\[
\iint T(X) \left| \sqrt{p_{\theta+h}(y \mid x)} - \sqrt{p_{\theta}(y \mid x)} \right|^2 dP^X_{\theta} d\nu
\]

\[
= \iint_{|T(X)|>1/(\sqrt{|h|})} T(X) \left| \sqrt{p_{\theta+h}(y \mid x)} - \sqrt{p_{\theta}(y \mid x)} \right|^2 dP^X_{\theta} d\nu
\]

\[
+ \iint_{|T(X)|\leq1/(\sqrt{|h|})} T(X) \left| \sqrt{p_{\theta+h}(y \mid x)} - \sqrt{p_{\theta}(y \mid x)} \right|^2 dP^X_{\theta} d\nu
\]

\[
\leq 2 \int |T(X)| dP^X_{\theta}
\]

\[
+ \frac{1}{|h|} \iint \left| \sqrt{p_{\theta+h}(y \mid x)} - \sqrt{p_{\theta}(y \mid x)} \right|^2 dP^X_{\theta} d\nu.
\]

Using Lemma 5.5, we have the assertion. □

**Proof of Theorem 4.1.** Assume without loss of generality that \( \theta \in \Theta \) is one dimensional parameter.

\[
\iint_{A_{\theta} \cap A_{\theta+h}} \left| \frac{\sqrt{p_{\theta+h}(y \mid x)} - \sqrt{p_{\theta}(y \mid x)}}{h} \right|^2 dP^X_{\theta} d\nu
\]

\[
- \left( \frac{s_{\theta+h}(x, y)}{s_{\theta}(x)} - \frac{s_{\theta}(x, y)}{s_{\theta}(x)} \frac{s_{\theta+h}(x)}{s_{\theta}(x)} \frac{\dot{s}_{\theta}(x)}{s_{\theta}(x)} \right)^2 dP^X_{\theta} d\nu
\]

(5.5) \[
\leq 2 \iint \left| \frac{s_{\theta+h}(x, y)}{h} - \frac{s_{\theta}(x, y)}{h} - \dot{s}_{\theta}(x, y) \right|^2 d\mu d\nu
\]

(5.6) \[
+ 2 \iint_{A_{\theta} \cap A_{\theta+h}} \left| \frac{s_{\theta+h}(x)}{h} s_{\theta+h}(y \mid x) - s_{\theta}(y \mid x) \dot{s}_{\theta}(x) \right|^2 d\mu d\nu.
\]

(5.6) is bounded above by

(5.7) \[
4 \int \left| \frac{s_{\theta+h}(x)}{h} - \dot{s}_{\theta}(x) \right|^2 d\mu
\]

(5.8) \[
+ \iint \left| \sqrt{p_{\theta+h}(y \mid x)} - \sqrt{p_{\theta}(y \mid x)} \right|^2 |\dot{l}_{\theta}(x)| dP^X_{\theta} d\nu.
\]

Applying Lemma 5.6, (5.8) tends to 0 as \(|h| \to 0\). In view of \( L^2 \)-differentiability, (5.5) and (5.7) approach 0 as \(|h| \to 0\). Therefore, \( L^2 \)-differentiability of \((\sqrt{p_{\theta}(y \mid x)})\) is proved in the domain of integration \((A_{\theta+h} \cap A_{\theta}) \times \mathcal{Y}\). And it is easily verified that \( L^2 \)-differentiability in the
integral domain \((\mathcal{X} - A_{\theta+h}) \times \mathcal{Y}\), and \((\mathcal{X} - A_{\theta}) \times \mathcal{Y}\). Hence, the assertion is proved.

**Theorem 5.7** Suppose that \((P^X_\theta)\) satisfies conditions (ii). Then \(E_\theta(Y \mid X)\) is \(L^1\)-differentiable in measure \(P^X_\theta\) with the derivative

\[
\dot{g}(X \mid \theta) = \int y\dot{p}_\theta(y \mid x) d\nu,
\]

where \(\dot{p}_\theta(y \mid x)\) is the \(L^1\)-derivative. We can easily verify that \(\dot{p}_\theta(y \mid x) = \left(\frac{\dot{p}_\theta(x,y) - p_\theta(x,y)}{p_\theta(x)^2}\right)1_{[p_\theta(x) > 0]}\) if the family \((\sqrt{p_\theta(x,y)})\) is \(L^2\)-differentiable.

**Proof of Theorem 5.7.**

\[
\frac{1}{|h|}\int_{A_\theta \cap A_{\theta+h}} |E_{\theta+h}(Y \mid x) - E_\theta(Y \mid x) - (h, \dot{g}(x \mid \theta))| dP^X_\theta
\]

\[
= \frac{1}{|h|} \int_{A_\theta \cap A_{\theta+h}} \left| \int y[p_{\theta+h}(y \mid x)p_\theta(x) - p_\theta(x,y) - (h, \dot{p}_\theta(y \mid x))\right] d\nu \right| d\mu
\]

\[
\leq \frac{1}{|h|} \int_{A_\theta \cap A_{\theta+h}} \left| \int y[p_{\theta+h}(y \mid x) - p_\theta(x) - (h, \dot{p}_\theta(x))]\right| d\nu \right| d\mu
\]

\[
+ \frac{1}{|h|} \int_{A_\theta \cap A_{\theta+h}} \left| \int y\dot{p}_\theta(y)\right| d\nu \right| d\mu
\]

\[
+ \int_{A_\theta \cap A_{\theta+h}} \left| \int y\dot{p}_\theta(y)\right| d\nu \right| d\mu
\]

Using Lemma 5.2 and Lemma 5.3, (5.10) and (5.11) tend to 0 as \(|h|\to 0\) respectively. By the same argument as Lemma 5.6, (5.12) tends to 0 as \(|h|\to 0\). Hence, (5.9) \(\to 0\) as \(|h|\to 0\). On the other hand, it is clear that \(\int_{\mathcal{X} - A_\theta} |E_{\theta+h}(Y \mid X) - E_\theta(Y \mid X) - (h, \dot{g}(X \mid \theta))| dP^X_\theta = 0\). Further, since \(\dot{p}_\theta(y \mid x)\) is written by \((\dot{p}_\theta(x,y)p_\theta(x) - p_\theta(x,y)\dot{p}_\theta(x))/p_\theta(x)^2\), it holds that

\[
\frac{1}{|h|} \int_{\mathcal{X} - A_{\theta+h}} |E_{\theta+h}(Y \mid x) - E_\theta(Y \mid x) - (h, \dot{g}(x \mid \theta))| dP^X_\theta
\]

\[
\leq \frac{1}{|h|} \int_{\mathcal{X} - A_{\theta+h}} |\dot{g}(x \mid \theta)| dP^X_\theta
\]

\[
\leq o(1) + (E_\theta g(X \mid \theta)^2)^{1/2} \left(\int_{\mathcal{X} - A_{\theta+h}} |\dot{g}(x \mid \theta)|^2 dP^X_\theta\right)^{1/2}
\]

\(\to 0\) as \(|h|\to 0\).
Hence, the theorem is proved.

**Lemma 5.8.** Under conditions (ii), it follows
\[
\frac{\partial}{\partial \theta} \int g(x : \theta)p_\theta(x)d\mu = \int \dot{g}(x : \theta)p_\theta(x)d\mu + \int g(x : \theta)\dot{p}_\theta(x)d\mu.
\]

**Proof of Lemma 5.8.**
(5.13) \[
\left| \int \frac{g(x : \theta + h)p_{\theta + h}(x) - g(x : \theta)p_\theta(x)}{h} \right| - \dot{g}(x : \theta)p_\theta(x) - g(x : \theta)\dot{p}_\theta(x)d\mu 
\]
(5.14) + \int |g(x : \theta + h)||\frac{p_{\theta + h}(x) - p_\theta(x)}{h} - \dot{p}_\theta(x)|d\mu
(5.15) + \int |g(x : \theta + h) - g(x : \theta)||\dot{p}_\theta(x)|d\mu.

By Theorem 5.7, (5.13) → 0 as \(|h| → 0\). By applying Lemma 5.2, (5.14) → 0 as \(|h| → 0\). (5.15) is bounded above by
(5.16) \[
(E_\theta g(X : \theta + h)^2 + E_\theta Y^2) \int_{|\dot{\theta}(x)|^2 > 1/\sqrt{|h|}} |\dot{\theta}(x)|^2 dP_\theta^X
\]
(5.17) + \frac{1}{\sqrt{|h|}} \int |g(x : \theta + h) - g(x : \theta)|dP_\theta^X.

(5.16) → 0 as \(|h| → 0\) by the same argument as (5.3). Since \(g(X : \theta)\) is \(L^1\)-differentiable, (5.17) → 0 as \(|h| → 0\). Thus the lemma is proved.

**Proof of Theorem 3.2.** It is sufficient to prove \(C(\theta) = -E_\theta(\dot{g}(X : \theta))\). Since \(T(X)\) is unbiased, \(0 = \int (g(x : \theta) - T(x))dP_\theta^X\). By using Lemma 5.8, it follows that
\[
0 = \frac{\partial}{\partial \theta} \int (g(x : \theta) - T(x))p_\theta(x)d\mu
\]
\[
= \int \dot{g}(x : \theta)p_\theta(x)d\mu + \int (g(x : \theta) - T(x))\dot{p}_\theta(x)d\mu.
\]
Since \(C(\theta) = \int (g(x : \theta) - T(x))\dot{p}_\theta(x)d\mu\), we obtain the assertion.

**Lemma 5.9.** Suppose that \((\sqrt{p_\theta(y \mid x)})\) is \(L^2\)-differentiable. Let \(T(X)\) be a statistics with \(E_\theta|T| < \infty\). Then, the following assertions hold:

(i) \[
\int \int |\sqrt{p_{\theta + h}(y \mid x)} - \sqrt{p_\theta(y \mid x)}|^2 dP_\theta^Xd\nu = O(|h|^2).
\]
(ii) \[
\lim_{|h| \to 0} \int \int T(X)|\sqrt{p_{\theta + h}(y \mid x)} - \sqrt{p_\theta(y \mid x)}|^2 dP_\theta^Xd\nu = 0.
\]
Proof of Lemma 5.9. (i) is easily verified from the definition of $L^2$-differentiability. (ii) is proved by the same argument as in the proof of Lemma 5.6.

Proof of Theorem 4.2. It is sufficient to verify in one dimensional parameter case.

\[
\int_{A_\theta \cap A_{\theta+h}} \left| \sqrt{p_{\theta+h}(x, y)} - \sqrt{p_{\theta}(x, y)} \right| h \\
- (s_\theta(y \mid x)\sqrt{p_{\theta}(x)} + \sqrt{p_{\theta}(y \mid x)s_\theta(x)}) \right|^2 d\mu d\nu
\]

\[\leq 3 \int_{A_\theta \cap A_{\theta+h}} \left| \sqrt{p_{\theta+h}(y \mid x)} - \sqrt{p_{\theta}(y \mid x)} \right|^2 dP^X d\nu \tag{5.18}\]

\[+ 3 \int_{A_\theta \cap A_{\theta+h}} \left| \sqrt{p_{\theta+h}(x)} - \sqrt{p_{\theta}(x)} \right|^2 \dot{s}_\theta(x) d\mu d\nu \tag{5.19}\]

\[+ 3 \int_{A_\theta \cap A_{\theta+h}} \left| \dot{s}_\theta(x) \right|^2 \left| \sqrt{p_{\theta+h}(y \mid x)} - \sqrt{p_{\theta}(y \mid x)} \right|^2 d\mu d\nu. \tag{5.20}\]

By the assumption, (5.18) tends to 0 as $|h| \to 0$, and in view of $\int_Y p_{\theta+h}(y \mid x) d\nu = 1$, (5.19) tends to 0 as $|h| \to 0$.

Applying Lemma 5.9, (5.20) tends to 0 as $|h| \to 0$. Therefore, $L^2$-differentiability of the family $(\sqrt{p_{\theta}(x, y)})$ is proved in the integral domain $(A_{\theta+h} \cap A_\theta) \times Y$. It is verified that $\int_{\mathcal{X} - A_\theta} \left| \frac{\sqrt{p_{\theta+h}(x,y)} - \sqrt{p_{\theta}(x,y)}}{h} - \dot{s}_\theta(x,y) \right|^2 d\mu d\nu = 0$ and $\lim_{|h| \to 0} \int_{\mathcal{X} - A_{\theta+h}} \left| \frac{\sqrt{p_{\theta+h}(x,y)} - \sqrt{p_{\theta}(x,y)}}{h} - \dot{s}_\theta(x,y) \right|^2 d\mu d\nu = 0$.

Hence the assertion is proved.

The following lemma gives sufficient conditions for continuous $L^2$-differentiability of densities in terms of ordinary differentiability of the likelihood. This result is proved by Bickel, Klaasen, Ritov and Wellner (1993), Proposition 1, p. 13.

Lemma 5.10. Suppose that $\Theta$ is open, and that for all $\theta$

(a) $p_\theta(x)$ is continuous differentiable in $\theta$ for $\mu$ almost all $x$ with gradient $\dot{p}_\theta(x)$.

(b) $|\dot{l}_\theta(x)| \in L^2(P^X_\theta)$ with $\dot{l}_\theta(x)$ as in (2.2).

(c) Fisher’s Information matrix $I(\theta) = \int \dot{l}_\theta(x)\dot{l}_\theta(x)' dP^X_\theta$ is nonsingular and continuous in $\theta$. 

Then the parametrization $\theta \rightarrow P^X_\theta$ is continuous $L^2$-differentiable with the derivative $s_\theta(x)$, where $s_\theta(x) = \frac{\dot{p}_\theta(x)}{2\sqrt{p_\theta(x)}}1_{[p_\theta(x)>0]}$.

Acknowledgements

The author thanks the editor and the referee for their helpful comments which have improved the presentation of the present paper.

References


