For JAW 65th birthday

A NOTE ON INSUFFICIENCY AND THE PRESERVATION OF FISHER INFORMATION

By David Pollard

Yale University

Kagan and Shepp (2005) presented an elegant example of a mixture model for which an insufficient statistic preserves Fisher information. This note uses the regularity property of differentiability in quadratic mean to provide another explanation for the phenomenon they observed. Some connections with Le Cam's theory for convergence of experiments are noted.

1. Introduction. Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a statistical experiment, a set of probability measures on some $(\mathcal{X}, \mathcal{A})$ indexed by a subset Θ of the real line.

The Fisher information function $\mathbb{I}_{\mathcal{P}}(\theta)$ can be defined under various regularity conditions. If S is a measurable map from $\mathcal{X}$ into another measure space $(\mathcal{Y}, \mathcal{B})$, each image measure $Q_\theta = SP_\theta$ (often called the distribution of S under P_θ, and sometimes denoted by $P_\theta S^{-1}$) is a probability measure on $\mathcal{B}$. The statistical experiment $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$ is less informative, in the sense that an observation $y \sim Q_\theta$ tells us less about θ than an observation $x \sim P_\theta$. In particular, $\mathbb{I}_{\mathcal{Q}}(\theta) \le \mathbb{I}_{\mathcal{P}}(\theta)$ for every θ. If S is a sufficient statistic the last inequality becomes an equality: there is no loss of Fisher information.

Kagan and Shepp (2005) (henceforth K&S) showed, by means of a simple example, that it is possible to have $\mathbb{I}_{\mathcal{Q}}(\theta) = \mathbb{I}_{\mathcal{P}}(\theta)$ for every θ without S being sufficient.

The purpose of this note is: (i) using the geometry of differentiabilty in quadratic mean, to reinterpret the phenomenon identified by K&S; (ii) to explain why the experiment $\mathcal{Q}_n$ obtained by n independent replications of $\mathcal{Q}$ is asymptotically equivalent (in Le Cam's sense) to the corresponding $\mathcal{P}_n$.

Most of the necessary theory is already available in the literature but is not widely known. The K&S example provides a good showcase for that theory.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35

Keywords and phrases: Fisher information; Hellinger differentiability of probability models; differentiability in quadratic mean; score function; Le Cam's distance between statistical models

2. The K&S example. What follows is a slightly simplified version of the K&S construction.

Start from a smooth probability density

$$g(w) = \tfrac{1}{2} w^2 e^{-w} \{w > 0\}$$

with respect to Lebesgue measure $\mathfrak{m}$ on the real line. The power w^2 is chosen so that

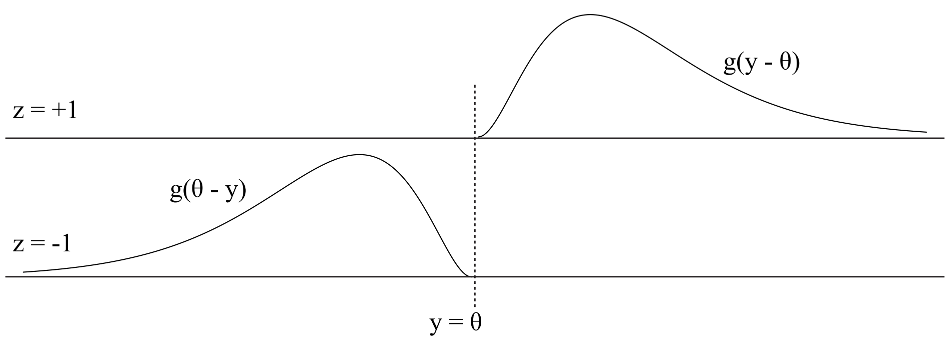
$$\frac{\dot g(w)^2}{g(w)} = g(w) \left(\frac{d \log g(w)}{dw} \right)^2 = \frac{1}{2} (2 - w)^2 e^{-w} \{w > 0\}$$

is Lebesgue integrable.

Let ν denote the probability measure that puts mass 1/2 at each of $+1$ and -1. For each $\theta \in \Theta = \mathbb{R}$ define a probability measure P_θ on (the Borel sigma-field of) $\mathcal{X} = \mathbb{R} \times \{-1, +1\}$ by means of its density

(1) $$f_\theta(x) = \{z = +1\} g(y - \theta) + \{z = -1\} g(\theta - y) \qquad \text{where } x = (y, z) \in \mathcal{X}$$

with respect to the measure $\lambda := \mathfrak{m} \otimes \nu$. That is, the coordinate z has marginal distribution ν and the conditional distribution of y given z is that of $\theta + zw$ where $w \sim g$ independently of z.



Define the statistic S as the coordinate projection, $S(y, z) = y$. Here are the pertinent facts. (See the next Section for some proofs.)

The distribution Q_θ of S has density

$$h_\theta(y) = \tfrac{1}{2} g(y - \theta) + \tfrac{1}{2} g(\theta - y) \qquad \text{with respect to } \mathfrak{m}.$$

Both $\mathcal{P}$ and $\mathcal{Q}$ have finite Fisher information,

(2) $$\mathbb{I}_{\mathcal{Q}}(\theta) = \mathbb{I}_{\mathcal{P}}(\theta) = \mathbb{I} := \int_{-\infty}^{\infty} \dot g(w)^2 / g(w) \, dw < \infty \qquad \text{for all } \theta.$$

There is no loss of Fisher information when z is ignored. However, the statistic S is not sufficient because

$$P_\theta(z = 1 \mid S = y) = \{y > \theta\},$$

which depends on θ. More formally, if S were sufficient there would exist some measurable function $\pi(y)$ for which $P_\theta(z = 1 \mid S = y) = \pi(y)$ a.e$[Q_\theta]$, for every θ.

Remark. K&S used a slightly more involved construction, with density

$$\begin{aligned} f(x,\theta) = \{z = +1\}\,&[0.7g(y-\theta) + 0.3g(\theta - y)] \\ &+ \{z = -1\}\,[0.3g(y-\theta) + 0.7g(\theta - y)] \qquad \text{where } x = (y,z) \in \mathcal{X} \end{aligned}$$

with respect to $\mathfrak{m} \otimes \mu$ where $\mu\{+1\} = \alpha = 1 - \mu\{-1\}$ and $\alpha \neq 1/2$. The analysis in this note can be extended to this f_θ.

3. DQM interpretation. K&S attributed the phenomenon in their version of the example in Section 2 to a failure of strict convexity of Fisher information with respect to mixtures of statistical experiments. There is another explanation involving the geometry of Hellinger derivatives, which I find more illuminating.

By a theorem of Hájek (1972, Lemma A.3), Lebesgue integrability of the function $\dot{g}^2/g$ in (2) implies that the set of densities $\mathcal{G} := \{g(y-\theta) : \theta \in \mathbb{R}\}$ (with respect to Lebesgue measure) is Hellinger differentiable with Hellinger derivative $\gamma(y - \theta)$ at θ, where

$$\gamma(w) := \frac{-\dot{g}(w)}{2\sqrt{g(w)}} = \frac{(2-w)}{2\sqrt{2}} e^{-w/2}\{w > 0\}.$$

That is,

$$\int \left| \sqrt{g(y-\theta-t)} - \sqrt{g(y-\theta)} - t\gamma(y-\theta) \right|^2 dy = o(|t|^2) \qquad \text{as } t \to 0.$$

This assertion is also easy to check by explicit calculations.

The family of densities $\mathcal{F} := \{f_\theta(x) : \theta \in \mathbb{R}\}$, for f_θ as in (1), inherits the Hellinger differentiability from $\mathcal{G}$:

$$\lambda \left| \sqrt{f_{\theta+t}(x)} - \sqrt{f_\theta(x)} - t\zeta_\theta(x) \right|^2 = o(|t|^2) \qquad \text{as } t \to 0, \tag{3}$$

for the Hellinger derviative

$$\zeta_\theta(x) := \{z = +1\}\gamma(y - \theta) - \{z = -1\}\gamma(\theta - y).$$

The significance of approximation (3) becomes clearer when it is rewritten as a differentiability property of the likelihood ratios. That is, it helps to work with the square root of the density of $P_{\theta+t}$ with respect to P_θ. Unfortunately, $P_{\theta+t}$ is not dominated by P_θ. In general, to eliminate such an embarrassment one needs to split $P_{\theta+t}$ into a singular part $P_{t,\theta}^{\perp}$, which concentrates on a set of zero P_θ measure, plus a part $P_{\theta+t}^{(abs)}$ that has a density $p_{t,\theta}$ with respect to P_θ. For reasons related to the asymptotic theory for repeated sampling, it is customary to make a small extra assumption about the behavior of $P_{t,\theta}^{\perp}\mathcal{X}$ as t tends to zero. Following Le Cam (1986, Section 17.3) and Le Cam and Yang (2000, Section 7.2), I will call the slightly stronger property ***differentiability in quadratic mean (DQM)***, to stress that the definition requires a little more than Hellinger differentiability.

> **Remark.** Beware: Some authors (for example, Bickel *et al.* 1993, page 457) use the term DQM as a synonym for Hellinger differentiability.

DEFINITION 4. *Say that $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, with $\Theta \subseteq \mathbb{R}$, is differentiable in quadratic mean (DQM) at θ with score function $\Delta_\theta(x)$ if, for $\theta + t \in \Theta$,*

(i) for the part $P_{t,\theta}^{\perp}$ of $P_{\theta+t}$ that is singular with respect to P_θ,

$$P_{t,\theta}^{\perp}(\mathcal{X}) = o(|t|^2) \qquad \text{as } |t| \to 0$$

(ii) $\Delta_\theta \in \mathcal{L}^2(P_\theta)$

(iii) the absolutely continuous part of $P_{\theta+t}$ has density $p_{t,\theta}(x)$ with respect to P_θ for which

$$\sqrt{p_{t,\theta}(x)} = 1 + \tfrac{1}{2}t\Delta_\theta(x) + r_{t,\theta}(x) \qquad \text{with } P_\theta\left(r_{t,\theta}^2\right) = o(|t|^2) \text{ as } t \to 0.$$

> **Remark.** The factor of $1/2$ in requirement (iii) ensures that $P_\theta\Delta_\theta^2$ is equal to the Fisher information $\mathbb{I}_{\mathcal{P}}(\theta)$ if the densities are suitably smooth in a pointwise sense.

Call $\mathcal{P}$ DQM if it is DQM at each θ in Θ.

The $\mathcal{P}$ from Section 2 is, in fact, DQM. For $t > 0$ the singular part $P_{t,\theta}^{\perp}$ has density $\{z = -1\}g(\theta - y)\{\theta < y < \theta + t\}$ with respect to λ, so that $P_{t,\theta}^{\perp}(\mathcal{X}) = O(|t|^3)$. The part of $P_{\theta+t}$ that is dominated by P_θ has density

$$\begin{aligned} p_{t,\theta}(x) &= \frac{f_{\theta+t}(x)}{f_\theta(x)}\{f_\theta(x) > 0\} \\ &= \{z = +1\}\frac{g(y-\theta-t)}{g(y-\theta)}\{y > \theta\} + \{z = -1\}\frac{g(\theta+t-y)}{g(\theta-y)}\{y < \theta\} \qquad (5) \end{aligned}$$

There is a similar expression for the case $t < 0$. The score function equals

$$\Delta_\theta(x) = 2\frac{\zeta_\theta(x)}{\sqrt{f_\theta(x)}}\{f_\theta(x) > 0\}$$
$$= \{z = +1\}\frac{\gamma(y-\theta)}{g(y-\theta)}\{y > \theta\} - \{z = -1\}\frac{\gamma(\theta-y)}{g(\theta-y)}\{y < \theta\} \tag{6}$$

The density $p_{t,\theta}$ and the score function $\Delta_\theta(x)$ are uniquely determined only up to a P_θ equivalence. With that thought in mind, observe that both $\{z = +1\} = \{y > \theta\}$ a.e.$[P_\theta]$ and $\{z = -1\} = \{y < \theta\}$ a.e.$[P_\theta]$. The score function Δ_θ is only changed on a P_θ-negligible set if we omit the two indicator functions involving z from (6). In effect, the score function $\Delta_\theta(x)$ depends on x only through the value of the statistic S. That property is exactly what we need to preserve Fisher information. The relevant facts are contained in the next theorem, which is proved in Section 5.

THEOREM 7. *Suppose $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ on $(\mathcal{X}, \mathcal{A})$ is DQM with score function Δ_θ. Suppose S is a measurable map from $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{Y}, \mathcal{B})$ and $Q_\theta = SP_\theta$ is the distribution of S under P_θ. Then:*

(i) The statistical experiment $\mathcal{Q} = \{Q_\theta : \theta \in \Theta\}$ is also DQM, with score function $\widetilde{\Delta}_\theta(y) = P_\theta(\Delta_\theta \mid S = y)$.
(ii) At each fixed θ, Fisher information is preserved (that is, $\mathbb{I}_{\mathcal{P}}(\theta) = \mathbb{I}_{\mathcal{Q}}(\theta)$) if and only if $\Delta_\theta(x) = \widetilde{\Delta}_\theta(Sx)$ a.e.$[P_\theta]$.

With only notational changes, the Theorem extends to the case where Θ is a subset of some Euclidean space; no extra conceptual difficulties arise in higher dimensions.

Credit where credit is due. The results stated in Theorem 7 have an interesting history. Part (i) was asserted ("Direct calculations show that the function $q^{1/2}(y;\theta)$ is differentiable in $L_2(\widetilde{\nu})$ and possess a continuous derivative ...") in Theorem 7.2 of Ibragimov and Has'minskii (1981, Chapter I, page 70), an English translation from the 1979 Russian edition. However, that Theorem also (incorrectly, as noted by K&S) asserted that Fisher information is preserved if and only if S is sufficient.

Pitman (1979, pages 19–21) established differentiability in mean, a property slightly different from (i), in order to deduce a result equivalent to (ii).

Le Cam and Yang (1988, Section 7) deduced an analogue of (i) (preservation of DQM under restriction to sub-sigma-fields) by an indirect argument using equivalence of DQM with the existence of a quadratic approximation to likelihood ratios of product measures (an LAN condition).

Bickel *et al.* (1993, page 461) proved result (i), citing Ibragimov and Has'minskii (1981), Le Cam and Yang (1988), and van der Vaart (1988, Appendix A3) for earlier proofs. The last of these was a revised ("I have not resisted the temptation to rewrite numerous parts of the original manuscript") version of van der Vaart's 1987 Ph.D. thesis. He cited Le Cam and Yang (1988) and a manuscript version of Bickel *et al.* (1993).

4. Large sample interpretation. Write $\mathbb{P}_{\theta,n}$ for the n-fold product measure P_θ^n, and $\mathbb{Q}_{n,\theta}$ for Q_θ^n, with P_θ and Q_θ as in Section 2. That is, the statistical experiment $\mathcal{P}_n = \{\mathbb{P}_{\theta,n} : \theta \in \Theta\}$ corresponds to taking n independent observations $x_1 = (y_1, z_1), \dots, x_n = (y_n, z_n)$ from P_θ and $\mathcal{Q}_n = \{\mathbb{Q}_{\theta,n} : \theta \in \Theta\}$ corresponds to $y_1, \dots, y_n$.

Classical theory establishes existence of estimators $\widehat{\theta}_n = \widehat{\theta}_n(y_1, \dots, y_n)$ for which $\sqrt{n}\left(\widehat{\theta}_n - \theta\right)$ converges in distribution under $\mathbb{P}_{\theta,n}$ to $N(0, \mathbb{I}^{-1})$. What more do we learn from the z_i's? Asymptotically speaking, not much.

For example, as shown by the Hájek-Le Cam convolution and asymptotic minimax theorems (Bickel *et al.*, 1993, Section 2.3), there are various senses in which $N(0, \mathbb{I}^{-1})$ is the best we can hope to achieve. Indeed, the score function essentially determines the behaviour of the likelihood function in $O(n^{-1/2})$ neighborhoods of θ, which controls the asymptotics at the "$\sqrt{n}$" level. The z_i's must be contributing at a less important level.

For $i = 1, \dots, n$, suppose $y_{L:n}$ is the largest y_i for which $z_i = -1$ and $y_{R:n}$ the smallest y_i for which $y_i = +1$. With $\mathbb{P}_{\theta,n}$ probability one we know that $y_{L:n} < \theta < yR$. The w^2 decay in $g(w)$ at zero, implies that both $\theta - y_{L:n}$ and $y_{R:n} - \theta$ are decreasing at an $n^{-1/3}$ rate. In fact both $n^{1/3}(\theta - y_{L:n})$ and $n^{1/3}(y_{R:n} - \theta)$ have nontrivial limit distributions under $\mathbb{P}_{\theta,n}$. The event $A_n = \{y_{L:n} < \widehat{\theta}_n < y_{R:n}\}$ has $\mathbb{P}_{\theta,n}$ probability that tends very rapidly to one.

Define $z^*_{i,n} = \operatorname{sgn}(y_i - \widehat{\theta}_n)$. That is,

$$z^*_{i,n} = \begin{cases} +1 & \text{if } y_i > \widehat{\theta}_n \\ -1 & \text{if } y_i < \widehat{\theta}_n \end{cases} .$$

Define $x^*_{i,n} = (y_i, z^*_{i,n})$. On the event A_n we have $x_i = x^*_{i,n}$ for $i = 1, \dots, n$. If $\mathbb{P}^*_{\theta,n}$ denotes the joint distribution of $x^*_{1,n}, \dots, x^*_{n,n}$ then

$$\sup_{\theta \in \Theta} \|\mathbb{P}^*_{\theta,n} - \mathbb{P}_{\theta,n}\|_{\text{TV}} \to 0 \qquad \text{rapidly.}$$

In the terminology of Le Cam's convergence of statistical experiments, $\mathcal{P}_n$ and $\mathcal{Q}_n$ are asymptotically equivalent; the vector $(y_1, \dots, y_n)$ is asymptotically sufficient for $\mathcal{P}_n$ in Le Cam's sense.

Remark. Rough calculations suggest that the Le Cam distance between $\mathcal{P}_n$ and $\mathcal{Q}_n$ tends to zero like $\exp(-Cn^{1/3})$ for some constant C. I omit the details because the actually rate is not important for the story I am telling.

Put another way, for every statistic $\psi_n(x_1,\dots,x_n)$ for $\mathcal{P}_n$ there is another statistic $\psi_n^*(y_1,\dots,y_n) = \psi_n(x_{1,n}^*,\dots,x_{n,n}^*)$ for $\mathcal{Q}_n$ that has the same asymptotic behavior.

5. Proof of Theorem 7. Recall that the Kolmogorov conditional expectation $P_\theta(\cdot \mid S = y)$ is abstractly defined, via the Radon-Nikodym theorem, as an increasing linear map (depending on θ) $\kappa : L_1(P_\theta) \to L^1(Q_\theta)$ with properties analogous to those enjoyed by a Markov kernel. If we identify an f in $L^1(P_\theta)$ with the (signed) measure μ_f for which $d\mu_f/dP_\theta = f$, then $g = \kappa f$ is the density of $S\mu_f$ with respect to Q_θ. To stress the analogy with Markov kernels I will write $\kappa_y f$, or even $\kappa_y f(x)$, instead of $(\kappa f)(y)$. Thus the defining property of κ can be rewritten as

$$Q_\theta f_1(y)\kappa_y f_2 = P_\theta f_1(Sx) f_2(x) \tag{8}$$

for measurable real functions f_1 on $\mathcal{Y}$ and f_2 on $\mathcal{X}$, at least when $f_1(Sx)f_2(x)$ is P_θ-integrable. A reader who chose to interpret κ_y as a Markov kernel would lose only a tiny amount of generality.

Of course if one regards κ as acting on $\mathcal{L}^1(P_\theta)$, instead of on the space $L^1(P_\theta)$ of P_θ-equivalence classes, then one should qualify assertions with the occasional a.e.$[P_\theta]$ caveats and regard κf as being defined only up to Q_θ equivalence. Following the usual custom, I will omit such qualifiers.

Proof of assertion (i). The following argument is adapted from van der Vaart (1988, Appendix A3).

To simplify notation, I will prove that $\mathcal{Q}$ is DQM only at $\theta = 0$, writing $P_t^\perp$ instead of $P_{t,0}^\perp$ and p_t instead of $p_{t,0}$. Keep in mind that κ_y now denotes the conditional expectation operator $P_0(\cdot \mid S = y)$. For each function $h(x)$ in $\mathcal{L}^2(P_0)$ I will write $\widetilde{h}(y)$ for its conditional mean $\kappa_y h(x)$ and

$$\text{var}_y h := \kappa_y \left(h(x) - \widetilde{h}(y)\right)^2 = \kappa_y h(x)^2 - \widetilde{h}(y)^2$$

for its conditional variance.

Start with the simplest case where P_t is actually dominated by P_0. Then

$$\xi_t(x) = \sqrt{dP_t/dP_0} = 1 + \tfrac{1}{2}t\Delta_0(x) + r_t(x) \qquad \text{with } P_0 r_t^2 = o(t^2)$$

and

(9) $\widetilde{\xi}_t(y) := \kappa_y \xi_t(x) = 1 + \frac{1}{2} t \widetilde{\Delta}_0(y) + \widetilde{r}_t(y) \qquad \text{with } Q_0 \widetilde{r}_t^2 \le P_0 r_t^2 = o(t^2).$

and, by the Radon-Nikodym property,

$$\eta_t(y) = \sqrt{dQ_t/dQ_0} = \sqrt{\kappa_y \xi_t(x)^2} \quad .$$

The proof of assertion (i) will work by showing that the difference $\delta_t(y) := \eta_t(y) - \widetilde{\xi}_t(y)$ is small, in the sense that $Q_0 \delta_t^2 = o(t^2)$. For then we will have

$$\eta_t(y) = 1 + \frac{1}{2} t \widetilde{\Delta}_0(y) + \left[\widetilde{r}_t(y) + \delta_t(y)\right] \qquad \text{with } Q_0 \left[\widetilde{r}_t(y) + \delta_t(y)\right]^2 = o(t^2),$$

which implies DQM for $\mathcal{Q}$ at 0.

The desired property for δ_t will be derived from the following facts about the conditional variance

(10) $$\sigma_t^2(y) := \text{var}_y(\xi_t) = \kappa_y \xi_t(x)^2 - \widetilde{\xi}_t(y)^2 = \eta_t(y)^2 - \widetilde{\xi}_t(y)^2.$$

(a) The representation $\sigma_t^2(y) = \kappa_y \left(\xi_t(x) - \widetilde{\xi}_t(y)\right)^2$ gives

$$\begin{aligned}
\sigma_t^2(y) &= \kappa_y \left(\tfrac{1}{2} t \left[\Delta_0(x) - \widetilde{\Delta}_0(y)\right] + \left[r_t(x) - \widetilde{r}_t(y)\right] \right)^2 \\
&\le 2 \left(\tfrac{1}{2} t\right)^2 \kappa_y \left[\Delta_0(x) - \widetilde{\Delta}_0(y)\right]^2 + 2\kappa_y \left[r_t(x) - \widetilde{r}_t(y)\right]^2 \\
&\le \tfrac{1}{2} t^2 \kappa_y \Delta_0^2 + 2\kappa_y r_t^2 .
\end{aligned}$$

> **Remark.** The cancellation of the leading 1 when $\widetilde{\xi}_t$ is subtracted from ξ_t seems to be vital to the proof. For general Hellinger differentiability, the cancellation would not occur.

(b) $\delta_t(y) \ge 0$ because $\eta_t(y)^2 - \widetilde{\xi}_t(y)^2 = \sigma_t^2(y) \ge 0$.

(c) Substitution of $\delta_t + \widetilde{\xi}_t$ for η_t in (10) gives

$$\sigma_t^2(y) = 2\delta_t(y)\widetilde{\xi}_t(y) + \delta_t(y)^2.$$

The rest is easy. For each $\epsilon > 0$ define

$$A_{t,\epsilon} := \{y \in \mathcal{Y} : \widetilde{\xi}_t(y) \ge \tfrac{1}{2},\ \sigma_t(y) \le \epsilon\}.$$

Integration of inequality (a) gives

$$Q_0 \sigma_t^2(y) \le \tfrac{1}{2} t^2 P_0 \Delta_0^2 + 2 P_0 r_t^2 = O(t^2) + o(t^2) \le C t^2 \qquad \text{for some constant } C,$$

which, together with (9), implies $Q_0 A_{t,\epsilon} \to 1$ as $t \to 0$.

On the set $A_{t,\epsilon}$ equality (c) ensures that $\delta_t(y) \le \sigma_t^2(y) \le \epsilon\sigma_t(y)$; on $A_{t,\epsilon}^c$ the nonnegativity of δ_t and equality (c) give $\delta_t^2 \le \sigma_t^2$. Thus

$$\begin{aligned} Q_0\delta_t(y)^2 &\le \epsilon^2 Q_0\sigma_t^2(y)\{y \in A_{t,\epsilon}\} + Q_0\sigma_t^2(y)\{y \notin A_{t,\epsilon}\} \\ &\le \epsilon^2 C t^2 + \tfrac{1}{2}t^2 Q_0\kappa_y\Delta_0^2 A_{t,\epsilon}^c + 2Q_0\kappa_y r_t^2 \qquad \text{by (a).} \end{aligned}$$

The Q_0-integrability of $\kappa_y\Delta_0^2$ and the Dominated Convergence theorem imply $Q_0\kappa_y\Delta_0^2 A_{t,\epsilon}^c \to 0$. It follows that $Q_0\delta_t^2 = o(t^2)$.

Finally, what happens when P_t is not dominated by P_0? The analysis for ξ_t^2, the density of the part of P_t that is dominated by P_0, is the same as before. The image measure $SP_t^\perp$ has total mass of order $o(t^2)$, part of which gets absorbed into $Q_t^\perp$. The part of $SP_t^\perp$ that is dominated by Q_0 contributes an extra nonnegative term, $\gamma_t(y)$, to the density $dQ_t^{(abs)}/dQ_0$. The $\eta_t^2(y)$ becomes $\kappa_y\xi_t^2(y) + \gamma_t(y)$. The extra term causes little trouble because

$$\sqrt{\kappa_y\xi_t^2} \le \eta_t \le \sqrt{\kappa_y\xi_t^2} + \sqrt{\gamma_t} \qquad \text{and} \qquad Q_0\gamma_t = o(t^2).$$

Proof of assertion (ii). Write $\mathbb{H}$ for the closed subspace of $L^2(P_\theta)$ consisting of (equivalence classes of) functions measurable with respect to the sub-sigma-field of $\mathcal{A}$ generated by S. Each member of $\mathbb{H}$ is of the form $f(Sx)$ for some f in $L^2(Q_\theta)$. The orthogonal projection of Δ_θ onto $\mathbb{H}$ equals $\widetilde{\Delta}_\theta(Sx)$. Thus

$$\mathbb{I}_{\mathcal{P}}(\theta) = P_\theta\Delta_\theta(x)^2 = P_\theta\widetilde{\Delta}_\theta(Sx)^2 + P_\theta\left[\Delta_\theta(x) - \widetilde{\Delta}_\theta(Sx)\right]^2.$$

The first term on the right-hand side equals $Q_\theta(\widetilde{\Delta}_\theta^2) = \mathbb{I}_{\mathcal{Q}}$; the last term is zero if and only iff $\Delta_\theta(x) = \widetilde{\Delta}_\theta(Sx)$ a.e.$[P_\theta]$.

References.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (L. LE CAM, J. NEYMAN and E. L. SCOTT, eds.) **I** 175–194. University of California Press, Berkeley.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory.* Springer, New York. (English translation from 1979 Russian edition).

KAGAN, A. and SHEPP, L. A. (2005). A sufficiency paradox: an insufficient statistic preserving the Fisher information. *The American Statistician* **59** 54–56.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

Le Cam, L. and Yang, G. L. (1988). On the preservation of local asymptotic normality under information loss. *Annals of Statistics* **16** 483–520.

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, 2nd ed. Springer-Verlag.

Pitman, E. J. G. (1979). *Some Basic Theory for Statistical Inference*. *Monographs on applied probability and statistics*. Chapman and Hall.

van der Vaart, A. (1988). *Statistical estimation in large parameter spaces*. Center for Mathematics and Computer Science CWI Tract 44.

Statistics Department
Yale University
New Haven, Connecticut