Chapter 7

Hellinger differentiability

SECTION *1 relates Hellinger differentiability to the classical regularity conditions for maximum likelihood theory.*

SECTION *2 discusses connections between Hellinger differentiability and pointwise differentiability of densities, leading to a sufficient condition for Hellinger differentiability.*

SECTION *3 derives the information inequality, as an illustration of the elegance brought into statistical theory by Hellinger differentiability.*

SECTION *4 explains how one can dispense with the domination assumption when defining Hellinger differentiability, at the cost of a natural extra assumption regarding non-dominated components. The slightly strengthened concept is called Differentiability in Quadratic Mean (DQM) to avoid confusion.*

SECTION *5 shows that DQM is preserved under measurable maps.*

Final two sections not yet edited.

Preliminary version. Editing in progress.

7.1 Heuristics

DQM::heuristics

The traditional regularity conditions for asymptotic statistical theory involve existence of two or three derivatives of density functions, together with domination assumptions to justify differentiation under integral signs. Le Cam (1970) noted that such conditions are unnecessarily stringent. He commented:

> Even if one is not interested in the maximum economy of assumptions one cannot escape practical statistical problems in which apparently "slight" violations of the assumptions occur. For instance the derivatives fail to exist at one point x which may depend on θ, or the distributions may not be mutually absolutely continuous or a variety of other difficulties may occur. The existing literature is rather unclear about what may happen in these circumstances. Note also that since the conditions are imposed upon probability densities they may be satisfied for one choice of such densities but not for certain other choices.

Probably Le Cam had in mind examples such as the double exponential density, $1/2\exp(-|x-\theta|)$, for which differentiability fails at the point $\theta = x$. He showed that the traditional conditions can, for some purposes, be replaced by a simpler assumption of ***Hellinger differentiability***: differentiability in norm of the square root of the density as an element of an $\mathcal{L}^2$ space.

norm.diff <1> **Definition.** *Write $\mathcal{L}^1_+(\lambda)$ for the set of nonnegative functions that are integrable with respect to a sigma-finite measure λ.*

*Say that a set $\mathcal{F} = \{f_\theta : \theta \in \Theta\} \subseteq \mathcal{L}^1_+(\lambda)$, indexed by a subset Θ of $\mathbb{R}^k$, is **Hellinger differentiable** at a point θ_0 of Θ if the map $\theta \mapsto \xi_\theta(x) := \sqrt{f_\theta(x)}$ is differentiable in $\mathcal{L}^2(\lambda)$ norm at θ_0, that is, if there exists a k-dimensional vector $\overset{\bullet}{\xi}_{\theta_0}(x)$ of functions in $\mathcal{L}^2(\lambda)$ such that*

hell.diff<2>
$$\xi_\theta(x) = \xi_{\theta_0}(x) + (\theta-\theta_0)'\overset{\bullet}{\xi}_{\theta_0}(x) + r_\theta(x) \qquad \textit{with } \|r_\theta\|_2 = o(|\theta-\theta_0|) \textit{ near } \theta_0.$$

Call $\overset{\bullet}{\xi}_{\theta_0}(x)$ the Hellinger derivative at θ_0.

In particular, if $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability measures dominated by λ, say that $\mathcal{P}$ is Hellinger differentiable at θ_0 if the set of densities $\{dP_\theta/d\lambda : \theta \in \Theta\}$ is Hellinger differentiable at θ_0.

Remark. Some authors—see, for example, Bickel, Klaassen, Ritov, and Wellner (1993, page 202)—adopt a slightly different definition,

hell.diff2<3>
$$\xi_\theta(x) = \xi_{\theta_0}(x) + \tfrac{1}{2}(\theta-\theta_0)'\Delta(x)\xi_{\theta_0}(x) + r_\theta(x),$$

replacing the Hellinger derivative $\overset{\bullet}{\xi}_{\theta_0}$ by $\frac{1}{2}\Delta(x)\xi_{\theta_0}(x)$. As explained in Section 4, the modification very cleverly adds an extra implicit regularity assumption to the definition, by requiring that $\overset{\bullet}{\xi}_{\theta_0}(x) = 0$ when $\xi_{\theta_0}(x) = 0$. The two definitions are not completely equivalent.

Classical statistical theory, especially when dealing with independent observations from a P_θ, makes heavy use of the function $\ell_\theta(x) := \log p_\theta(x)$, where $p_\theta = dP_\theta/d\lambda$. The vector $\overset{\bullet}{\ell}_\theta(x)$ of partial derivatives with respect to θ is called the ***score function***. The variance matrix $\mathbb{I}_\theta$ of the score function is called the ***Fisher information matrix*** for the model. The classical regularity conditions justify differentiation under the integral sign to get

zero.deriv<4>
$$P_\theta \overset{\bullet}{\ell}_\theta(x) = \lambda \overset{\bullet}{p}_\theta(x) = \frac{\partial}{\partial\theta}\lambda p_\theta(x) = 0,$$

whence $\mathbb{I}_\theta := \text{var}_\theta\left(\overset{\bullet}{\ell}_\theta\right) = P_\theta\left(\overset{\bullet}{\ell}_\theta \overset{\bullet}{\ell}{}_\theta'\right)$.

Under assumptions of Hellinger differentiability, the derivative $\overset{\bullet}{\xi}_\theta$ takes over the role of the score vector. Ignoring problems related to division by zero and distinctions between pointwise and $\mathcal{L}^2(\lambda)$ differentiability, we would have

$$\frac{2\xi^\bullet_\theta(x)}{\xi_\theta(x)} \overset{?}{=} \frac{2}{\sqrt{f_\theta(x)}}\frac{\partial}{\partial\theta}\sqrt{f_\theta(x)} = \frac{1}{f_\theta(x)}\frac{\partial f_\theta(x)}{\partial\theta} = \ell^\bullet_\theta(x).$$

Thus the Δ in the modified definition <3> corresponds to the score function.

The equality <4> corresponds to the assertion $P_\theta\left(\xi^\bullet_\theta/\xi_\theta\right) = \lambda\left(\xi_\theta\xi^\bullet\right) = 0$, which Section **??** will show to be a consequence of Hellinger differentiability and the fact that $\|\xi_\theta\|_2 =$ for all θ. The Fisher information $\mathbb{I}_\theta$ at θ corresponds to the matrix

$$P_{\theta_0}\left(\ell^\bullet_\theta\ell^{\bullet\prime}_\theta\right) \overset{?}{=} 4P_{\theta_0}\left(\xi^\bullet_\theta\xi^{\bullet\prime}_\theta/\xi^2_\theta\right) \overset{?}{=} 4\lambda\left(\xi^\bullet_\theta\xi^{\bullet\prime}_\theta\right).$$

Here I flag both equalities as slightly suspect, not just for the unsupported assumption of equivalence between pointwise and Hellinger differentiabilities, but also because of a possible 0/0 cancellation. For the moment it is better to insert an explicit indicator function, $\{\xi_\theta > 0\}$, to protect against 0/0. To avoid possible ambiguity or confusion, I will write $\mathbb{I}_\theta$ for $4\lambda(\xi^\bullet_\theta\xi^{\bullet\prime}_\theta)$ and $\mathbb{I}^\circ_\theta$ for $4\lambda(\xi^\bullet_\theta\xi^{\bullet\prime}_\theta\{\xi_\theta > 0\})$, to hint at equivalent forms for $\mathbb{I}_\theta$ without yet giving precise conditions under which all three exist and are equal. See Section **??** for an explanation of when the distinction is necessary.

The classical assumptions also justify further interchanges of integrals and derivatives, to derive an alternative representation $\mathbb{I}_\theta = -\mathbb{P}_\theta\ell^{\bullet\bullet}_\theta$ for the information matrix. It might seem obvious that there can be no analog of this representation for Hellinger differentiability. Indeed, how could an assumption of one-times differentiability, in norm, imply anything about a second derivative? Surprisingly, there is a way, if we think of second derivatives as coefficients of quadratic terms in local approximations. As will be shown in Section **??**, the fact that $\|\xi_\theta\|_2 = 1$ for all θ leads to a quadratic approximation for a log-likelihood ratio—a sort of Taylor expansion to quadratic terms without the usual assumption of twice continuous differentiability. Remarkable.

7.2 A sufficient condition for Hellinger differentiability

DQM::pwise

How does Hellinger differentiability relate to the classical assumption of pointwise differentiability?

Roughly speaking, the difference between the two concepts is like the difference between convergence in $\mathcal{L}^2$ and convergence almost surely. In fact, it is easy (Problem [1]) to adapt a standard counterexample to show that Hellinger differentiability does not imply pointwise differentiability.

Consider the case where Θ is one-dimensional, and f_θ is both Hellinger differentiable and differentiable a.e. $[\lambda]$ at $\theta = 0$. Choose a sequence $\{\theta_n\}$ tending to zero so fast that $\sum_n \|r_{\theta_n}\|_2/|\theta_n| < \infty$, which implies $r_{\theta_n}(x) = o(|\theta_n|)$ a.e. $[\lambda]$. For almost all x,

$$\begin{aligned}
\xi_{\theta_n}(x) &= \xi_0(x) + \theta_n \xi^\bullet_0(x) + o(|\theta_n|)\\
\xi_{\theta_n}(x)^2 &= \xi_0(x)^2 + \theta_n f_0'(x) + o(|\theta_n|).
\end{aligned}$$

If $f_0(x) \neq 0$, the second equation can be rewritten as

$$\xi_{\theta_n}(x) = \xi_0(x)\left(1 + \theta_n \frac{f_0'(x)}{\xi_0(x)^2} + o(|\theta_n|)\right)^{1/2} = \xi_0(x) + \tfrac{1}{2}\theta_n \frac{f_0'(x)}{\xi_0(x)} + o(|\theta_n|).$$

It follows (cf. differentiation of $\sqrt{f_\theta(x)}$ by first principles) that $f_0'(x) = 2\xi_0(x)\xi^\bullet_0(x)$. At an x where $f_0(x) = 0$, this argument fails. Instead we would have

$$\begin{aligned}
\xi_{\theta_n}(x)^2 &= \theta_n^2 \xi^\bullet_0(x)^2 + o(|\theta_n|^2)\\
\xi_{\theta_n}(x)^2 &= \theta_n f_0'(x) + o(|\theta_n|).
\end{aligned}$$

We then deduce that $f_0'(x) = 0$ but apparently we no longer have any control over $\xi^\bullet_0(x)$. However, if 0 is an interior point of the parameter space Θ we could repeat the argument with $\{\theta_n\}$ replaced by $\{-\theta_n\}$, obtaining for almost all x for which $\xi_0(x) = 0$ that

$$\xi_{\pm\theta_n}(x) = \pm\theta_n \xi^\bullet_0(x) + o(|\theta_n|).$$

Nonnegativity of ξ_θ would then force $\xi^\bullet_0(x) = 0$.

In summary: If the $f_\theta(x)$ are pointwise differentiable at $\theta = 0$ for almost all x and if 0 is an interior point of Θ then the only possible candidate (up to an almost sure equivalence) for the Hellinger derivative at 0 is

$$\xi^\bullet_0(x) = \tfrac{1}{2}\frac{f_0'(x)}{\xi_0(x)}$$

What more do we need in order to show that this $\xi^\bullet_0$ is, in fact, an $\mathcal{L}^2(\lambda)$ derivative of θ_θ at $\theta = 0$? The answer requires careful attention to the problem of when functions of a real variable can be recovered as integrals of their derivatives.

abs.cty.def <5> **Definition.** *A real valued function H defined on an interval $[a,b]$ of the real line is said to be* ***absolutely continuous*** *if to each $\epsilon > 0$ there exists a $\delta > 0$ such that $\sum_i |H(b_i) - H(a_i)| < \epsilon$ for all finite collections of nonoverlapping subintervals $[a_i, b_i]$ of $[a,b]$ for which $\sum_i (b_i - a_i) < \delta$.*

Absolute continuity of a function defined on the whole real line is taken to mean absolute continuity on each finite subinterval.

The following connection between absolute continuity and integration of derivatives is one of the most celebrated results of classical analysis (UGMTP §3.4).

fundamental <6> **Theorem.** *A real valued function H defined on an interval $[a,b]$ is absolutely continuous if and only if the following three conditions hold.*

(i) The derivative $H'(t)$ exists at Lebesgue almost all points of $[a,b]$.

(ii) The derivative H' is Lebesgue integrable

(iii) $H(t) - H(a) = \int_a^t H'(s)\,ds$ for each t in $[a,b]$

Put another way, a function H is absolutely continuous on an interval $[a,b]$ if and only if there exists an integrable function h for which

ac.integral<7>
$$H(t) = \int_a^t h(s)\,ds \qquad \text{for all } t \text{ in } [a,b]$$

The function H must then have derivative $h(t)$ at almost all t. As a systematic convention we could take h equal to the measurable function

$$H^\bullet(t) = \begin{cases} H'(t) & \text{at points } t \text{ where the derivative exists,} \\ 0 & \text{elsewhere.} \end{cases}$$

I will refer to $H^\bullet$ as the ***density*** of H. Of course it is actually immaterial how $H^\bullet$ is defined on the Lebesgue negligible set of points at which the derivative does not exist, but the convention helps to avoid ambiguity.

Now consider a *nonnegative* function H that is differentiable at a point t. If $H(t) > 0$ then the chain rule of elementary calculus implies that the function $2\sqrt{H}$ is also differentiable at t, with derivative $H'(t)/\sqrt{H(t)}$. At points where $H(t) = 0$, the question of differentiability becomes more delicate, because the map $y \mapsto \sqrt{y}$ is not differentiable at the origin. If t is an internal point of the interval and $H(t) = 0$ then we must have $H'(t) = 0$. Thus $H(y) = o(|y-t|)$ near t. If $\sqrt{H}$ had a derivative at t then $\sqrt{H(y)} = o(|y-t|)$ near t, and hence $H(y) = o(|y-t|^2)$. Clearly we need to take some care with the question of differentiability at points where H equals zero.

Even more delicate is the fact that absolute continuity of a nonnegative function H need not imply absolute continuity of the function $\sqrt{H}$, without further assumptions—even if H is everywhere differentiable (Problem [2]).

sqrt.ac <8> **Lemma.** *Suppose a nonnegative function H is absolutely continuous on an interval $[a,b]$, with density $H^\bullet$. Let $\Delta(t) := {}^{1\!}/_{2} H^\bullet(t)\{H(t)>0\}/\sqrt{H(t)}$. If $\int_a^b |\Delta(t)|\,dx < \infty$ then $\sqrt{H}$ is absolutely continuous, with density Δ, that is,*

$$\sqrt{H(t)} - \sqrt{H(a)} = \int_a^t \Delta(s)\,ds \qquad \textit{for all } t \textit{ in } [a,b]$$

Proof Fix an $\eta > 0$. The function $H_\eta := \eta + H$ is bounded away from zero, and hence $\sqrt{H_\eta}$ has derivative $H_\eta' = H'/(2\sqrt{H+\eta})$ at each point where the derivative H' exists. Moreover, absolute continuity of $\sqrt{H_\eta}$ follows directly from the Definition <5>, because

$$|\sqrt{H_\eta(b_i)} - \sqrt{H_\eta(a_i)}| = \frac{|H_\eta(b_i) - H_\eta(a_i)|}{\sqrt{H_\eta(b_i)} + \sqrt{H_\eta(a_i)}} \le \frac{|H(b_i) - H(a_i)|}{2\sqrt{\eta}}$$

for each interval $[a_i, b_i]$. From Theorem <6>, for each t in $[a,b]$,

$$\sqrt{H(t)+\eta} - \sqrt{H(a)+\eta} = \int_a^t \frac{H^\bullet(s)}{2\sqrt{H(s)+\eta}}\,ds.$$

As η decreases to zero, the left-hand side converges to $\sqrt{H(t)} - \sqrt{H(a)}$. The integrand on the right-hand side converges to $\Delta(s)$ at points where $H(s) > 0$. For almost all s in $\{H = 0\}$ the derivative $H'(s)$ exists and equals zero; the integrand converges to $0 = \Delta(s)$ at those points. By Dominated Convergence, the right-hand side converges to $\int_a^t \Delta(s)\,ds$.

□

The integral representation for the square root of an absolutely continuous function is often the key to proofs of Hellinger differentiability. For simplicity of notation, the following sufficient condition is stated only for a one-dimensional Θ with 0 as an interior point.

Hdiff.suff <9> **Theorem.** *Suppose $\mathcal{F} = \{f_\theta(x) : |\theta| < \delta\} \subseteq \mathcal{L}^1_+(\lambda)$ for some $\delta > 0$. Suppose also that*

(i) the map $(x,\theta) \mapsto f_\theta(x)$ is product measurable;

(ii) for λ almost all x, the function $\theta \mapsto f_\theta(x)$ is absolutely continuous on $[-\delta,\delta]$, with almost sure derivative $f^\bullet_\theta(x)$;

(iii) for λ almost all x, the function $\theta \mapsto f_\theta(x)$ is differentiable at $\theta = 0$;

(iv) for each θ the function $\xi^\bullet_\theta(x) := \frac{1}{2} f^\bullet_\theta(x)\{f_\theta(x) > 0\}/\sqrt{f_\theta(x)}$ belongs to $\mathcal{L}^2(\lambda)$ and $\lambda(\xi^\bullet_\theta)^2 \to \lambda(\xi^\bullet_0)^2$ as $\theta \to 0$.

Then $\mathcal{F}$ has Hellinger derivative $\xi^\bullet_0(x)$ at $\theta = 0$.

> **Remark.** Assumption (iii) might appear redundant, because (ii) implies differentiability of $\theta \mapsto f_\theta(x)$ at Lebesgue almost all θ, for λ-almost all x. A mathematical optimist (or Bayesian) might be prepared to gamble that 0 does not belong to the bad negligible set; a mathematical pessimist might prefer Assumption (iii).

Proof As before, write $\xi_\theta(x)$ for $\sqrt{f_\theta(x)}$ and define $r_\theta(x) := \xi_\theta(x) - \xi_0(x) - \theta\xi^\bullet_0(x)$. We need to prove that $\lambda r^2_\theta = o(|\theta|^2)$ as $\theta \to 0$.

Assumption (i) and the convention about densities imply joint measurability of $(x, \theta) \mapsto f^\bullet_\theta(x)$.

For simplicity of notation, consider only positive θ. The arguments for negative θ are analogous. Write $\mathfrak{m}$ for Lebesgue measure on $[-\delta, \delta]$.

With no loss of generality (or by a suitable decrease in δ) we may assume that $\lambda(\xi^\bullet_\theta)^2$ is bounded, so that, by Tonelli, $\infty > \mathfrak{m}^\theta\lambda^x(\xi^\bullet_\theta(x))^2 = \lambda^x\mathfrak{m}^\theta(\xi^\bullet_\theta(x))^2$, implying $\mathfrak{m}^\theta(\xi^\bullet_\theta(x))^2 < \infty$ a.e. $[\lambda]$. From Lemma <8> it then follows that

$$\frac{\xi_\theta(x) - \xi_0(x)}{\theta} = \frac{1}{\theta}\int_0^\theta \xi^\bullet_s(x)\,ds \qquad \text{a.e. } [\lambda].$$

By Jensen's inequality for the uniform distribution on $[0, \theta]$, and (iv),

limsup.diff<10>

$$\lambda\left|\frac{\xi_\theta(x) - \xi_0(x)}{\theta}\right|^2 \le \frac{1}{\theta}\int_0^\theta \lambda\xi^\bullet_s(x)^2\,ds \to \lambda(\xi^\bullet_0)^2 \qquad \text{as } \theta \to 0.$$

Define nonnegative, measurable functions

$$g_\theta(x) := 2\left|\xi_\theta(x) - \xi_0(x)\right|^2/\theta^2 + 2\xi^\bullet_0(x)^2 - \left|r_\theta(x)/\theta\right|^2.$$

By (iii), $r_\theta(x)/\theta \to 0$ at almost all x where $\xi_0(x) > 0$, and hence $g_\theta(x) \to 4\xi^\bullet_0(x)^2$. At almost all points where $\xi_0(x) = 0$ we have $\xi^\bullet_0(x) = 0$, so that $\xi_\theta(x) = r_\theta(x)$ and $g_\theta(x) \ge 0$. Thus $\liminf g_\theta(x) \ge 4\xi^\bullet_0(x)^2$ a.e. $[\lambda]$. By Fatou's Lemma (applied along subsequences), followed by an appeal to <10>,

$$4\lambda(\xi^\bullet_0)^2 \le \liminf_{\theta\to 0} \lambda g_\theta \le 4\lambda(\xi^\bullet_0)^2 - \limsup_{\theta\to 0} \lambda\left|r_\theta(x)/\theta\right|^2.$$

That is, $\lambda r^2_\theta = o(\theta^2)$, as required for Hellinger differentiability.

□

shift.family <11> **Example.** Let q be a probability density with respect to Lebesgue measure $\mathfrak{m}$ on the real line. Suppose q is absolutely continuous, with density $q^\bullet$ for which $\mathbb{I}_q := \mathfrak{m}\left(\{q > 0\}q^{\bullet 2}/q\right) < \infty$. Define Q_θ to have density $f_\theta(x) := q(x-\theta)$ with respect to λ, for each θ in $\mathbb{R}$. The conditions of Theorem <9> are satisfied, with

$$\xi_\theta^\bullet(x) = -\tfrac{1}{2}\frac{q^\bullet(x-\theta)}{\sqrt{q(x-\theta)}}\{q(x-\theta) > 0\} \qquad \text{and} \qquad 4\mathfrak{m}(\xi_\theta^\bullet)^2 \equiv \mathbb{I}_q.$$

The family $\mathcal{Q} := \{Q_\theta : \theta \in \mathbb{R}\}$ is Hellinger differentiable at $\theta = 0$. In fact, the same argument works at every θ; the family is everywhere Hellinger differentiable, with Hellinger derivative $\xi_\theta^\bullet$ at θ.

It is traditional to call $\mathbb{I}_q$ the Fisher information for q, even though it would be more more precise to call it the Fisher information for the shift family generated by q.

□

7.3 Differentiability of unit vectors

DQM::unit.vector

Suppose τ is a map from $\mathbb{R}^k$ into some inner product space $\mathcal{H}$ (such as $\mathcal{L}^2(\lambda)$). Suppose also that τ is differentiable (in norm) at θ_0,

$$\tau_\theta = \tau_{\theta_0} + (\theta - \theta_0)'\tau_{\theta_0}^\bullet + r_\theta \qquad \text{with } \|r_\theta\| = o(|\theta - \theta_0|) \text{ near } \theta_0.$$

For simplicity of notation, suppose $\theta_0 = 0$.

The Cauchy-Schwarz inequality gives $|\langle \tau_0, r_\theta\rangle| \le \|\tau_0\|\,\|r_\theta\| = o(|\theta|)$. It would usually be a blunder to assume naively that the bound must therefore be of order $O(|\theta|^2)$; typically, higher-order differentiability assumptions are needed to derive approximations with smaller errors. However, if $\|\tau_\theta\|$ is constant—that is, if τ_θ is constrained to take values lying on the surface of a sphere—then the naive assumption turns out to be no blunder. Indeed, in that case, it is easy to show that in general $\langle \tau_0, r_\theta\rangle$ equals a quadratic in θ plus an error of order $o(|\theta|^2)$. The sequential form of the assertion will be more convenient for the calculations in Section **??**.

UNITvector <12> **Lemma.** *Let $\{\alpha_n\}$ be a sequence of constants tending to zero. Let τ_0, τ_1, ... be elements of norm one for which $\tau_n = \tau_0 + \alpha_n W + \rho_n$, with W a fixed element of $\mathcal{H}$ and $\|\rho_n\| = o(\alpha_n)$. Then $\langle \tau_0, W\rangle = 0$ and $2\langle \tau_0, \rho_n\rangle = -\alpha_n^2\|W\|^2 + o(\alpha_n^2)$.*

Proof Because both τ_n and τ_0 have unit length,

$$\begin{aligned} 0 = \|\tau_n\|^2 - \|\tau_0\|^2 = 2\alpha_n\langle\tau_0, W\rangle & \qquad\qquad \text{order } O(\alpha_n) \\ + 2\langle\tau_0, \rho_n\rangle & \qquad\qquad \text{order } o(\alpha_n) \\ + \alpha_n^2\|W\|^2 & \qquad\qquad \text{order } O(\alpha_n^2) \\ + 2\alpha_n\langle W, \rho_n\rangle + \|\rho_n\|^2 & \qquad\qquad \text{order } o(\alpha_n^2). \end{aligned}$$

The $o(\alpha_n)$ and $o(\alpha_n^2)$ rates of convergence in the second and fourth lines come from the Cauchy-Schwarz inequality. The exact zero on the left-hand side of the equality exposes the leading $2\alpha_n\langle\tau_0, W\rangle$ as the only $O(\alpha_n)$ term on the right-hand side. It must be of smaller order, $o(\alpha_n)$ like the other terms, which can happen only if $\langle\tau_0, W\rangle = 0$, leaving

$$0 = 2\langle\tau_0, \rho_n\rangle + \alpha_n^2\|W\|^2 + o(\alpha_n^2),$$

as asserted.

□

> **Remark.** Without the fixed length property, the difference $\|\tau_n\|^2 - \|\tau_0\|^2$ might contain terms of order α_n. The inner product $\langle\tau_0, \rho_n\rangle$, which inherits $o(\alpha_n)$ behaviour from $\|\rho_n\|$, might then not decrease at the $O(\alpha_n^2)$ rate.

unit2 <13> **Corollary.** *If $\mathcal{P}$ has a Hellinger derivative $\xi^\bullet_{\theta_0}$ at 0, and if 0 is an interior point of Θ, then $\lambda\left(\xi_0\xi_0^\bullet\right) = 0$ and $8\lambda\left(\xi_0 r_\theta\right) = -\theta'\mathbb{I}_0\theta + o(|\theta|^2)$ near 0.*

Proof Start with the second assertion, in its equivalent form for sequences $\theta_n \to 0$. Write θ_n as $|\theta_n|u_n$, with u_n a unit vector in $\mathbb{R}^k$. By a subsequencing argument, we may assume that $u_n \to u$, in which case,

$$\xi_{\theta_n} = \xi_0 + |\theta_n|u_n'\xi_0^\bullet + r_{\theta_n} = \xi_0 + |\theta_n|u'\xi_0^\bullet + \left(r_{\theta_n} + |\theta_n|(u_n - u)'\xi_0^\bullet\right).$$

Invoke the Lemma (with $W = u'\xi_0^\bullet$) to deduce that $u'\lambda\left(\xi_0\xi_0^\bullet\right) = 0$ and

$$\begin{aligned} -4|\theta_n|^2\lambda\left(u'\xi_0^\bullet\right)^2 + o(|\theta_n|^2) &= 8\lambda\left(\xi_0\left(r_{\theta_n} + |\theta_n|(u_n - u)'\xi_0^\bullet\right)\right) \\ &= 8\lambda\left(\xi_0 r_{\theta_n}\right) + 8|\theta_n|(u_n - u)'\lambda\left(\xi_0\xi_0^\bullet\right). \end{aligned}$$

Because 0 is an interior point, for every unit vector u there are sequences $\theta_n \to 0$ through Θ for which $u = \theta_n/|\theta_n|$. Thus $u'\lambda\left(\xi_0\xi_0^\bullet\right) = 0$ for every unit vector u, implying that $\lambda\left(\xi_0\xi_0^\bullet\right) = 0$. The last displayed equation reduces the sequential analog of the asserted approximation.

□

Remark. If 0 were not an interior point of the parameter space, there might not be enough directions u along which $\theta_n \to 0$ through Θ, and it might not follow that $\lambda(\xi_0 \dot{\xi}_0) = 0$. Roughly speaking, the set of such directions is called the ***contingent*** of Θ at θ_0. If the contingent is 'rich enough', we do not need to assume that 0 is an interior point. See Le Cam and Yang (2000, Section 7.2) and Le Cam (1986, page 575) for further details.

7.4 Information inequality

DQM::info

The information inequality for the model $\mathcal{P} := \{P_\theta : \theta \in \Theta\}$ bounds the variance of a statistic $T(x)$ from below by an expression involving the expected value of the statistic and the Fisher information: under suitable regularity conditions,

$$\text{var}_\theta(T) \geq \dot{\gamma}_\theta' \mathbb{I}_\theta^{-1} \dot{\gamma}_\theta \qquad \text{where } \gamma_\theta := P_\theta T(x) \text{ and } \dot{\gamma}_\theta := \frac{d}{d\theta}\gamma_\theta.$$

The classical proof of the inequality imposes assumptions that derivatives can be passed inside integral signs, typically justified by more primitive assumptions involving pointwise differentiability of densities and domination assumptions about their derivatives.

By contrast, the proof of the information inequality based on an assumption of Hellinger differentiability replaces the classical requirements by simple properties of $\mathcal{L}^2(\lambda)$ norms and inner products. The gain in elegance and economy of assumptions illustrates the typical benefits of working with Hellinger differentiability. The main technical ideas are captured by the following Lemma. Once again, with no loss of generality I consider only behavior at $\theta = 0$.

Remark. The measure P_θ might itself be a product measure, representing the joint distribution of a sample of independent observations from some distribution μ_θ. As shown by Problem [5], Hellinger differentiability of $\theta \mapsto \mu_\theta$ at $\theta = 0$ would then imply Hellinger differentiability of $\theta \mapsto P_\theta$ at $\theta = 0$. We could substitute an explicit product measure for P_θ in the next Lemma, but there would be no advantage to doing so.

Tdiff<14> **Lemma.** *Suppose a dominated family $\mathcal{P}$ has Hellinger derivative $\dot{\xi}_0$ at 0 and that $\sup_{\theta \in U} P_\theta T(x)^2 < \infty$, for some neighborhood U of 0. Then the function $\theta \mapsto \gamma_\theta := P_\theta^x T(x)$ has derivative $\dot{\gamma}_0 = 2\lambda(\xi_0 \dot{\xi}_0 T)$ at 0.*

Remark. Notice that $P_\theta T$ is well defined throughout U, because of the bound on the second moment. Also $\left(\lambda|\xi_0 \dot{\xi}_0 T|\right)^2 \leq \left(\lambda \xi_0^2 T^2\right)\left(\lambda|\dot{\xi}_0|^2\right) < \infty$.

Proof Write C^2 for $\sup_{\theta\in U} P_\theta T(x)^2$, so that $\|\xi_\theta T\|_2 \leq C$ for each θ in U. For simplicity, I consider only the one-dimensional case. The proof for $\mathbb{R}^k$ differs only notationally.

□