Bernoulli **11**(2), 2005, 263–282

Cramér–Rao revisited

ANDRIES J. LENSTRA

Department of Mathematics, Oberlin College, Oberlin OH 44074, USA
E-mail: andries.lenstra@oberlin.edu

In a right-angled triangle, the hypotenuse is the longest side. So, if all (hypotenuse) vectors from a given set of vectors have the same orthogonal projection onto a certain subspace, we have a lower bound for their lengths. Interpreting the square of such a length as the variance of an unbiased estimator produces an information bound. The Cramér–Rao bound and the van Trees inequality can be seen as consequences of this bound. Another consequence is an inequality for the minimax variance, that is, the maximal variance in shrinking neighbourhoods, minimized over all unbiased estimators. This bound is non-asymptotic and requires almost no regularity conditions.

Keywords: Cramér–Rao inequality; nonparametric information bounds; tangential differentiation; unbiased estimation; van Trees inequality

1. Summary and introduction

There are viewpoints from which the existence of bounds for the precision of unbiased estimation is obvious. In Section 2 we present such a viewpoint and the corresponding bound; in Section 3 we examine the relation between that bound and the number of independent draws, and see the use of taking the mean. The question for which unbiased estimators our bound holds is answered in Section 4; the answer is such that a new information bound appears, which holds for all unbiased estimators. In Section 5 the empirical distribution function is found to be optimal for the new criterion. Next we connect our viewpoint with the customary viewpoint via properties of the parametrization map that are often present but seldom used, if ever. The Cramér–Rao bound then follows, in Section 6, from the bound in Section 2 and the chain rule; that is why the Fisher information is inverted. Once the right setting has been established, the same devices give, in Section 7, the van Trees inequality. A condition that makes our van Trees proof work is contained in Section 8; the Appendix explains the differentiability concept we use.

The difference between the present approach and the usual set-up is that, roughly speaking, in the latter one differentiates $\vartheta \mapsto P_\vartheta$, while in the former it is $P_\vartheta \mapsto \vartheta$ which is differentiated. We are not the first to do this inversion; it is implicit in, for example, Klaassen *et al.* (1988). For more references see Janssen (2003), which itself is a demonstration of recent interest. New are, we think, the extent to which this approach has been simplified and exploited in an almost self-contained exposition, and – as hinted at above – the supplement we provide to the inversion of the parametrization map: the inverse function theorem for Banach spaces.

2. One draw

Let $\mathcal{P}$ be a set of probability measures on the measurable space $(\mathcal{X}, \mathcal{A})$ and $\kappa : \mathcal{P} \to \mathbb{R}$ be a *parameter of interest*. If, for a measurable map $t : (\mathcal{X}, \mathcal{A}) \to (\mathbb{R},$ Borel sets of $\mathbb{R})$, we have

$$\int_{\mathcal{X}} t(x)\,\mathrm{d}P(x) = \kappa(P) \qquad \forall P \in \mathcal{P},$$

then t is called an *unbiased estimator for* κ. Let X, given $P \in \mathcal{P}$, be any random element of $\mathcal{X}$ with probability distribution P and abbreviate

$$T := t \circ X. \tag{1}$$

Then $\int_{\mathcal{X}} t(x)\,\mathrm{d}P(x)$ can be written as $\mathrm{E}_P T$, the P-expectation of T. In this section we show the existence of a lower bound $B_P \in \mathbb{R}$ for the numbers $\mathrm{var}_P T := \mathrm{E}_P((T - \mathrm{E}_P T)^2)$, $t \in \mathcal{T}_r$, where $\mathcal{T}_r$ is the set of unbiased estimators for κ that have the special property of being 'regular at P'. In Section 4 this 'highly undesirable' (cf. Fabian and Hannan 1977) restriction will be removed and a bound will emerge for *all* unbiased estimators.

The notion involved in the regularity is tangential differentiation. It is explained in the Appendix; the only things that matter in this section are that tangent spaces are closed linear subspaces of the whole normed linear space at hand, that tangential derivatives are maps on tangent spaces, and that no map has more than one tangential derivative at any point. Here is the definition of regularity.

Let μ be a measure on $(\mathcal{X}, \mathcal{A})$ and let $u \in \mathcal{U} \subset L^2_{\mathcal{X},\mu}$. A measurable $t : \mathcal{X} \to \mathbb{R}$ is [$\mathcal{U}$-]*regular at* u if there is a neighbourhood U of u in $\mathcal{U}$ such that

- for every $s \in U$ the map $ts : x \in \mathcal{X} \mapsto t(x)s(x)$ lies in $L^2_{\mathcal{X},\mu}$, and
- the map

  $$s \in U \subset \mathcal{U} \subset L^2_{\mathcal{X},\mu} \mapsto \langle ts, s\rangle \in \mathbb{R}$$

  (with $\langle a, b\rangle = \int_{\mathcal{X}} a(x)b(x)\,\mathrm{d}\mu(x)$, the inner product in the real Hilbert space $L^2_{\mathcal{X},\mu}$) is tangentially differentiable at u, while
- the tangential derivative of this map at u is represented by $2tu$, that is, it is equal to the map

  $$h \in \dot{\mathcal{U}}(u) \mapsto \langle h, 2tu\rangle,$$

  where $\dot{\mathcal{U}}(u)$ denotes the tangent space of $\mathcal{U}$ at u (just as the derivative at u of the map $s \in \mathbb{R} \mapsto cs^2$ is equal to the map $h \in \mathbb{R} \mapsto h \cdot 2cu$).

Theorem 1. *If* $(\mathcal{X}, \mathcal{A})$ *is a measurable space,* μ *a measure on* $(\mathcal{X}, \mathcal{A})$*,* $\mathcal{P}$ *a set of probability measures on* $(\mathcal{X}, \mathcal{A})$ *that have a density with respect to* μ*,* P *is a member of* $\mathcal{P}$*, we identify each* $Q \in \mathcal{P}$ *with the corresponding* (density of $Q)^{1/2} \in L^2_{\mathcal{X},\mu}$*,* $\kappa : \mathcal{P} \to \mathbb{R}$ *is a parameter of interest, and* $\mathcal{T}_r$ *is the set of the unbiased estimators for* κ *that are* $\mathcal{P}$*-regular at* P*, then all members of* $\{2(t + c)P \in L^2_{\mathcal{X},\mu} : t \in \mathcal{T}_r, c \in \mathbb{R}\}$ *have the same projection onto the tangent*

space $\dot{\mathcal{P}}(P)$ *of* $\mathcal{P}$ *at* P*, so that, with* $\mathrm{d}\kappa(P)/\mathrm{d}P$ *defined as equal to this unique projection for* $\mathcal{T}_r \neq \emptyset$*, we have*

$$\|2(t+c)P\|^2_{L^2_{\mathcal{X},\mu}} \geqslant \left\| \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P} \right\|^2_{L^2_{\mathcal{X},\mu}} \qquad \forall t \in \mathcal{T}_r,\ c \in \mathbb{R}, \tag{2}$$

and, in particular,

$$\mathrm{var}_P\, T \geqslant \frac{1}{4} \left\| \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P} \right\|^2_{L^2_{\mathcal{X},\mu}} \qquad \forall t \in \mathcal{T}_r. \tag{3}$$

Proof. If $t, t' \in \mathcal{T}_r$, then $2tP$ and $2t'P$ represent the same map on $\dot{\mathcal{P}}(P)$, namely, the tangential derivative at P of $Q \in \mathcal{P} \mapsto \langle tQ, Q\rangle = \kappa(Q) = \langle t'Q, Q\rangle$, and their projections onto $\dot{\mathcal{P}}(P)$ do the same; therefore, these projections coincide. Further, the definition of tangential differentiability implies that $Q \in \mathcal{P} \mapsto \langle cQ, Q\rangle = c$ has derivative $h \in \dot{\mathcal{P}}(P) \mapsto 0$ and that, as we shall show in Section 4, the map $x \in \mathcal{X} \mapsto c$ is regular at P. So $h \in \dot{\mathcal{P}}(P) \mapsto 0 \in \mathbb{R}$ is represented by $2cP$. It follows that P is perpendicular to $\dot{\mathcal{P}}(P)$. Now (2) results from Pythagoras and (3) from taking $c = -\mathrm{E}_P T$ in (2). □

We see that if $\mathcal{T}_r \neq \emptyset$, the map κ is tangentially differentiable at P and the vector $\mathrm{d}\kappa(P)/\mathrm{d}P$ is the unique member of $\dot{\mathcal{P}}(P)$ representing the derivative at hand.

This section concludes with an observation in the case where the tangent spaces are large and a picture for the case where they are not.

In the situation of Theorem 1 we call the set $\mathcal{P}$ of probability measures *nonparametric* (for a different definition, see Groeneboom and Wellner 1992) if, for every measurable $g : \mathcal{X} \to \mathbb{R}$ with $gQ \in L^2_{\mathcal{X},\mu}$ for all $Q \in \mathcal{P}$, it is true that

$$gQ \perp Q \Rightarrow gQ \in \dot{\mathcal{P}}(Q) \qquad \forall Q \in \mathcal{P}.$$

Let $\mathcal{T}_R$ be the set of unbiased estimators for κ that are *regular*, that is, regular everywhere. For any $Q \in \mathcal{P}$, $t, t' \in \mathcal{T}_R$ we have $2(t - \kappa(Q))Q \perp Q$ and

$$(t-t')Q \perp \begin{cases} Q & \text{because both } t \text{ and } t' \text{ are unbiased,} \\ \dot{\mathcal{P}}(Q) & \text{because the projection on } \dot{\mathcal{P}}(Q) \text{ is zero.} \end{cases}$$

Indeed, if $\mathcal{P}$ is nonparametric, the implications are

$$tQ \overset{L^2_{\mathcal{X},\mu}}{=} t'Q \quad \text{and} \quad 2(t-\kappa(Q))Q \overset{L^2_{\mathcal{X},\mu}}{=} \frac{\mathrm{d}\kappa(Q)}{\mathrm{d}Q} \qquad \forall t, t' \in \mathcal{T}_R,\ \forall Q \in \mathcal{P}.$$

These findings we express as follows:

Proposition 2. *In nonparametric models, regular unbiased estimators are essentially unique and attain the bound* (3) *everywhere.*

In Section 5 we give an example. Observe that for a nonparametric $\mathcal{P}$ and $n > 1$, the set

$\mathcal{P}^{(n)} := \{Q^n : Q \in \mathcal{P}\}$ of probability measures on $\mathcal{X}^n$ need *not* be nonparametric, if only because estimating on the basis of a subsample will, in general, increase the variance, while not necessarily being less regular; see Section 5. From the next sections we shall also learn that in estimating on the basis of a draw from P^n 'the mean of optimal is optimal'.

Figure 1 illustrates the non-nonparametric case. The tangent space and all vectors shown have been translated over P, except P itself. The space $L^2_{\mathcal{X},\mu}$ is depicted as $\mathbb{R}^3$, that is, for the case $\mathcal{X} = \{1, 2, 3\}$. Of the unit sphere only its intersection with the horizontal plane in $\mathbb{R}^3$ is shown; the set $\mathcal{P}$ happens to lie in this intersection and has a one-dimensional tangent space $\dot{\mathcal{P}}(P)$ at P. The difference of $2tP$ and $2t'P$ is seen to be orthogonal to the horizontal plane, which here is equal to the space spanned by P and $\dot{\mathcal{P}}(P)$.

3. Independent draws

Here we examine what happens to the lower bound in Theorem 1 if the estimating is based on n independent draws from P, that is, one draw from P^n.

In the formulation of the theorem below we use the facts that a tangential derivative of an $\mathbb{R}$-valued function is a bounded linear $\mathbb{R}$-valued map and that by the Riesz–Fréchet representation theorem for such maps on real Hilbert spaces there are always vectors that represent them. From what we saw in the previous section it is clear that the resulting new

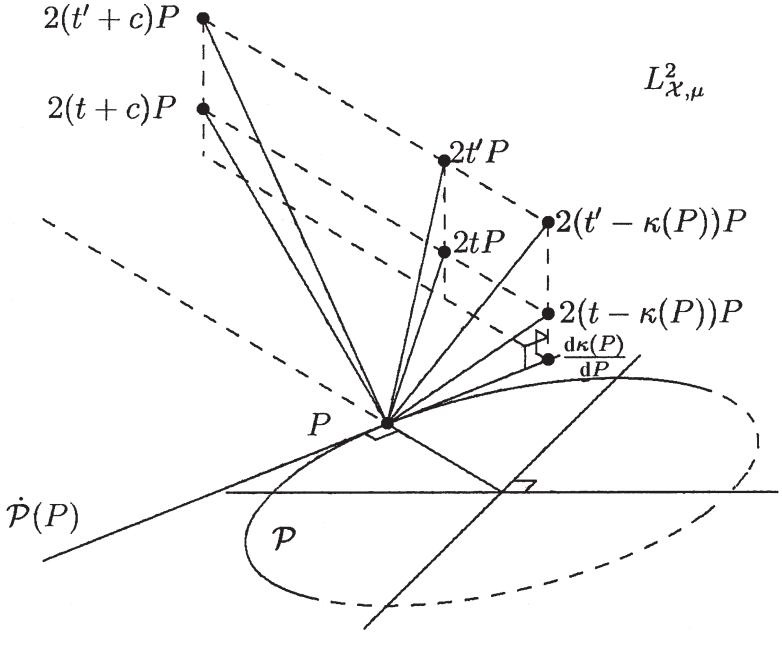

Figure 1. A view of Theorem 1 for a non-nonparametric model $\mathcal{P}$.

definitions extend the old ones. We require that μ be σ-finite, so that μ^n is defined and dominates P^n if μ dominates P.

Theorem 3. *If $(\mathcal{X}, \mathcal{A})$ is a measurable space, μ a σ-finite measure on $(\mathcal{X}, \mathcal{A})$, $\mathcal{P}$ a set of probability measures on $(\mathcal{X}, \mathcal{A})$ that have a density with respect to μ, P is a member of $\mathcal{P}$, n of $\mathbb{N}\backslash\{0\}$, we identify each $Q \in \mathcal{P}$ with the corresponding* (density of $Q)^{1/2} \in L^2_{\mathcal{X},\mu}$ *and each $Q^n \in \mathcal{P}^{(n)}$ with the corresponding* (density of $Q^n)^{1/2} \in L^2_{\mathcal{X}^n,\mu^n}$, *and $\kappa : \mathcal{P} \to \mathbb{R}$ and $\kappa_n : \mathcal{P}^{(n)} \to \mathbb{R}$ are related by $\kappa_n : Q^n \mapsto \kappa(Q)$ for all $Q \in \mathcal{P}$, then the tangential differentiability of κ at P and the tangential differentiability of κ_n at P^n are equivalent, and if both hold, we have*

$$\left\| \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P^n} \right\|^2_{L^2_{\mathcal{X}^n,\mu^n}} = \frac{1}{n} \left\| \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P} \right\|^2_{L^2_{\mathcal{X},\mu}},$$

where $\mathrm{d}\kappa(P)/\mathrm{d}P^n$ is defined as equal to the unique member of the tangent space $\dot{\mathcal{P}}^{(n)}(P^n)$ of $\mathcal{P}^{(n)}$ at P^n representing the tangential derivative of κ_n at P^n, and $\mathrm{d}\kappa(P)/\mathrm{d}P :=$ the unique member of the tangent space $\dot{\mathcal{P}}(P)$ of $\mathcal{P}$ at P representing the tangential derivative of κ at P.

Proof. Denote for members $a_1, \ldots, a_n \in L^2_{\mathcal{X},\mu}$ the function $(x_1, \ldots, x_n) \in \mathcal{X}^n \mapsto a_1(x_1) \ldots a_n(x_n) \in \mathbb{R}$ by $a_1 \otimes \cdots \otimes a_n$ (which, with our identification, leads to $Q^n \equiv Q^{\otimes n}$). Then Fubini is seen to give $a_1 \otimes \cdots \otimes a_n \in L^2_{\mathcal{X}^n,\mu^n}$ and even

$$\langle a_1 \otimes \cdots \otimes a_n, b_1 \otimes \cdots \otimes b_n \rangle_{L^2_{\mathcal{X}^n,\mu^n}} = \prod_{i=1}^{n} \langle a_i, b_i \rangle_{L^2_{\mathcal{X},\mu}} \quad \forall a_i, b_i \in L^2_{\mathcal{X},\mu},\ i = 1, \ldots, n. \tag{4}$$

With $f : a \in L^2_{\mathcal{X},\mu} \mapsto a^{\otimes n} \in L^2_{\mathcal{X}^n,\mu^n}$ we have

$$f(a+h) - f(a) = h \otimes a^{\otimes n-1} + a \otimes h \otimes a^{\otimes n-2} + \ldots + a^{\otimes n-1} \otimes h$$

$$+ \text{terms with more than one factor } h;$$

(4) implies the latter terms are $o(\|h\|)$, and T_a: $h \in L^2_{\mathcal{X},\mu} \mapsto h \otimes a^{\otimes n-1} + a \otimes h \otimes a^{n-2} + \ldots + a^{\otimes n-1} \otimes h$ is bounded, while T_a is also linear. According to the Appendix, f is everywhere differentiable with derivative $f'(a)$ at a equal to T_a, and by Theorem A.1 the restriction $f|_{\mathcal{P}}$ is tangentially differentiable at P with tangential derivative at P equal to $T_P|_{\dot{\mathcal{P}}(P)}$. As $f(P) = P^n$, it follows from the chain rule (Theorem A.3) that if κ_n is tangentially differentiable at P^n, then so is $\kappa = \kappa_n \circ f|_{\mathcal{P}}$ at P.

For the second half of the first statement of the theorem we construct a map $g : \mathcal{V} \subset L^2_{\mathcal{X}^n,\mu^n} \to L^2_{\mathcal{X},\mu}$. (We owe this construction to Arnoud van Rooij.) Let $b \in L^2_{\mathcal{X}^n,\mu^n}$, $b \geqslant 0$. Then for all $x_1 \in \chi$ the function $(x_2, \ldots, x_n) \mapsto b(x_1, x_2, \ldots, x_n) \in \mathbb{R}$ is measurable and

$$(z(b))(x_1) := \int_{\mathcal{X}^{n-1}} b(x_1, \ldots, x_n) P(x_2) \ldots P(x_n)\, \mathrm{d}\mu(x_2) \ldots \mathrm{d}\mu(x_n)$$

defines a number in $[0, \infty]$, while $x_1 \in \mathcal{X} \mapsto (z(b))(x_1)$ is measurable. By Cauchy–Schwarz and (4) we have

$$\begin{aligned}\int_{\mathcal{X}} |z(b)h| \,\mathrm{d}\mu &\leqslant \int_{\mathcal{X}^n} |b|.|h| \otimes P^{\otimes n-1} \,\mathrm{d}\mu^n \\ &\leqslant \|b\|_{L^2_{\mathcal{X}^n,\mu^n}} \|h\|_{L^2_{\mathcal{X},\mu}} \|P\|^{n-1}_{L^2_{\mathcal{X},\mu}} \\ &= \|b\|_{L^2_{\mathcal{X}^n,\mu^n}} \|h\|_{L^2_{\mathcal{X},\mu}}, \qquad \forall h \in L^2_{\mathcal{X},\mu}, \end{aligned} \tag{5}$$

and in particular we see that, if $\mathcal{X} = \cup_{i=1}^{\infty} A_i$ for $A_i \in \mathcal{A}$ with $\mu(A_i) < \infty$, every $\int z(b) 1_{A_i} \,\mathrm{d}\mu$ is finite. It follows that $z(b)1_{A_i}$ is μ-almost everywhere finite, and so is $z(b)$. If, therefore, in the integrand of the integral that defined $(z(b))(x_1)$ we take an arbitrary $b \in L^2_{\mathcal{X}^n,\mu^n}$, the integral exists and is finite for μ-almost every $x_1 \in L^2_{\mathcal{X},\mu}$. Call the resulting μ-almost everywhere defined function $z(b)$ again. We prove $z(b) \in L^2_{\mathcal{X},\mu}$.

Take the A_i from above and let $h_i := 1_{|z(b)| \leqslant i} 1_{\cup_{j=1}^{i} A_j} |z(b)|$. Then $h_i \in L^2_{\mathcal{X},\mu}$; because (5) still holds for the new b, we have $\|b\| \, \|h_i\| \geqslant \int_{\mathcal{X}} |z(b)| h_i \,\mathrm{d}\mu = \int_{\mathcal{X}} h_i^2 \,\mathrm{d}\mu$. We obtain $\int h_i^2 \,\mathrm{d}\mu \leqslant \|b\|^2$ for all i; the μ-almost everywhere convergence $h_i \uparrow |z(b)|$ entails $\int_{\mathcal{X}} (z(b))^2 \,\mathrm{d}\mu \leqslant \|b\|^2$.

From (5) with $h := z(b)$ it now follows that the linear map $b \mapsto z(b)$ is bounded and therefore everywhere differentiable; since, away from $\{0\}$, taking the norm is differentiable as well, we have that

$$g : b \in \mathcal{V} \subset L^2_{\mathcal{X}^n,\mu^n} \mapsto \frac{z(b)}{\|z(b)\|} \in L^2_{\mathcal{X},\mu}$$

is differentiable on the open set $\mathcal{V} := \{b : z(b) \neq 0\} \subset L^2_{\mathcal{X}^n,\mu^n}$. Observe $z(Q^n) = (\int QP \,\mathrm{d}\mu)^{n-1} Q$ for all $Q \in \mathcal{P}$, so $z(P^n) = P$ and $P^n \in \mathcal{V}$. Again by Theorem A.1, the restriction $g|_{\mathcal{P}^{(n)} \cap \mathcal{V}}$ is differentiable at P^n along the tangent space of $\mathcal{P}^{(n)} \cap \mathcal{V}$ at P^n, which is $\dot{\mathcal{P}}^{(n)}(P^n)$ because $\mathcal{V}$ is open.

If $Q \in \mathcal{P}$ is such that $z(Q^n) \neq 0$, then $(\int QP \,\mathrm{d}\mu)^{n-1} > 0$ because $QP \geqslant 0$, so that $\|z(Q^n)\| = (\int QP \,\mathrm{d}\mu)^{n-1}$ and $g(Q^n) = Q$. We conclude that $f|_{\mathcal{P}} \circ g|_{\mathcal{P}^{(n)} \cap \mathcal{V}} = \mathrm{id}_{\mathcal{P}^{(n)} \cap \mathcal{V}}$ and $\kappa_n|_{\mathcal{P}^{(n)} \cap \mathcal{V}} = \kappa \circ g|_{\mathcal{P}^{(n)} \cap \mathcal{V}}$. As $g(P^n) = P$, it follows from the latter conclusion and Theorem A.3 that if κ is tangentially differentiable at P, then so is $\kappa_n|_{\mathcal{P}^{(n)} \cap \mathcal{V}}$ at P^n, and from the openness of $\mathcal{V}$ again that if $\kappa_n|_{\mathcal{P}^{(n)} \cap \mathcal{V}}$ is tangentially differentiable at P^n, then so is κ_n at P^n. This proves the first statement of the theorem.

From $f|_{\mathcal{P}} \circ g|_{\mathcal{P}^{(n)} \cap \mathcal{V}} = \mathrm{id}_{\mathcal{P}^{(n)} \cap \mathcal{V}}$ and the chain rule we infer that

$$T_P\big(\dot{\mathcal{P}}(P)\big) = \dot{\mathcal{P}}^{(n)}(P^n). \tag{6}$$

We can now determine the relation between the norms of $\mathrm{d}\kappa(P)/\mathrm{d}P$ and $\mathrm{d}\kappa(P)/\mathrm{d}P^n$. Suppose the corresponding derivatives exist. For them, application of the chain rule to $\kappa = \kappa_n \circ f|_{\mathcal{P}}$ yields

$$\left\langle T_P(h), \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P^n} \right\rangle_{L^2_{\mathcal{X}^n,\mu^n}} = \left\langle h, \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P} \right\rangle_{L^2_{\mathcal{X},\mu}} \qquad \forall h \in \dot{\mathcal{P}}(P). \tag{7}$$

We have $\langle P, P \rangle_{L^2_{\mathcal{X},\mu}} = 1$ and, from Theorem 1, $\langle h, P \rangle_{L^2_{\mathcal{X},\mu}} = 0 = \langle P, h \rangle_{L^2_{\mathcal{X},\mu}}$ for all $h \in \dot{\mathcal{P}}(P)$, so that from (4) we obtain

$$\langle T_P(h),\, T_P(k)\rangle_{L^2_{\mathcal{X}^n,\mu^n}} = n\langle h,\, k\rangle_{L^2_{\mathcal{X},\mu}} \qquad \forall h,\, k \in \dot{\mathcal{P}}(P), \tag{8}$$

from which, with $k = \mathrm{d}\kappa(P)/\mathrm{d}P$, we see that

$$\left\langle T_P(h),\, T_P\left(\frac{1}{n}\frac{\mathrm{d}\kappa(P)}{\mathrm{d}P}\right)\right\rangle_{L^2_{\mathcal{X}^n,\mu^n}} = \left\langle h,\, \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P}\right\rangle_{L^2_{\mathcal{X},\mu}} \qquad \forall h \in \dot{\mathcal{P}}(P).$$

This gives, with (7) and (6),

$$T_P\left(\frac{1}{n}\frac{\mathrm{d}\kappa(P)}{\mathrm{d}P}\right) = \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P^n},$$

so that

$$\frac{1}{n^2}\left\| T_P\left(\frac{\mathrm{d}\kappa(P)}{\mathrm{d}P}\right)\right\|^2 = \left\|\frac{\mathrm{d}\kappa(P)}{\mathrm{d}P^n}\right\|^2$$

and the last statement of the theorem follows from (8). □

Continuation 4. *If, under the circumstances of Theorem 3, $t : \mathcal{X} \to \mathbb{R}$ is $\mathcal{P}$-regular at P, then $\sum_{i=1}^n \alpha_i t_i : \mathcal{X}^n \to \mathbb{R}$, where $t_i : (x_1, \ldots, x_n) \mapsto t(x_i)$, is $\mathcal{P}^{(n)}$-regular at P^n for all $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$.*

Proof. Suppose $\sum_{i=1}^n \alpha_i = 1$. Let $\kappa(Q) := \langle tQ, Q\rangle$, $Q \in \mathcal{P}$. What needs proof is the requirement that $2(\sum_{i=1}^n \alpha_i t_i)P^n$ should represent the tangential derivative of κ_n at P^n. For all i, we have $t_i P^n = P^{i-1} \otimes tP \otimes P^{n-i}$ and therefore

$$\begin{aligned}
\langle T_P(h),\, 2t_i P^n\rangle_{L^2_{\mathcal{X}^n,\mu^n}} &= \langle h \otimes P^{n-1} + \ldots + P^{n-1} \otimes h,\, 2P^{i-1} \otimes tP \otimes P^{n-i}\rangle_{L^2_{\mathcal{X}^n,\mu^n}} \\
&\overset{\text{proof Th. 3}}{=} \langle h,\, 2tP\rangle_{L^2_{\mathcal{X},\mu}} \\
&\overset{t \text{ regular}}{=} \left\langle h,\, \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P}\right\rangle_{L^2_{\mathcal{X},\mu}} \\
&\overset{(7)}{=} \left\langle T_P(h),\, \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P^n}\right\rangle_{L^2_{\mathcal{X}^n,\mu^n}} \qquad \forall h \in \dot{\mathcal{P}}(P),
\end{aligned}$$

so that $\langle \cdot,\, 2t_i P^n\rangle_{L^2_{\mathcal{X}^n,\mu^n}}$ and $\langle \cdot,\, \mathrm{d}\kappa(P)/\mathrm{d}P^n\rangle_{L^2_{\mathcal{X}^n,\mu^n}}$ are seen to agree on the image $T_P\big(\dot{\mathcal{P}}(P)\big) \overset{(6)}{=} \dot{\mathcal{P}}^{(n)}(P^n)$. But then $\langle \cdot,\, \mathrm{d}\kappa(P)/\mathrm{d}P^n\rangle_{L^2_{\mathcal{X}^n,\mu^n}}$ agrees with $\langle \cdot,\, 2(\sum_{i=1}^n \alpha_i t_i)P^n\rangle_{L^2_{\mathcal{X}^n,\mu^n}}$ on $\dot{\mathcal{P}}^{(n)}(P^n)$. □

Corollary 5. *If $\mathcal{P}$ is nonparametric, μ σ-finite, and t regular and unbiased, then the mean $n^{-1}\sum_{i=1}^n t_i$ is a regular unbiased estimator $\mathcal{X}^n \to \mathbb{R}$ that achieves the bound* (3) *everywhere; thus, it has the smallest variance everywhere among the regular unbiased estimators with the same domain.*

In the next section it will turn out that such a mean is also optimal, for a quantity related to the variance, among *all* unbiased estimators.

4. Sufficient conditions for regularity; a new bound

We identify each $Q \in \mathcal{P}$ with the corresponding (density of $Q)^{1/2} \in L^2_{\mathcal{X},\mu}$. Let $tQ \in L^2_{\mathcal{X},\mu}$ for all Q in a neighbourhood of P. According to the Appendix, the map $Q \mapsto \langle tQ, Q\rangle$ is tangentially differentiable at P with derivative $h \in \dot{\mathcal{P}}(P) \mapsto \langle h, 2tP\rangle$ at P if and only if, for every $(\epsilon_n)_{n=1}^{\infty}$ in $\mathbb{R}\backslash\{0\}$ and $(h_n)_{n=1}^{\infty}$ in $L^2_{\mathcal{X},\mu}$, with $\epsilon_n \to 0$, $h_n \to h$, and $P + \epsilon_n h_n \in \mathcal{P}$, one would have, as $n \to 0$,

$$\frac{\langle t(P+\epsilon_n h_n),\, P+\epsilon_n h_n\rangle - \langle tP,\, P\rangle}{\epsilon_n} \to \langle h,\, 2tP\rangle.$$

This is the same as

$$\langle h_n,\, 2tP\rangle + \langle h_n,\, \epsilon_n t h_n\rangle \to \langle h,\, 2tP\rangle$$

and, because of $\langle h_n, 2tP\rangle \to \langle h, 2tP\rangle$, equivalent to

$$\langle h_n,\, \epsilon_n t h_n\rangle \to 0.$$

This convergence certainly holds if $t(x) = c$ for all $x \in \mathcal{X}$, as $(\|h_n\|^2)_{n=1}^{\infty}$ is bounded. Thus, constant real functions on $\mathcal{X}$ are indeed regular.

The convergence $\langle h_n, \epsilon_n t h_n\rangle \to 0$ is also implied by boundedness of $(tP_n)_{n=1}^{\infty}$, where $P_n := P + \epsilon_n h_n$. In order to see this, let $B > 0$ be a bound for $(\|t(P_n - P)\|_{n=1}^{\infty})$ and observe that, for every $c > 0$ and $n = 1, 2, \ldots$, we have

$$\begin{aligned}\langle h_n,\, \epsilon_n t h_n\rangle &= \langle h_n,\, t(P_n - P)\rangle \\ &= \langle h_n,\, t1_{|t|\leqslant c}(P_n - P)\rangle + \langle h_n 1_{|t|>c},\, t(P_n - P)\rangle,\end{aligned}$$

so that, by Cauchy–Schwarz,

$$|\langle h_n,\, \epsilon_n t h_n\rangle| \leqslant \|h_n\| \cdot c\|P_n - P\| + \|h_n 1_{|t|>c}\| \cdot \|t(P_n - P)\|$$

and therefore, for all $c > 0$,

$$\limsup_{n\to\infty} |\langle h_n,\, \epsilon_n t h_n\rangle| \leqslant \|h\| \cdot c \cdot 0 + \|h1_{|t|>c}\| \cdot B,$$

while $\lim_{c\to\infty}\|h1_{|t|>c}\| = 0$ by Lebesgue. That tP_n is bounded if P_n approaches P in $\mathcal{P} \subset L^2_{\mathcal{X},\mu}$ for all $(P_n)_{n=1}^{\infty}$ in $\mathcal{P}$ for which there are $\epsilon_n \neq 0$, $\epsilon_n \to 0$, and an $h \in L^2_{\mathcal{X},\mu}$ such that $(P_n - P)/\epsilon_n \to h$, is the same as local boundedness at P (i.e., boundedness on an $L^2_{\mathcal{X},\mu}$ neighbourhood of P in $\mathcal{P}$) of $Q \in \mathcal{P} \mapsto \mathrm{E}_Q T^2$ (for T, cf. (1)), because if $P_n \to P$, then $(P_n - P)/\sqrt{\|P_n - P\|} \to 0$. We have proved:

Lemma 6. *A measurable $\mathcal{X} \xrightarrow{t} \mathbb{R}$ is regular at Q for every Q in a neighbourhood of P if the second moments $\mathrm{E}_Q T^2$, $Q \in \mathcal{P}$, are locally bounded at P.*

Lemma 6 enables us to provide an information bound for all unbiased estimators, and to see what happens when an unbiased estimator beats the bound (3).

Theorem 7. *If* $(\mathcal{X}, \mathcal{A})$ *is a measurable space,* μ *a measure on* $(\mathcal{X}, \mathcal{A})$*,* $\mathcal{P}$ *a set of probability measures on* $(\mathcal{X}, \mathcal{A})$ *that have a density with respect to* μ*,* P *is a member of* $\mathcal{P}$*, we identify each* $Q \in \mathcal{P}$ *with the corresponding* (density of $Q)^{1/2} \in L^2_{\mathcal{X},\mu}$*, and* $\kappa : \mathcal{P} \to \mathbb{R}$ *is a parameter of interest which is locally bounded at* P *(e.g., by being tangentially differentiable at* P*), then, with*

$$\operatorname{maxVar}_P T := \lim_{n\to\infty} \sup_{\substack{Q\in\mathcal{P} \\ \text{distance}(Q,P)<1/n}} \operatorname{var}_Q T$$

and

$$B_P := \lim_{n\to\infty} \sup_{\substack{Q\in\mathcal{P} \\ \text{distance}(Q,\,P)<1/n}} \frac{1}{4}\left\| \frac{\mathrm{d}\kappa(Q)}{\mathrm{d}Q} \right\|^2_{L^2_{\mathcal{X},\mu}} \quad \text{or } B_P := \infty,$$

where the former definition of B_P *applies if and only if* $Q \in \mathcal{P} \mapsto \kappa(Q)$ *is tangentially differentiable on a neighbourhood of* P*, in which case* $\mathrm{d}\kappa(Q)/\mathrm{d}Q$ *denotes the unique vector in the tangent space of* $\mathcal{P}$ *at* Q *that represents the derivative at* Q*, we have, with* $\mathcal{T} :=$ *the set of all unbiased estimators* $\mathcal{X} \to \mathbb{R}$ *for* κ*,*

$$\operatorname{maxVar}_P T \geqslant B_P \qquad \forall t \in \mathcal{T}; \tag{9}$$

if, moreover, κ *is indeed tangentially differentiable at* P*, then*

$$\operatorname{var}_P T < \frac{1}{4}\left\| \frac{\mathrm{d}\kappa(P)}{\mathrm{d}P} \right\|^2_{L^2_{\mathcal{X},\mu}} \;\Rightarrow\; \operatorname{maxVar}_P T = \infty \qquad \forall t \in \mathcal{T}.$$

Proof. If $Q \mapsto \mathrm{E}_Q T^2$ is locally bounded at P, then by Lemma 6 there is a neighbourhood U of P such that t is regular at Q for all $Q \in U$, so that by Theorem 1 we have $\operatorname{var}_Q T \geqslant \frac{1}{4}\|\mathrm{d}\kappa(Q)/\mathrm{d}Q\|^2_{L^2_{\mathcal{X},\mu}}$ for all $Q \in U$; if $Q \mapsto \mathrm{E}_Q T^2$ is not locally bounded at P, then neither is $Q \mapsto \operatorname{var}_Q T$ because $\mathrm{E}_Q T = \kappa(Q)$ and κ is locally bounded at P, so that $\operatorname{maxVar}_P T = \infty$. □

In defining $\operatorname{maxVar}_P T$, the $L^2_{\mathcal{X},\mu}$ distance could be replaced by the total variation distance, as these two metrics are equivalent. So we do not need μ for $\operatorname{maxVar}_P T$.

If κ is not locally bounded, then it is not total variation continuous and there are no uniformly consistent estimators for κ based on independent draws from P; cf. Bickel *et al.* (1993, p. 20).

Proposition 8. *If under these circumstances* κ *is tangentially differentiable everywhere, then every unbiased estimator* $\mathcal{X} \to \mathbb{R}$ *that achieves the bound* (3) *everywhere (cf. Corollary 5)*

achieves the bound (9) *everywhere; thus, everywhere it has the smallest maxVariance among all unbiased estimators for* κ *with the same domain.*

Corollary 9. *If* $(\mathcal{X}, \mathcal{A})$ *is a measurable space,* $\mathcal{P}$ *is a set of probability measures on* $(\mathcal{X}, \mathcal{A})$, $\kappa : \mathcal{P} \to \mathbb{R}$ *is a parameter of interest, and* $t : \mathcal{X} \to \mathbb{R}$ *is an unbiased estimator for* κ *such that for all total variation convergent* $P_k \to P_0$ *in* $\mathcal{P}$ *there are a measure* μ *on* $(\mathcal{X}, \mathcal{A})$ *and a subset* $\mathcal{P}_\mu \subset \{P \in \mathcal{P} : P \text{ has a density w.r.t. } \mu\}$ *containing* $(P_k)_{k=0}^\infty$ *for which* $\mathrm{d}\kappa|_{\mathcal{P}_\mu}(P)/\mathrm{d}P$ *exists everywhere and* t, *as an unbiased estimator for* $\kappa|_{\mathcal{P}_\mu}$, *achieves the bound* (3) *everywhere, then* t *has the smallest maxVariance everywhere among all unbiased estimators for* κ *with the same domain.*

Proof. Let $P_0 \in \mathcal{P}$; let $s : \mathcal{X} \to \mathbb{R}$ be an unbiased estimator for κ. Choose $P_1, P_2, \ldots$ with $P_k \to P_0$ and $\mathrm{var}_{P_k} T \to \mathrm{maxVar}_{P_0} T$ and choose μ, $\mathcal{P}_\mu$ as above. Then Proposition 8 gives $\mathrm{maxVar}_{P_0}^{\mathcal{P}_\mu} S \geqslant \mathrm{maxVar}_{P_0}^{\mathcal{P}_\mu} T$, where '$\mathcal{P}_\mu$' has been added in order to indicate that for the suprema at hand only members of $\mathcal{P}_\mu \subset \mathcal{P}$ are considered, and $\mathrm{maxVar}_{P_0} S \geqslant \mathrm{maxVar}_{P_0} T$ follows from $\mathrm{maxVar}_{P_0} S \equiv \mathrm{maxVar}_{P_0}^{\mathcal{P}} S \geqslant \mathrm{maxVar}_{P_0}^{\mathcal{P}_\mu} S$ and $\mathrm{maxVar}_{P_0}^{\mathcal{P}_\mu} T = \mathrm{maxVar}_{P_0}^{\mathcal{P}} T \equiv \mathrm{maxVar}_{P_0} T$. □

In the next section there will be an application of Corollary 9.

For regularity there is a separate condition saying that one of the vectors representing the derivative should be $2tP$. One might wonder if this condition followed from the mere tangential differentiability of $P \mapsto \langle tP, P\rangle$, for, if it did, then the regularity conditions could be restricted to the first condition: having a finite second moment. This is because the differentiability condition is just a condition on the statistical problem at hand: the parameter of interest κ. However, there is no such implication, as we now show.

Let $\mathcal{X} := \mathbb{R}$ and μ be equal to Borel–Lebesgue measure on the Borel sets of $[0, 1]$, $\mu : \{2\} \mapsto 1$, and $\mu : \mathbb{R}\backslash([0, 1] \cup \{2\}) : \mapsto 0$. Let P correspond to the density $1_{\{2\}}$ and P_n to the density $(1/n^2)1_{[0,1-1/n]} + (1 - (1/n^2)(1 - 1/n))1_{\{2\}}$ and identify P, P_n with $1_{\{2\}} \in L^2_{\mathbb{R},\mu}$ and $(1/n)1_{[0,1-1/n]} + (1 - (1/n^2)(1 - 1/n))^{1/2}1_{\{2\}} \in L^2_{\mathbb{R},\mu}$ respectively; let $\mathcal{P} := \{P, P_2, P_3, \ldots\}$. Then $\|P_n - P\|^2_{L^2_{\mathbb{R},\mu}} = 2 - 2(1 - (1/n^2)(1 - 1/n))^{1/2}$, for all $x \in \mathbb{R}$ we have $(P_n(x) - P(x))/\|P_n - P\| \to 1_{[0,1)}(x)$, and by Lebesque we also have $(P_n - P)/\|P_n - P\| \to 1_{[0,1)}$ in $L^2_{\mathbb{R},\mu}$. So $\mathbb{R}1_{[0,1)} \subset \dot{\mathcal{P}}(P)$; we prove equality.

Suppose there are $Q_n \in \mathcal{P}$, $\epsilon_n \neq 0$, $\epsilon_n \to 0$ with $(Q_n - P)/\epsilon_n \to h \in \dot{\mathcal{P}}(P)$, $h \neq 0$. Then $Q_n \to P$ and $Q_n \neq P$ eventually. So there are subsequences of $(Q_n)_{n=1}^\infty$ and $(P_n)_{n=1}^\infty$, again denoted by $(Q_n)_{n=1}^\infty$ and $(P_n)_{n=1}^\infty$, with $Q_n = P_n$ for all n, and we have $(P_n - P)/\epsilon_n \to h$, $(P_n - P)/\|P_n - P\| \to 1_{[0,1)}$. We see that $(\|P_n - P\|)/|\epsilon_n| \to \|h\|$ and, for a subsequence, $(\|P_n - P\|)/\epsilon_n \to \|h\|$, say, so that $h = \|h\|1_{[0,1)} \in \text{closed } \mathbb{R}1_{[0,1)}$.

Take an arbitrary $\lambda > 0$; there are many $t : \mathbb{R} \to (0, \infty)$ for which $\langle 1_{[0,1)}, tP_n\rangle = \int_0^1 tP_n \,\mathrm{d}\mu \to \lambda$, $tP_n \in L^2_{\mathbb{R},\mu}$, and $t(x) = 0$ for all $x \notin [0, 1]$; take such a t. Then $Q \in \mathcal{P} \mapsto \langle tQ, Q\rangle$ is tangentially differentiable at P, but the derivative at P is not equal to $h \in \dot{\mathcal{P}}(P) \mapsto \langle h, 2tP\rangle = 0$, because it is equal to $\mu 1_{[0,1)} \in \dot{\mathcal{P}}(P) \mapsto \mu\lambda \in \mathbb{R}$. This follows from what happens to $\langle h_n, \epsilon_n t h_n\rangle$ as $n \to \infty$. Namely, if we choose $h_n := (P_n - P)/\|P_n - P\|$ and $h := 1_{[0,1)}$, then

$$\begin{aligned}\langle h_n, \epsilon_n t h_n\rangle &= \left\langle \frac{P_n - P}{\|P_n - P\|},\, t(P_n - P)\right\rangle \\ &= \left(\frac{P_n - P}{\|P_n - P\|}\right)\left(\frac{1}{4}\right)\langle 1_{[0,1)},\, tP_n\rangle \\ &\to 1 \cdot \lambda \qquad \text{as } n \to \infty;\end{aligned}$$

arbitrary $h_n \to h \in \dot{\mathcal{P}}(P)$ reduce to this special case, as we saw above: every subsequence of (h_n) has a subsequence that behaves as a subsequence of $\|h\|(P_n - P)/(\|P_n - P\|)$ or its negative.

5. An example

Let μ be a measure on ($\mathbb{R}$, Borel sets of $\mathbb{R}$) and $\mathcal{P}$ be the set of all (square roots of densities of) probability measures on $\mathbb{R}$ that have a density with respect to μ. Let $r \in \mathbb{R}$ and $\kappa : P \in \mathcal{P} \mapsto P((-\infty, r])$. Then $t := 1_{(-\infty, r]}$ is an unbiased estimator for κ, which is regular by $\|tP\| \leqslant 1$ for all $P \in \mathcal{P}$ and Lemma 6. By Proposition 2 it is essentially unique and its variance $\kappa(P)(1 - \kappa(P))$ is equal to the bound (3), because $\mathcal{P}$ is nonparametric (if $gP \perp P$ and $\chi(x) = 2/(1 + \mathrm{e}^{-2x})$, then

$$\eta \mapsto \left[x \mapsto \left(\frac{\chi(\eta g(x))(P(x))^2}{\displaystyle\int_{\mathbb{R}} \chi(\eta g(y))(P(y))^2 \,\mathrm{d}\mu(y)} \right)^{1/2} \right]$$

is a map into $\mathcal{P}$, continuous on a neighbourhood of 0 and differentiable at 0 with value at 1 of the derivative at 0 equal to $\frac{1}{2}gP$, so gP is even a tangent vector in the strict sense of representing the derivative of a curve).

For μ σ-finite and $n \geqslant 2$, however, $\mathcal{P}^{(n)}$ need not be nonparametric. In order to see this, suppose there are $P \in \mathcal{P}$ and $r \in \mathbb{R}$ such that $\kappa(P) \in (0, 1)$. Then $\mathrm{var}_P\, T \neq 0$ and $\mathrm{var}_{P^n} T_1 \neq \mathrm{var}_{P^n} n^{-1}\sum_{i=1}^n T_i$, while both $t_1 : (x_1, \ldots, x_n) \mapsto t(x_1)$ and $n^{-1}\sum_{i=1}^n t_i$ are unbiased estimators that are regular according to Continuation 4; apply Proposition 2.

If we change $\mathcal{P}$ into the set of *all* probability measures on ($\mathbb{R}$, Borel sets of $\mathbb{R}$), then it follows that among all unbiased estimators $\mathbb{R}^n \to \mathbb{R}$ for κ: $P \in \mathcal{P} \mapsto P((-\infty, r])$ the mean $n^{-1}\sum_{i=1}^n t_i$ has the smallest maxVariance everywhere, that is, the empirical distribution function at r is optimal for that quantity. Namely, for any $P_k^n \to P_0^n$, we can take $\mu := (\sum 2^{-k} P_k)^n$ and $\mathcal{P}_\mu := (\{P \in \mathcal{P} : P \text{ has a density w.r.t. } \sum 2^{-k} P_k\})^{(n)}$ in Corollary 9; apply Corollary 5.

6. Fisher information and Cramér–Rao

If the parameter of interest $\kappa : \mathcal{P} \to \mathbb{R}$ is the inverse of a given *parametrization*

$\vartheta \in \Theta \subset \mathbb{R} \mapsto P_\vartheta \in \mathcal{P}$ *of* (=bijection on) $\mathcal{P}$, i.e., if $\kappa : P_\vartheta \mapsto \vartheta$, we might try, in Theorem 1, to differentiate with respect to ϑ instead of P.

In general, if H is a Hilbert space, $\Theta \subset \mathbb{R}$ is open, $\vartheta_0 \in \Theta$, and $\vartheta \mapsto v_\vartheta \in H$ is everywhere differentiable and continuously differentiable at ϑ_0 (for the topology on the derivatives, take the topology of the vectors $\mathrm{d}v_\vartheta/\mathrm{d}\vartheta \in H$, $\vartheta \in \Theta$, that are the images of $1 \in \mathbb{R}$ under these derivatives) with $\mathrm{d}v_{\vartheta_0}/\mathrm{d}\vartheta_0 \neq 0$, then

(i) there is an open set $U \subset \Theta$ with $\vartheta_0 \in U$ such that $\vartheta \in U \mapsto v_\vartheta$ is a homeomorphism between U and $V := \{v_\vartheta \in H : \vartheta \in U\}$
(ii) whose inverse, $(v_\vartheta \mapsto \vartheta)|_V$, has a tangential derivative everywhere, in particular at v_{ϑ_0}, while
(iii) for the unique vector $\mathrm{d}\vartheta_0/\mathrm{d}v_{\vartheta_0}$ in the tangent space of V at v_{ϑ_0} representing this derivative we have

$$\frac{\mathrm{d}\vartheta_0}{\mathrm{d}v_{\vartheta_0}} = \left\| \frac{\mathrm{d}v_{\vartheta_0}}{\mathrm{d}\vartheta_0} \right\|^{-2} \cdot \frac{\mathrm{d}v_{\vartheta_0}}{\mathrm{d}\vartheta_0},$$

according to the inverse mapping theorem for Banach spaces and the chain rule (see Lang 1995; Lenstra 1998); in this section we take $H := L^2_{\mathcal{X},\mu}$ and $v_\vartheta := P_\vartheta$, that is, (density of $P_\vartheta)^{1/2}$. (For conditions on the mappings $\vartheta \mapsto$ (density of $P_\vartheta)(x)$, $x \in \mathcal{X}$, that ensure the required differentiability of $\vartheta \mapsto$ (density of $P_\vartheta)^{1/2}$, see Bickel *et al.* 1993.)

The result is a version of the Cramér–Rao inequality that resembles what Theorem 7.3 in Ibragimov and Has'minskii (1981) says for the unbiased one-dimensional case, and is mentioned because of its proof:

Theorem 10 (Cramér–Rao). *If* $(\mathcal{X}, \mathcal{A})$ *is a measurable space,* μ *is a measure on* $(\mathcal{X}, \mathcal{A})$, $\Theta \subset \mathbb{R}$ *is open,* $(P_\vartheta)_{\vartheta \in \Theta}$ *is an indexed family of probability measures on* $(\mathcal{X}, \mathcal{A})$ *that have a density with respect to* μ, ϑ_0 *is a member of* Θ, *we identify each* P_ϑ *with the corresponding* (density of $P_\vartheta)^{1/2} \in L^2_{\mathcal{X},\mu}$, $\vartheta \mapsto P_\vartheta$ *is injective, everywhere differentiable, and continuously (see above) differentiable at* ϑ_0 *with non-zero derivative d,* $\mathcal{T}_{\mathrm{lb}}$ *is the set of the unbiased estimators for* $P_\vartheta \mapsto \vartheta$ *for which* $\vartheta \in \Theta \mapsto \mathrm{E}_{P_\vartheta} T^2$ *is* $\mathbb{R}$*-locally bounded at* ϑ_0, $\mathrm{d}P_{\vartheta_0}/\mathrm{d}\vartheta_0$ *denotes the image in* $L^2_{\mathcal{X},\mu}$ *of* $1 \in \mathbb{R}$ *under d, and* $I_{\vartheta \mapsto P_\vartheta}(\vartheta_0)$ *denotes the Fisher information at* ϑ_0 *for the parametrization* $\vartheta \mapsto P_\vartheta$, *that is, the number* $4\|\mathrm{d}P_{\vartheta_0}/\mathrm{d}\vartheta_0\|^2$, *then we have, with* $\mathrm{var}_{\vartheta_0} T := \mathrm{var}_{P_{\vartheta_0}} T$,

$$\mathrm{var}_{\vartheta_0} T \geqslant (I_{\vartheta \mapsto P_\vartheta}(\vartheta_0))^{-1} \qquad \forall t \in \mathcal{T}_{\mathrm{lb}}.$$

Proof. Choose an open $U \subset \Theta$ containing ϑ_0 such that U and its image $V := \{P_\vartheta \in L^2_{\mathcal{X},\mu} : \vartheta \in U\}$ are as U and V in (i)–(iii) above. For $t \in \mathcal{T}_{\mathrm{lb}}$ the homeomorphism between U and V guarantees that $Q \in V \mapsto \mathrm{E}_Q T^2$ is $L^2_{\mathcal{X},\mu}$-locally bounded at P_{ϑ_0}, and Lemma 6 that t, which is also an unbiased estimator for $\kappa|_V$, is V-regular at P_{ϑ_0}. Now (3) gives $\mathrm{var}_{P_{\vartheta_0}} T \geqslant \frac{1}{4}\|\mathrm{d}\kappa|_V(P_{\vartheta_0})/\mathrm{d}P_{\vartheta_0}\|^2$ and we obtain the desired result from (iii). □

7. The van Trees inequality

The van Trees inequality is 'a Bayesian version of the Cramér–Rao bound', which is not to say that its importance is confined to Bayesians; see Gill and Levit (1995). Our van Trees inequality reads:

Theorem 11 (van Trees). *If*

- $\mathcal{X}$ *is a Polish space,* $\mathcal{A}$ *is the* σ*-algebra of its Borel sets,* μ *is a* σ*-finite measure on* $(\mathcal{X}, \mathcal{A})$*,* $(P_\vartheta)_{\vartheta\in\mathbb{R}}$ *is an indexed family of probability measures on* $(\mathcal{X}, \mathcal{A})$ *that have a density with respect to* μ*, we identify each* P_ϑ *with the corresponding* $(\text{density of } P_\vartheta)^{1/2} \in L^2_{\mathcal{X},\mu}$*,* $\vartheta \in \mathbb{R} \mapsto P_\vartheta$ *is injective and everywhere continuously differentiable with non-zero derivative,* $\mathrm{d}P_\vartheta/\mathrm{d}\vartheta$ *denotes the image in* $L^2_{\mathcal{X},\mu}$ *of* $1 \in \mathbb{R}$ *under the derivative at* ϑ*, and* $I_{\vartheta'\mapsto P_{\vartheta'}}(\vartheta) := 4\|\mathrm{d}P_\vartheta/\mathrm{d}\vartheta\|^2_{L^2_{\mathcal{X},\mu}}$*,*
- W *is a probability measure on* $\Theta \equiv \mathbb{R}$ *that has a density* w *with respect to* $\nu :=$ *Borel-Lebesgue measure on* $\mathbb{R}$ *(so that, for every* $\lambda \in \mathbb{R}$*, the function* $w_\lambda := w(.-\lambda)$ *is such a density too, of law* W_λ*, say, and* $\lambda \in \mathbb{R} \mapsto W_\lambda$ *is a parametrization), we identify each* W_λ *with the corresponding* $\sqrt{w_\lambda} \in L^2_{\mathbb{R},\nu}$*,* $\lambda \in \mathbb{R} \mapsto W_\lambda$ *is differentiable at 0 with non-zero derivative (so that the derivative exists everywhere and is constant),* $\mathrm{d}W_\lambda/\mathrm{d}\lambda$ *denotes the image in* $L^2_{\mathbb{R},\nu}$ *of* $1 \in \mathbb{R}$ *under the derivative at* λ*, and* $I_{\lambda'\mapsto W_{\lambda'}}(\lambda)$ *denotes* $4\|\mathrm{d}W_\lambda/\mathrm{d}\lambda\|^2_{L^2_{\mathbb{R},\nu}}$*,*
- $t : \mathcal{X} \to \mathbb{R}$ *is measurable and such that*
  (t1) *all expectations* $\mathrm{E}_\vartheta T^2$*,* $\vartheta \in \mathbb{R}$*, are finite, and*
  (t2) *the function* $\lambda \in \mathbb{R} \mapsto \mathrm{E}_{w_\lambda} T^2$ *is locally bounded at* $\lambda = 0$
  (where

$$\mathrm{E}_\vartheta(f \circ X) := \int_{\mathcal{X}} f(x)\,\mathrm{d}P_\vartheta(x),$$

$$\mathrm{E}_{w_\lambda}(g \circ (X, \Theta)) := \int_{\mathbb{R}} \mathrm{E}_\vartheta\, g(X, \vartheta)\,\mathrm{d}W_\lambda(\vartheta),$$

  for all $\vartheta, \lambda \in \mathbb{R}$ *and measurable* $f : \mathcal{X} \to \mathbb{R}$*,* $g : \mathcal{X} \times \mathbb{R} \to \mathbb{R}$*),*

then we have

$$\mathrm{E}(T - \Theta)^2 \geqslant \left(\int_{\mathbb{R}} I_{\vartheta'\mapsto P_{\vartheta'}}(\vartheta) w(\vartheta)\,\mathrm{d}\vartheta + I_{\lambda\mapsto W_\lambda}(0)\right)^{-1}$$

for any random element (X, Θ) *of* $\mathcal{X} \times \mathbb{R}$ *such that* Θ *is distributed according to* W *and* $(P_\vartheta)_{\vartheta\in\Theta}$ *is a conditional distribution of* X *given* Θ *(so* $\mathrm{E} \equiv \mathrm{E}_{w_0}$*) under any extra condition that justifies* (d1) *and* (d2) *below for at least one choice of* $\prod_w L^2_{\mathcal{X},\mu}$*, for example, the condition provided by the next section.*

(As to the existence of the integrals and of random elements as indicated: the mapping $\vartheta \mapsto I_{\vartheta'\mapsto P_{\vartheta'}}(\vartheta)$ is measurable because it is continuous; the same is true for every

$\vartheta \in \mathbb{R} \mapsto P_\vartheta(A)$, $A \in \mathcal{A}$, because $\vartheta \mapsto$ (density of $P_\vartheta)^{1/2}$ and (density of $P)^{1/2} \in L^2_{\mathcal{X},\mu} \mapsto$ density of $P \in L^1$ are continuous; see Bickel *et al.* 1993, p. 464.)

Proof. We shall present a geometric, or rather Pythagorean, setting in which the above integral of squares of lengths is the square of a length itself, and the same holds for the resulting sum of squares of lengths. This setting will be completed by the construction of a suitable $\bar{t}$ that plays the role t played in the preceding sections; the resulting truth translates into the van Trees inequality.

If $(x_\vartheta)_{\vartheta\in\mathbb{R}}$, $(y_\vartheta)_{\vartheta\in\mathbb{R}} \in (L^2_{\mathcal{X},\mu})^{\mathbb{R}}$ are Borel–Borel measurable, the function $\vartheta \in \mathbb{R} \mapsto \langle x_\vartheta, y_\vartheta\rangle_{L^2_{\mathcal{X},\mu}} \in \mathbb{R}$ is measurable by the separability of $L^2_{\mathcal{X},\mu}$, so that we may define

$$\langle (x_\vartheta)_{\vartheta\in\mathbb{R}}, (y_\vartheta)_{\vartheta\in\mathbb{R}}\rangle_w := \int_{\mathbb{R}} \langle x_\vartheta, y_\vartheta\rangle_{L^2_{\mathcal{X},\mu}} w(\vartheta)\,\mathrm{d}\vartheta;$$

this is a semi-inner product on the subspace of all measurable members $(x_\vartheta)_{\vartheta\in\mathbb{R}}$ for which $\|(x_\vartheta)_{\vartheta\in\mathbb{R}}\|_w := \langle (x_\vartheta)_{\vartheta\in\mathbb{R}}, (x_\vartheta)_{\vartheta\in\mathbb{R}}\rangle_w^{1/2}$ is finite. Mod out by $\{x : \|x\|_w = 0\}$ and obtain an inner product space. Let $\prod_w L^2_{\mathcal{X},\mu}$ denote a Hilbert space that contains this quotient space as an inner product subspace. Clearly, for all $\lambda \in \mathbb{R}$ there is a member of $\prod_w L^2_{\mathcal{X},\mu}$ which contains $(P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}}$. We denote it by $(P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}}$ again and observe that

$$\lambda \in \mathbb{R} \mapsto v_\lambda := ((P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}}, W_\lambda) \in \prod_w L^2_{\mathcal{X},\mu} \times L^2_{\mathbb{R},\nu} := H$$

is a mapping into the product H of the two Hilbert spaces $\prod_w L^2_{\mathcal{X},\mu}$ and $L^2_{\mathbb{R},\nu}$; this H is again a Hilbert space with the usual inner product $\langle (a, b), (c, d)\rangle = \langle a, c\rangle_w + \langle b, d\rangle_{L^2_{\mathbb{R},\nu}}$. Because every $\lambda \mapsto P_{\vartheta+\lambda}$, $\vartheta \in \mathbb{R}$, as well as $\lambda \mapsto W_\lambda$ is continuously differentiable at $\lambda = 0$, one could imagine that

(d1) the same holds for $\lambda \mapsto v_\lambda$, and
(d2) for the image $\mathrm{d}v_\lambda/\mathrm{d}\lambda$ of $1 \in \mathbb{R}$ under the derivative at λ one has

$$\frac{\mathrm{d}v_\lambda}{\mathrm{d}\lambda} \equiv \frac{\mathrm{d}((P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}}, W_\lambda)}{\mathrm{d}\lambda} = \left(\left(\frac{\mathrm{d}P_{\vartheta+\lambda}}{\mathrm{d}\lambda}\right)_{\vartheta\in\mathbb{R}}, \frac{\mathrm{d}W_\lambda}{\mathrm{d}\lambda}\right) \in H,$$

at least at $\lambda = 0$.

Let, in that case, the square of the length of the above vector at $\lambda = 0$ be denoted by $\frac{1}{4}I_{\lambda\mapsto v_\lambda}(0)$. Then one also has

$$I_{\lambda\mapsto v_\lambda}(0) = \int I_{\vartheta'\mapsto P_{\vartheta'}}(\vartheta) w(\vartheta)\,\mathrm{d}\vartheta \;+\; I_{\lambda\mapsto W_\lambda}(0) \neq 0.$$

Now, if $U \subset \mathbb{R}$ and $V \subset H$ are as in (i)–(iii) in Section 6 with $\vartheta_0 := 0$ etc., $v_\lambda \in V \mapsto \bar{t}v_\lambda \in H$ is some function on V, and $v_\lambda \mapsto \langle \bar{t}v_\lambda, v_\lambda\rangle$ would have a tangential derivative at v_0 that is represented by $2\bar{t}v_0$ and is equal to that of $v_\lambda \mapsto \lambda$ at v_0, then from the proof of Theorem 1 and (iii) in Section 6 it will be seen that

$$\|\bar{t}v_0\|^2 \geqslant \left(I_{\lambda\mapsto v_\lambda}(0)\right)^{-1},$$

which has the same right-hand side as the van Trees inequality. If, moreover,

$$\|\bar{t}v_0\|^2 = \mathrm{E}(T-\Theta)^2,$$

for a measurable $t : \mathcal{X} \to \mathbb{R}$, then the van Trees inequality would have been established for this t. We shall demonstrate that in order for a measurable $t : \mathcal{X} \to \mathbb{R}$ to provide a $v_\lambda \in V \mapsto \bar{t}v_\lambda \in H$ with the above properties, conditions (t1) and (t2) suffice. Indeed, let

$$\bar{t} := ((t - \mathrm{E}_\vartheta T)_{\vartheta\in\mathbb{R}},\ \vartheta \in \mathbb{R} \mapsto \vartheta - \mathrm{E}_\vartheta T),$$

and let the juxtaposition $\bar{t}v_\lambda$ denote the coordinatewise pointwise product $(((t - \mathrm{E}_\vartheta T)P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}},\ \vartheta \mapsto (\vartheta - \mathrm{E}_\vartheta T)W_\lambda(\vartheta))$. We show that $\bar{t}v_\lambda \in H$. First, $((t - \mathrm{E}_\vartheta T)P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}}$ is a member of $(L^2_{\mathcal{X},\mu})^{\mathbb{R}}$, because $\int_{\mathcal{X}} (tP_{\vartheta+\lambda})^2\,\mathrm{d}\mu = \mathrm{E}_{\vartheta+\lambda}T^2 < \infty$, and a measurable member because $P_\vartheta \in L^2_{\mathcal{X},\mu} \mapsto tP_\vartheta \in L^2_{\mathcal{X},\mu}$ is measurable (taking $t := 1_A$, $A \in \mathcal{A}$, makes $P_\vartheta \mapsto tP_\vartheta$ continuous, and if $t_n \to t$, $|t_n| \uparrow |t|$ pointwise, then $t_nP_\vartheta \to tP_\vartheta$ in $L^2_{\mathcal{X},\mu}$) and $\vartheta \mapsto (\mathrm{E}_\vartheta T,\ P_{\vartheta+\lambda}) \in \mathbb{R} \times L^2_{\mathcal{X},\mu} \mapsto (\mathrm{E}_\vartheta T)P_{\vartheta+\lambda} \in L^2_{\mathcal{X},\mu}$ is measurable as the composition of a measurable and a continuous map. Further, we have, with $p_\vartheta :=$ density of P_ϑ,

$$\begin{aligned}\|((t - \mathrm{E}_\vartheta T)P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}}\|^2_w &= \int_{\mathbb{R}} \left(\int_{\mathcal{X}} (t - \mathrm{E}_\vartheta T)^2 p_{\vartheta+\lambda}\,\mathrm{d}\mu\right) w(\vartheta)\,\mathrm{d}\vartheta \\ &= \int_{\mathbb{R}} \left(\mathrm{E}_{\vartheta+\lambda}T^2 - 2(\mathrm{E}_\vartheta T)(\mathrm{E}_{\vartheta+\lambda}T) + (\mathrm{E}_\vartheta T)^2\right) w(\vartheta)\,\mathrm{d}\vartheta,\end{aligned}$$

for which we observe $\int_{\mathbb{R}} \mathrm{E}_{\vartheta+\lambda}T^2 w(\vartheta)\,\mathrm{d}\vartheta = \int_{\mathbb{R}} \mathrm{E}_\vartheta T^2 w(\vartheta-\lambda)\,\mathrm{d}\vartheta = \mathrm{E}_{w_\lambda}T^2$,

$$\left(\int_{\mathbb{R}} (\mathrm{E}_\vartheta T)(\mathrm{E}_{\vartheta+\lambda}T) w(\vartheta)\,\mathrm{d}\vartheta\right)^2 \leqslant \int_{\mathbb{R}} (\mathrm{E}_\vartheta T)^2 w(\vartheta)\,\mathrm{d}\vartheta \int_{\mathbb{R}} (\mathrm{E}_{\vartheta+\lambda}T)^2 w(\vartheta)\,\mathrm{d}\vartheta,$$

and $(\mathrm{E}_\vartheta T)^2 \leqslant \mathrm{E}_\vartheta T^2$, so that

$$\|((t - \mathrm{E}_\vartheta T)P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}}\|^2_w \leqslant \mathrm{E}_{w_\lambda}T^2 + 2(\mathrm{E}_{w_0}T^2\mathrm{E}_{w_\lambda}T^2)^{1/2} + \mathrm{E}_{w_0}T^2,$$

bounded on a neighbourhood of $\lambda = 0$ by (t2); in particular, $((t - \mathrm{E}_\vartheta T)P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}} \in \prod_w L^2_{\mathcal{X},\mu}$ on this neighbourhood.

As to $\vartheta \mapsto (\vartheta - \mathrm{E}_\vartheta T)W_\lambda(\vartheta)$, the last coordinate of $\bar{t}v_\lambda$, we see that

$$\int_{\mathbb{R}} (\vartheta - \mathrm{E}_\vartheta T)^2 w_\lambda(\vartheta)\,\mathrm{d}\vartheta = \int_{\mathbb{R}} \left(\vartheta^2 - 2\vartheta\mathrm{E}_\vartheta T + (\mathrm{E}_\vartheta T)^2\right) w_\lambda(\vartheta)\,\mathrm{d}\vartheta,$$

for which we observe that $\int_{\mathbb{R}} \vartheta^2 w_\lambda(\vartheta)\,\mathrm{d}\vartheta = \int_{\mathbb{R}} (\vartheta+\lambda)^2 w_0(\vartheta)\,\mathrm{d}\vartheta = \mathrm{E}(\Theta+\lambda)^2 \leqslant 4\mathrm{E}(\Theta^2 + 1) < \infty$ for all $|\lambda| \leqslant 1$ (if $\mathrm{E}\Theta^2 = \infty$, then there is nothing to prove in van Trees), so that with the same arguments as above we may conclude that the $L^2_{\mathbb{R},\nu}$ norm of the last coordinate is bounded on a neighbourhood of $\lambda = 0$. In particular, the last coordinate is in $L^2_{\mathbb{R},\nu}$ on this neighbourhood. For $\bar{t}v_\lambda$ it follows not only that it is in H, but also that it is locally bounded at $\lambda = 0$, and therefore (Section 4; see the next paragraph) that $v_\lambda \mapsto \langle \bar{t}v_\lambda, v_\lambda\rangle$ has a tangential derivative at v_0 represented by $2\bar{t}v_0$. As to this derivative, inspection shows that

$$\begin{aligned}\langle \bar{t}v_\lambda, v_\lambda\rangle &= \int_{\mathbb{R}} \left(\int_{\mathcal{X}} (t - \mathrm{E}_\vartheta T) p_{\vartheta+\lambda} \, \mathrm{d}\mu \right) w(\vartheta) \, \mathrm{d}\vartheta + \int_{\mathbb{R}} (\vartheta - \mathrm{E}_\vartheta T) w_\lambda(\vartheta) \, \mathrm{d}\vartheta \\ &= \int_{\mathbb{R}} (\mathrm{E}_{\vartheta+\lambda} T - \mathrm{E}_\vartheta T) w(\vartheta) \, \mathrm{d}\vartheta + \mathrm{E}(\Theta + \lambda) - \int_{\mathbb{R}} (\mathrm{E}_\vartheta T) w_\lambda(\vartheta) \, \mathrm{d}\vartheta \\ &= \lambda - \mathrm{E}(T - \Theta),\end{aligned}$$

so it is equal to that of $v_\lambda \mapsto \lambda$. The equality $\|\bar{t}v_0\|^2 = \mathrm{E}(T - \Theta)^2$ follows from the norm calculations by taking $\lambda = 0$.

We still have to show that the reasoning of Section 4 which proved Lemma 6 extends to the new juxtaposition. For this, we may restrict ourselves to what happens in the Hilbert space $\prod_w L^2_{\mathcal{X},\mu}$ and prove that, now with

$$\bar{t} := (t - \mathrm{E}_\vartheta T)_{\vartheta\in\mathbb{R}}, \qquad v_\lambda := (P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}}, \qquad \bar{t}v_\lambda := \big((t - \mathrm{E}_\vartheta T) P_{\vartheta+\lambda}\big)_{\vartheta\in\mathbb{R}},$$

the map $v_\lambda \mapsto \langle \bar{t}v_\lambda, v_\lambda\rangle$ has a tangential derivative at v_0 represented by $2\bar{t}v_0$, if $\bar{t}v_{\lambda_n}$ is bounded when v_{λ_n} approaches v_0 in $\prod_w L^2_{\mathcal{X},\mu}$. *Mutatis mutandis*, so with v_0 instead of P, v_{λ_n} instead of P_n, $\bar{t}$ instead of t, and $\prod_w L^2_{\mathcal{X},\mu}$ instead of $L^2_{\mathcal{X},\mu}$, we repeat what we did for Lemma 6 and see that the first obstacle is where the old t is split into $t = t1_{|t|\leqslant c} + t1_{|t|>c}$. This time the splitting requires more care because of the desired measurability. Let $c > 0$, $A_i := t^{-1}([i - c, i + c])$ for all $i \in \mathbb{Z}$, and $\mathrm{E}_\vartheta T = i_\vartheta + r_\vartheta$ with $i_\vartheta \in \mathbb{Z}$ and $r_\vartheta \in [0, 1)$ for all $\vartheta \in \mathbb{R}$. Then $\vartheta \mapsto i_\vartheta$ and $\vartheta \mapsto r_\vartheta$ are measurable and if $(x_\vartheta)_{\vartheta\in\mathbb{R}} \in (L^2_{\mathcal{X},\mu})^{\mathbb{R}}$ is measurable, then $\vartheta \mapsto x_\vartheta 1_{A_{i_\vartheta}}$ and $\vartheta \mapsto x_\vartheta 1_{A^c_{i_\vartheta}}$ are also measurable, because $\{\vartheta \in \mathbb{R} : x_\vartheta 1_{A_{i_\vartheta}} \in B\} = \cup_i \{\vartheta \in \mathbb{R} : x_\vartheta 1_{A_i} \in B, i_\vartheta = i\}$ and $x \mapsto x1_{A_i}$ is continuous for each i. Define $1_{|\bar{t}|\leqslant} := (1_{A_{i_\vartheta}})_{\vartheta\in\mathbb{R}}$ and $1_{|\bar{t}|>} := (1_{A^c_{i_\vartheta}})_{\vartheta\in\mathbb{R}}$. Any measurable $(x_\vartheta)_{\vartheta\in\mathbb{R}} \in (L^2_{\mathcal{X},\mu})^{\mathbb{R}}$, then, can be split into a sum

$$(x_\vartheta)_{\vartheta\in\mathbb{R}} = (x_\vartheta)_{\vartheta\in\mathbb{R}} 1_{|\bar{t}|\leqslant} + (x_\vartheta)_{\vartheta\in\mathbb{R}} 1_{|\bar{t}|>}$$

such that the terms (coordinatewise pointwise products) are measurable; in our case

$$\langle h_n, \epsilon_n \bar{t} h_n\rangle_w = \langle h_n, \bar{t}1_{|\bar{t}|\leqslant}(v_{\lambda_n} - v_0)\rangle_w + \langle h_n 1_{|\bar{t}|>}, \bar{t}(v_{\lambda_n} - v_0)\rangle_w,$$

so that

$$|\langle h_n, \epsilon_n \bar{t} h_n\rangle_w| \leqslant \|h_n\|_w \cdot (c + 1)\|v_{\lambda_n} - v_0\|_w + \|h_n 1_{|\bar{t}|>}\|_w \cdot \|\bar{t}(v_{\lambda_n} - v_0)\|_w$$

and

$$\limsup_{n\to\infty} |\langle h_n, \epsilon_n \bar{t} h_n\rangle_w| \leqslant \|h\|_{\prod_w L^2_{\mathcal{X},\mu}} \cdot (c + 1) \cdot 0 + \|\lim_{n\to\infty} (h_n 1_{|\bar{t}|>})\|_{\prod_w L^2_{\mathcal{X},\mu}} \cdot B$$

(the limit in the last term exists because $(h_n 1_{|\bar{t}|>})_{n=1}^\infty$ is a Cauchy sequence in the inner product space); the juxtaposition $h1_{|\bar{t}|>}$ is not necessarily already defined. Let $\epsilon > 0$ and let n_1 be such that $\|h_m - h_n\|_w < \epsilon$ for all $m, n \geqslant n_1$; then

$$\|h_n 1_{|\bar{t}|>}\|_w \leqslant \|h_{n_1} 1_{|\bar{t}|>}\|_w + \|(h_{n_1} - h_n)1_{|\bar{t}|>}\|_w$$

and

$$\|\lim_{n\to\infty}(h_n 1_{|\bar{t}|>})\|_{\prod_w L^2_{\mathcal{X},\mu}} \leqslant \|h_{n_1} 1_{|\bar{t}|>}\|_w + \epsilon.$$

Two applications of Lebesgue's dominated convergence have the first term of the right-hand side converge to 0 as $c \to \infty$; we conclude that $\langle h_n, \epsilon_n \bar{t} h_n\rangle_w \to 0$ as $n \to \infty$. □

8. Sufficient conditions for van Trees

Here is a condition guaranteeing the desired behaviour (d1)–(d2).

Proposition 12. *In the situation of Theorem 11 let the following hold: there are an* $M : \mathbb{R} \to \mathbb{R}$ *and a* $\lambda_0 > 0$ *with* $\int_{\mathbb{R}} (M(\vartheta))^2 w(\vartheta)\,\mathrm{d}\vartheta < \infty$ *and*

$$\sup_{\xi\in[\vartheta-\lambda_0,\vartheta+\lambda_0]} \left\| \frac{\mathrm{d}P_\xi}{\mathrm{d}\xi} \right\| \leqslant M(\vartheta) \qquad \forall \vartheta \in \mathbb{R}.$$

Then $\lambda \in \mathbb{R} \mapsto \upsilon_\lambda \in H$ *is differentiable on a neighbourhood* N *of* 0 *with*

$$\frac{\mathrm{d}\upsilon_\lambda}{\mathrm{d}\lambda} = \left(\left(\frac{\mathrm{d}P_{\vartheta+\lambda}}{\mathrm{d}\lambda} \right)_{\vartheta\in\mathbb{R}}, \frac{\mathrm{d}W_\lambda}{\mathrm{d}\lambda} \right), \qquad \forall \lambda \in N,$$

and continuously differentiable at 0*, for any choice of* $\prod_w L^2_{\mathcal{X},\mu}$.

Proof. With $\pi : \vartheta \mapsto P_\vartheta \in L^2_{\mathcal{X},\mu}$ we have, by definition, $\mathrm{d}P_\xi/\mathrm{d}\xi = (\pi'(\xi))(1)$ and, by the chain rule, $\mathrm{d}P_{\vartheta+\lambda}/\mathrm{d}\lambda = (\pi'(\vartheta+\lambda))(1)$ for all $\vartheta, \lambda, \xi \in \mathbb{R}$, so that

$$\left\| \frac{\mathrm{d}P_{\vartheta+\lambda}}{\mathrm{d}\lambda} \right\| \leqslant M(\vartheta) \qquad \forall \lambda \in [-\lambda_0, \lambda_0],\ \vartheta \in \mathbb{R},$$

and therefore $(\mathrm{d}P_{\vartheta+\lambda}/\mathrm{d}\lambda)_{\vartheta\in\mathbb{R}} \in \prod_w L^2_{\mathcal{X},\mu}$ for all $\lambda \in [-\lambda_0, \lambda_0]$. Further, the mean value theorem (see Lenstra 1998) gives the inequality in

$$\begin{aligned} \left\| \frac{P_{\vartheta+\lambda+h} - P_{\vartheta+\lambda}}{h} \right\|_{L^2_{\mathcal{X},\mu}} &\leqslant \sup_{\xi\in[\vartheta+\lambda,\vartheta+\lambda+h]} \|\pi'(\xi)\|_{L(\mathbb{R},L^2_{\mathcal{X},\mu})} \\ &= \sup_{\xi\in[\vartheta+\lambda,\vartheta+\lambda+h]} \left\| \frac{\mathrm{d}P_\xi}{\mathrm{d}\xi} \right\|_{L^2_{\mathcal{X},\mu}}, \end{aligned}$$

where $L(\mathbb{R}, L^2_{\mathcal{X},\mu})$ is the Banach space of all bounded linear maps $\mathbb{R} \to L^2_{\mathcal{X},\mu}$, and we conclude that

$$\left\| \frac{P_{\vartheta+\lambda+h} - P_{\vartheta+\lambda}}{h} - \frac{\mathrm{d}P_{\vartheta+\lambda}}{\mathrm{d}\lambda} \right\|_{L^2_{\mathcal{X},\mu}} \leqslant 2M(\vartheta) \qquad \forall \vartheta \in \mathbb{R},\ \forall \lambda,\ h \in [-\lambda_0/2, \lambda_0/2],$$

so that by Lebesgue's dominated convergence the mapping $\pi_1 : \lambda \in (-\lambda_0/2, \lambda_0/2) \mapsto (P_{\vartheta+\lambda})_{\vartheta\in\mathbb{R}} \in \prod_w L^2_{\mathcal{X},\mu}$ is everywhere differentiable with $(\pi_1'(\lambda))(1) =$

$(\mathrm{d}P_{\vartheta+\lambda}/\mathrm{d}\lambda)_{\vartheta\in\mathbb{R}}$ for these λ. Again by Lebesgue we have $(\mathrm{d}P_{\vartheta+\lambda}/\mathrm{d}\lambda)_{\vartheta\in\mathbb{R}} \overset{\lambda\to 0}{\longrightarrow} (\mathrm{d}P_{\vartheta}/\mathrm{d}\vartheta)_{\vartheta\in\mathbb{R}}$ in $\prod_w L^2_{\mathcal{X},\mu}$. This says that the differentiability of π_1 is continuous at 0. The theorem now follows from $\upsilon_\lambda = (\pi_1(\lambda), W_\lambda)$ and the properties of π_1 and $\lambda \mapsto W_\lambda$. □

Appendix: Tangential differentiation

Let X and Y be normed linear spaces and U an open set in X. The map $f : U \to Y$ is *differentiable* at a point $x \in U$ if there exists a bounded linear map $f'(x) : X \to Y$, the *derivative of f at x*, such that

$$f(x+h) - f(x) = (f'(x))(h) + o(\|h\|), \qquad \|h\| \to 0,$$

or, equivalently,

$$\left\| \frac{f(x_n) - f(x)}{\epsilon_n} - (f'(x))\left(\frac{x_n - x}{\epsilon_n}\right) \right\|_Y \to 0, \qquad n \to \infty,$$

for all $(\epsilon_n)_{n=1}^\infty$ in $\mathbb{R}\backslash\{0\}$ and $(x_n)_{n=1}^\infty$ in U such that $\epsilon_n \to 0$ and $((x_n - x)/\epsilon_n)_{n=1}^\infty$ is bounded, so that, if it exists, $f'(x)$ is unique. Now let V be an arbitrary set in X, not necessarily open, and $x \in V$. If $(\epsilon_n)_{n=1}^\infty$ in $\mathbb{R}\backslash\{0\}$ and $(x_n)_{n=1}^\infty$ in V are such that $\epsilon_n \to 0$ and $\lim_{n\to\infty} (x_n - x)/\epsilon_n$ exists, then this limit is a *tangent vector of V at x*. The closed linear span of the tangent vectors is the *tangent space of V at x*.

Let $f : V \to Y$ be a map of V into Y and $\dot{V}$ be the tangent space of V at x. Suppose there exists a bounded linear map $f'(x) : \dot{V} \to Y$ for which

$$\left\| \frac{f(x_n) - f(x)}{\epsilon_n} - (f'(x))\left(\lim_{n\to\infty} \frac{x_n - x}{\epsilon_n}\right) \right\|_Y \to 0, \qquad n \to \infty,$$

for all $(\epsilon_n)_{n=1}^\infty$ in $\mathbb{R}\backslash\{0\}$ and $(x_n)_{n=1}^\infty$ in V such that $\epsilon_n \to 0$ and $((x_n - x)/\epsilon_n)_{n=1}^\infty$ converges. Then $f'(x)$ is the only bounded linear map of $\dot{V}$ into Y with this property, as there is uniqueness on the vectors that generate $\dot{V}$. The map f is said to be *differentiable at x along $\dot{V}$* or *tangentially differentiable at x*, and $f'(x)$ is the *derivative of f at x along $\dot{V}$* or the *tangential derivative of f at x*.

Two observations connect the two kinds of differentiability that we have exposed here. The first is obvious from the equivalence of 'bounded' to 'continuous' for linear maps, and the boundedness of convergent sequences.

Theorem A.1. *Let X and Y be normed linear spaces and U open in X, $V \subset U$. Let $x \in V$. If $f : U \to Y$ is differentiable at x, then $f|_V : V \to Y$ is differentiable at x along the tangent space $\dot{V}$ of V at x with derivative $(f|_V)'(x) = f'(x)|_{\dot{V}}$.*

In particular, differentiability at x implies differentiability at x along the tangent space $\dot{U}$ of U at x, that is, along X. On the other hand, as to the convergence of bounded sequences: in $\mathbb{R}^n$ every bounded sequence has a convergent subsequence. These facts lead to the second observation:

Theorem A.2. *If the domain of f is an open set in $X = \mathbb{R}^n$ and x lies in that domain, then differentiability of f at x and differentiability of f at x along X are equivalent.*

The first differentiability is often referred to as *Fréchet differentiability*, the second differentiability is known as *Hadamard differentiability along $\dot{V}$*. The latter is strong enough to support the chain rule:

Theorem A.3 (Chain rule). *Let X, Y and Z be normed linear spaces and $V \subset X$, $W \subset Y$. If $f : V \to W$ is tangentially differentiable at $x \in V$ and $g : W \to Z$ is tangentially differentiable at $f(x)$, then $g \circ f$ is tangentially differentiable at x and $(g \circ f)'(x) = g'(f(x)) \circ f'(x)$.*

Proof. Immediate from the definition, which shows that

$$\lim_{n\to\infty} \frac{f(x_n) - f(x)}{\epsilon_n} = (f'(x))\left(\lim_{n\to\infty} \frac{x_n - x}{\epsilon_n}\right),$$

so that under $f'(x)$ the image of a tangent vector of V at x is a tangent vector of $f(V)$, and therefore of W, at $f(x)$. □

Acknowledgements

This research was supported by The Netherlands Foundation for Scientific Research (NWO); it was carried out at Eurandom in Eindhoven, the Korteweg–de Vries Institute for Mathematics of the Universiteit van Amsterdam, and the Department of Mathematics and Statistics of Texas Tech University in Lubbock, TX. The author found the device of Section 2 while reading the introduction to Klaassen *et al.* (1988) as given in van der Vaart (1987). Richard Gill suggested including the van Trees inequality and made helpful comments, as did Chris Klaassen, Arnoud van Rooij and Aad van der Vaart.

References

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

Fabian, V. and Hannan, J. (1977) On the Cramér–Rao inequality. *Ann. Statist.*, **5**, 197–205.

Gill, R.D. and Levit, B.Y. (1995) Applications of the van Trees inequality: a Bayesian Cramér–Rao bound. *Bernoulli*, **1**, 59–79.

Groeneboom, P. and Wellner, J.A. (1992) *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser Verlag.

Ibragimov, I.A. and Has'minskii, R.Z. (1981) *Statistical Estimation*. New York: Springer-Verlag.

Janssen, A. (2003) A nonparametric Cramér–Rao inequality for estimators of statistical functionals. *Statist. Probab. Lett.*, **64**, 347–358.

Klaassen, C.A.J., van der Vaart, A.W. and van Zwet, W.R. (1988) On estimating a parameter and its

score function. In S.S. Gupta and J.O. Berger (eds), *Statistical Decision Theory and Related Topics IV*, Vol. 2, pp. 281–288. New York: Springer-Verlag.

Lang, S. (1995) *Differential and Riemannian Manifolds*. New York: Springer-Verlag.

Lenstra, A.J. (1998) On the efficient influence function and the efficient score function. In *Analyses of the Nonparametric Mixed Proportional Hazards Model*, Article V. Doctoral thesis, Universiteit van Amsterdam.

van der Vaart, A.W. (1987) *Statistical Estimation in Large Parameter Spaces*. Doctoral thesis, Rijksuniversiteit Leiden.

Received May 2002 and revised June 2004