M-estimators and their consistency

This handout is adapted from Section 3.3 of 18.466 lecture notes on mathematical statistics, available on OCW.

A sequence of estimators T_n, one for each sample size n, possibly only defined for n large enough, is called *consistent* if for $X_1, X_2, \ldots$, i.i.d. (P_θ), $T_n = T_n(X_1, \ldots, X_n)$ converges in probability as $n \to \infty$ to a function $g(\theta)$ being estimated. We will consider consistency of estimators more general than maximum likelihood estimators in two ways, first that the function being maximized may not be a likelihood, and second that it only needs to be approximately maximized.

It will be assumed that the parameter space Θ is a locally compact separable metric space with a metric d, such as an open or closed subset of a Euclidean space. $(X, \mathcal{A}, P)$ will be any probability space, and $h = h(\theta, x)$ is a measurable function on $\Theta \times X$ with values in the extended real number system $[-\infty, \infty]$. One example will be the negative of a log likelihood function, $h(\theta, x) \equiv -\log f(\theta, x)$. This will be called the *log likelihood case*. Let $X_1, X_2, \ldots$ be independent random variables with values in X and distribution P, specifically, coordinates on the countable product $(X^\infty, \mathcal{A}^\infty, P^\infty)$ of copies of $(X, \mathcal{A}, P)$ (RAP, Sec. 8.2). A statistic $T_n = T_n(X_1, ..., X_n)$ with values in Θ will be called an *M-estimator* if

$$\tfrac{1}{n}\sum_{i=1}^n h(T_n, X_i) \;=\; \inf_{\theta\in\Theta} \;\tfrac{1}{n}\sum_{i=1}^n h(\theta, X_i).$$

Thus, in the log likelihood case, an M-estimator is a maximum likelihood estimator.

The *outer probability* $P^*(C)$ of a not necessarily measurable set C is defined by

$$P^*(C) \;:=\; \inf\{P(A) :\; A \supset C,\; A \;\text{ measurable}\}.$$

Let f_n be a sequence of not necessarily measurable functions from a probability space into a metric space S with metric d. Then f_n is said to converge to f_0 *almost uniformly* if for every $\varepsilon > 0$, $P^*(\sup_{m\ge n} d(f_m, f_0) > \varepsilon) \to 0$ as $n \to \infty$. If $d(f_m, f_0)$ is a measurable random variable, as it will be in nearly all actual applications, then almost uniform convergence is the same as almost sure convergence.

Statistics $T_n = T_n(X_1, ..., X_n)$ with values in Θ will be called a sequence of *approximate M-estimators* if as $n \to \infty$,

$$\tfrac{1}{n}\sum_{i=1}^n h(T_n, X_i) - \inf_{\theta\in\Theta} \;\tfrac{1}{n}\sum_{i=1}^n h(\theta, X_i) \to 0 \tag{3.3.1}$$

almost uniformly.

It will be proved that T_n converges almost uniformly to some θ_0 under a list of assumptions as follows.

(A-1) $h(\theta, x)$ is a separable stochastic process, meaning that there is a set $A \subset X$ with $P(A) = 0$ and a countable subset $S \subset \Theta$ such that for every open set $U \subset \Theta$ and every closed set $J \subset [-\infty, \infty]$,

$$\{x :\; h(\theta, x) \in J \;\text{ for all }\; \theta \in S \cap U\} \subset A \cup \{x :\; h(\theta, x) \in J \;\text{ for all }\; \theta \in U\}.$$

This will be true with A empty if each function $h(\cdot, x)$ is continuous on Θ and S is dense in Θ, but the assumption is valid in more general situations. An alternate, equivalent formulation of separability is that for some countable S and almost all x, the graph of $h(\cdot, x)$ restricted to S is dense in the whole graph. For example. if Θ is an interval in $\mathbb{R}$, and for almost all x, $h(\cdot, x)$ is either left-continuous or right-continuous at each θ, then $h(\cdot, \cdot)$ is a separable process.

It is known that by changing $h(\theta, x)$ only for x in a set of probability 0 (depending on θ), one can assume that h is separable (by a theorem of Doob, proved in Appendix C, Theorem C.2 of the 18.466 notes). But in statistics, where the probability P is unknown, the separability is more clearly attainable in case h has at least a one-sided continuity property as just mentioned.

Instead of continuity, here is a weaker assumption:

(A-2) For each x in X, the function $h(\cdot, x)$ is lower semicontinuous on Θ, meaning that $h(\theta, x) \leq \liminf_{\phi \to \theta} h(\phi, x)$ for all θ.

Often, but not always, the functions $h(\cdot, x)$ will be continuous on Θ. Consider for example the uniform distributions $U[\theta, \theta + 1]$ on $\mathbb{R}$ for $\theta \in \mathbb{R}$. The density $f(\theta, x) := 1_{[\theta, \theta+1]}(x)$ is not continuous in θ, but it is upper semicontinuous,

$$f(\theta, x) \geq \limsup_{\phi \to \theta} f(\phi, x).$$

It follows that the functions $h(\theta, x) = -\log f(\theta, x)$ are lower semicontinuous (they have values $+\infty$ for $x \notin [\theta, \theta + 1]$). This is a reason for choosing the densities to be indicator functions of closed intervals; if we had taken $f(\theta, x) = 1_{(\theta, \theta+1)}(x)$, then $h(\theta, x)$ would no longer be lower semicontinuous.

For any real function f, as usual let $f^+ := \max(f, 0)$ and $f^- := -\min(f, 0)$. A function $h(\cdot, \cdot)$ of θ and x will be called *adjusted* for P if

$$Eh(\theta, x)^- < \infty \text{ for all } \theta \in \Theta, \text{ and} \tag{3.3.2}$$

$$Eh(\theta, x)^+ < \infty \text{ for some } \theta \in \Theta. \tag{3.3.3}$$

To say that h is adjusted is equivalent to saying that $Eh(\theta, \cdot)$ is well-defined (possibly $+\infty$) and not $-\infty$ for all θ, and for some θ, also $Eh(\theta, \cdot) < +\infty$, so it is some finite real number.

If $a(\cdot)$ is a measurable real-valued function on X such that $h(\theta, x) - a(x)$ is adjusted for P, then $h(\cdot, \cdot)$ will be called *adjustable* for P and $a(\cdot)$ will be called an *adjustment function* for h and P. The next assumption is:

(A-3) $h(\cdot, \cdot)$ is adjustable for P.

From here on, if $h(\theta, x)$ is adjustable but not adjusted, let $\gamma(\theta) := \gamma_a(\theta) := E[h(\theta, x) - a(x)]$ for a suitable adjustment function $a(\cdot)$. As an example, let $h(\theta, x) := |x - \theta|$ for $\theta, x \in \mathbb{R}$. If P is a law on $\mathbb{R}$, such as the Cauchy distribution with density $(\pi(1 + x^2))^{-1}$, with $\int |x| dP(x) = +\infty$, then h itself is not adjusted and an adjustment function is needed. Let $a(x) := |x|$ in this case. Then for each θ, $|x - \theta| - |x|$ is bounded in absolute value (by $|\theta|$), so $\gamma(\theta)$ is defined and finite for all θ. Thus $|x|$ is in fact an adjustment function for any P.

The example illustrates an idea of Huber (1967,1981) who seems to have invented the notion of adjustment. An estimator is defined by minimizing or approximately minimizing $\frac{1}{n}\sum_{i=1}^{n} h(\theta, X_i)$. If $\int h(\theta, x)dP(x)$ is finite, it is the limit of the sample averages by the strong law of large numbers. But if it isn't finite, it may be made finite by subtracting an adjustment function $a(x)$ from h. Since $a(\cdot)$ doesn't depend on θ, this change doesn't affect the minimization for each n. Thus, such estimators can be treated for more general probability measures P which on the real line, for example, can have long tails, allowing robust estimation. In fact, in the last example, P can be an arbitrary (and so arbitrarily heavy-tailed) distribution on $\mathbb{R}$.

3.3.4 Proposition. If a_1 is an adjustment function for $h(\cdot, \cdot)$ and P, then another measurable real-valued function $a_2(\cdot)$ on X is also an adjustment function if and only if $a_1 - a_2$ is integrable for P, and $\{\theta : \ \gamma(\theta) \in \mathbb{R}\}$ does not depend on the choice of adjustment function $a(\cdot)$.

Proof. "If" is clear. To prove "only if," we have $E((h(\theta, x) - a_i(x))^-) < \infty$ for all θ and $i = 1, 2$, while $E((h(\theta_i, x) - a_i(x))^+) < \infty$ for some θ_i and $i = 1, 2$. We can write for $\theta = \theta_1$ or θ_2,

$$(a_1 - a_2)(x) \ = \ h(\theta, x) - a_2(x) - [h(\theta, x) - a_1(x)]$$

for P-almost all x. To check this we need to take account that h can have values $\pm\infty$. For any θ, $h(\theta, x) > -\infty$ for P-almost all x since h is adjustable. We have $h(\theta_1, x) < +\infty$ and $h(\theta_2, x) < +\infty$ for P-almost all x. Thus the given expression for $(a_1 - a_2)(x)$ is well-defined for P-almost all x and $\theta = \theta_1$ or θ_2. We then have

$$E((a_1 - a_2)^+) \ \leq \ E[(h(\theta_2, x) - a_2(x))^+] + E[(h(\theta_2, x) - a_1(x))^-] \ < \ \infty,$$

$$E((a_1 - a_2)^-) \ \leq \ E[(h(\theta_1, x) - a_2(x))^-] + E[(h(\theta_1, x) - a_1(x))^+] \ < \ \infty,$$

so $E|a_1 - a_2| < \infty$ as stated. Thus, the sets of θ for which $E((h(\theta, x) - a_i(x))^+) < \infty$, or equivalently $E|h(\theta, x) - a_i(x)| < \infty$, don't depend on i, as stated. This finishes the proof of the proposition. □

The next assumption is:

(A-4) There is a $\theta_0 \in \Theta$ such that $\gamma(\theta) > \gamma(\theta_0)$ for all $\theta \neq \theta_0$.

Here θ_0 is called the *M-functional* of P. In the log likelihood case it is sometimes called the *pseudo-true* value of θ. Then $h(\theta, x) = -\log f(\theta, x)$ where for fixed θ, f is a density or probability mass function for a probability measure P_θ. The distribution P of the observations may not be in the parametric family of laws P_θ, and if not, no true value of θ exists, but often a pseudo-true value exists.

By Proposition 3.3.4, θ_0 does not depend on the choice of adjustment function. After some more assumptions, it will be shown that T_n converges to θ_0.

If Θ is not compact, let ∞ be the point adjoined in its one-point compactification (RAP, 2.8.1) and let $\liminf_{\theta\to\infty}$ mean $\sup_K \inf_{\theta\notin K}$ where the supremum is over all compact K. The next assumption is

(A-5) For some adjustment function $a(\cdot)$, there is a continuous function $b(\cdot) > 0$ on Θ such that

$$\inf\{(h(\theta,x) - a(x))/b(\theta) : \ \theta \in \Theta\} \ \geq \ -u(x) \tag{3.3.5}$$

for some integrable function $u(\cdot) \geq 0$, and if Θ is not compact, then

$$\liminf_{\theta\to\infty} b(\theta) \ > \ \gamma_a(\theta_0) \ \text{ and} \tag{3.3.6}$$

$$E\{\liminf_{\theta\to\infty}(h(\theta,x) - a(x))/b(\theta)\} \ \geq \ 1. \tag{3.3.7}$$

This completes the list of assumptions. Here (3.3.5) and (3.3.7) may depend on the choice of adjustment function. In the example where $X = \Theta = \mathbb{R}$, $h(\theta,x) = |x-\theta|$ and $a(x) := |x|$, all the assumptions hold if $b(\theta) := |\theta| + 1$ and P is any law on $\mathbb{R}$ with a unique median. Consistency, to be proved below, will imply that sample medians converge to the true median in this case.

Some consequences of the assumptions will be developed. The first one follows directly from Proposition 3.3.4 and the definitions:

3.3.8 Lemma. For any adjustable $h(\cdot,\cdot)$ and adjustment function $a(\cdot)$ for it, and any $\theta \in \Theta$ for which $\gamma_a(\theta) \in \mathbb{R}$, $h(\theta,\cdot)$ is also an adjustment function.

A sequence of sets $U_k \subset \Theta$ will be said to converge to a point θ if $\sup\{d(\theta,\phi) : \ \phi \in U_k\} \to 0$ as $k \to \infty$. Next, we have

3.3.9 Lemma. If (A-1), (A-2), and (A-3) hold and $a(\cdot)$ is an adjustment function for which (3.3.5) holds, with $b(\cdot)$ continuous, then

(A-2′) for any θ, as an open neighborhood U_k of θ converges to $\{\theta\}$,

$$E(\inf\{h(\phi,x) - a(x) : \ \phi \in U_k\}) \ \to \ \gamma(\theta) \leq +\infty.$$

Proof. Separability (A-1) applied to sets $J = [q,+\infty)$ for all rational q and joint measurability of $h(\cdot,\cdot)$ imply that the infimum in (A-2′) is equal almost surely to a measurable function of x. By (A-2), the integrand on the left converges to $h(\theta,x) - a(x)$, and it is larger for smaller neighborhoods U_k, so in this sense the convergence is monotone. Since $b(\cdot)$ is continuous and positive, it is bounded on any neighborhood U_k with compact closure, say $0 < b(\phi) \leq M$ for all $\phi \in U_k$. Then by (3.3.5), $h(\phi,x) - a(x) \geq -Mu(x)$ for all $\phi \in U_k$ and all x. Thus the stated convergence holds by monotone convergence (RAP, 4.3.2) for a fixed sequence of neighborhoods of θ such as $\{\phi : \ d(\phi,\theta) < 1/n\}$ where d is a metric for the topology of Θ. So, for any $\varepsilon > 0$, there is a neighborhood U_k of θ such that the expression being shown to converge is larger than $\gamma(\theta) - \varepsilon$ if $\gamma(\theta)$ is finite, or larger than $1/\varepsilon$ if $\gamma(\theta) = +\infty$, and the same will hold for any smaller neighborhood. □

Note that (3.3.1), the definition of approximate M-estimator, is not affected by subtracting $a(x)$ from $h(\theta,x)$.

By the alternate formulation given for separability (A-1), $h(\theta,x) - a(x)$ is separable and since $b(\theta)$ is continuous and strictly positive, $(h(\theta,x) - a(x))/b(\theta)$ is also separable.

For any adjustable $h(\cdot,\cdot)$ and adjustment function $a(\cdot)$ for it, let $h_a(\theta,x) := h(\theta,x) - a(x)$. If (A-5) holds, this notation will mean that $a(\cdot)$ has been chosen so that it holds.

3.3.10 Lemma. If (A-1), (A-3), (A-4), and (A-5) hold, then there is a compact set $C \subset \Theta$ such that for every sequence T_n of approximate M-estimators, almost surely there will be some n_0 such that $T_n \in C$ for all $n \geq n_0$, in the sense that

$$1_{\{T_n \in C\}} \to 1 \quad \text{almost uniformly as } n \to \infty. \tag{3.3.11}$$

Proof. If Θ is compact there is no problem. Otherwise, by (3.3.6) there is a compact C and an ε with $0 < \varepsilon < 1$ such that

$$\inf\{b(\theta) : \ \theta \notin C\} \ \geq \ (\gamma(\theta_0) + \varepsilon)/(1 - \varepsilon).$$

(Note: the $1 - \varepsilon$ in the denominator is useful when $\gamma(\theta_0) + \varepsilon > 0$ and otherwise makes little difference as $\varepsilon \downarrow 0$.) By (3.3.5), (3.3.7), (A-1), and monotone convergence as in the last proof, C can be chosen large enough so that

$$E(\inf\{h_a(\theta, x)/b(\theta) : \ \theta \notin C\}) \ \geq \ 1 - \varepsilon/2.$$

Then by the strong law of large numbers (RAP, Sec. 8.3), where a function with expectation $+\infty$ can be replaced by a smaller function with large positive expectation, a.s. for n large enough

$$\tfrac{1}{n} \inf\{\textstyle\sum_{i=1}^n h_a(\theta, X_i)/b(\theta) : \ \theta \notin C\} \ \geq \ \tfrac{1}{n} \textstyle\sum_{i=1}^n \inf\{h_a(\theta, X_i)/b(\theta) : \ \theta \notin C\} \ > \ 1 - \varepsilon.$$

Note that the infima are measurable since by separability of $h(\cdot, \cdot)$, measurability of $a(\cdot)$ and continuity of $b(\cdot)$, they can be restricted to a countable (dense) set in the complement of C. Then for any $\theta \notin C$,

$$\tfrac{1}{n} \textstyle\sum_{i=1}^n h_a(\theta, X_i) \ \geq \ (1 - \varepsilon) b(\theta) \ \geq \ \gamma(\theta_0) + \varepsilon. \tag{3.3.12}$$

On the other hand, for n large enough

$$\inf_\theta \tfrac{1}{n} \textstyle\sum_{i=1}^n h_a(\theta, X_i) \ \leq \ \tfrac{1}{n} \textstyle\sum_{i=1}^n h_a(\theta_0, X_i) \ \leq \ \gamma(\theta_0) + \varepsilon/2,$$

so as soon as the expression in (3.3.1) is less than $\varepsilon/2$, the same will hold for h_a since terms $a(X_i)$ cancel, and $T_n \in C$. □

3.3.13 Theorem. Let $\{T_n\}$ be a sequence of approximate M-estimators. Assume either (a) (A-1) through (A-5) hold, or (b) (A-1), (A-2′), (A-3) and (A-4) hold, and for some compact C, (3.3.11) holds. Then $T_n \to \theta_0$ almost uniformly.

Proof. Assumptions (a) imply (A-2′) by Lemma 3.3.9, and (3.3.11) by Lemma 3.3.10. So assumptions (b) hold in either case. By (3.3.11), Θ can be assumed to be a compact set C: take any point ψ of C and when T_n takes a value outside of C, redefine it as ψ. It can also be assumed that $\theta_0 \in C$ by adjoining it if necessary, and the proof below will show that θ_0 had to be in C.

Let U be an open neighborhood of θ_0. It follows from (A-2′) that $\gamma(\cdot)$ is lower semi-continuous. Thus its infimum on the compact set $C \setminus U$ is attained: let θ_k be a sequence in $C \setminus U$ on which γ converges to its infimum; we can assume that θ_k converges to some θ_∞, and then γ attains its minimum on $C \setminus U$ at θ_∞. By (A-4), $\inf_{C \setminus U} \gamma = \gamma(\theta_\infty) > \gamma(\theta_0)$.

Let $\varepsilon := (\gamma(\theta_\infty) - \gamma(\theta_0))/4$, or if $\gamma(\theta_\infty) = +\infty$ let $\varepsilon := 1$. By (A-2′), each $\theta \in C \setminus U$ has an open neighborhood U_θ such that

$$E(\inf\{h_a(\phi, x) : \ \phi \in U_\theta\}) \ \geq \ \gamma(\theta_0) + 3\varepsilon.$$

Again, the infimum is measurable since by separability it can be restricted to a countable dense set in U_θ. Take finitely many points $\theta(j)$, $j = 1, \ldots, N$, such that the neighborhoods $U_j := U_{\theta(j)}$ cover $C \setminus U$. By the strong law of large numbers, as in the proof of Lemma 3.3.10, we have a.s. for n large enough and each $j = 1, \ldots, N$,

$$\inf\{\tfrac{1}{n}\textstyle\sum_{i=1}^n h_a(\phi, X_i) : \ \phi \in U_j\} \ \geq \ \tfrac{1}{n}\textstyle\sum_{i=1}^n \inf\{h_a(\phi, X_i) : \ \phi \in U_j\} \ \geq \ \gamma(\theta_0) + 2\varepsilon$$

and $n^{-1}\sum_{i=1}^n h_a(\theta_0, X_i) \ \leq \ \gamma(\theta_0) + \varepsilon$. It follows that

$$\begin{aligned} \inf\{\tfrac{1}{n}\textstyle\sum_{i=1}^n h_a(\theta, X_i) : \ \theta \in C \setminus U\} \ &\geq \ \tfrac{1}{n}\textstyle\sum_{i=1}^n h_a(\theta_0, X_i) + \varepsilon \\ &\geq \ \inf\{\tfrac{1}{n}\textstyle\sum_{i=1}^n h_a(\theta, X_i) : \ \theta \in U\} + \varepsilon, \end{aligned} \tag{3.3.14}$$

so $\Pr\{T_n \in U \text{ for all } n \text{ large enough}\} = 1$. This completes the proof. □

Next let's recall the notion of likelihood ratio. Let P and Q be two probability measures on the same sample space S. Then there always exists some measure μ such that both P and Q have densities with respect to μ, where μ is a σ-finite measure, in other words there is a countable sequence of sets A_n whose union is all of S with $\mu(A_n) < \infty$ for each n. For example, if the sample space is a Euclidean space $\mathbb{R}^d$ and P and Q both have densities, then we can take μ to be Lebesgue measure (volume), $d\mu(x) = dx_1 dx_2 \cdots dx_d$. If P and Q are both discrete probabilities concentrated on a countable set S such as the nonnegative integers, we can take μ to be counting measure on S, where $\mu(A)$ is the number of elements in A for any $A \subset S$. In complete generality, we can always take $\mu = P + Q$, by the Radon-Nikodym theorem in measure theory.

Suppose then that P has a density $f = dP/d\mu$ and Q has a density $g = dQ/d\mu$ with respect to μ. Then the likelihood ratio of Q to P is defined as $R_{Q/P}(x) = g(x)/f(x)$, or as $+\infty$ if $g(x) > f(x) = 0$, or as 0 if $g(x) = f(x) = 0$. Then the likelihood ratio is well-defined and unique in the sense that if R and S are two functions with the properties of $R_{Q/P}$, possibly defined for different μ's, then $R = S$ except possibly on some set A with $P(A) = Q(A) = 0$. This is shown in Appendix A of the 18.466 Mathematical Statistics notes on the MIT OCW site.

To apply Theorem 3.3.13 to the case of maximum likelihood estimation the following will help. Let P and Q be two laws on a sample space $(X, \mathcal{B})$. Let

$$I(P, Q) \ := \ \int \log(R_{P/Q}) dP \ = \ -\int \log(R_{Q/P}) dP,$$

called the *Kullback-Leibler* information of P with respect to Q. Here we have $R_{P/Q} \equiv 1/R_{Q/P}$ with $1/0 := +\infty$ and $1/+\infty := 0$.

3.3.15 Theorem. Let $(X, \mathcal{B})$ be a sample space and P, Q any two laws on it. Then $I(P, Q) \geq 0$ and $I(P, Q) = 0$ if and only if $P = Q$.

Proof. By derivatives, it's easy to check that $\log x \le x-1$ for all $x \ge 0$, with $\log x = x-1$ if and only if $x = 1$. Thus

$$I(P,Q) \;=\; \int -\log(R_{Q/P})dP \;\ge\; \int 1 - R_{Q/P}dP \;\ge\; 0,$$

with equality if and only if $R_{Q/P} = 1$ a.s. for P, and then $Q = P$. □

Although $I(P,Q)$ is sometimes called a metric or distance, it is not symmetric in P and Q, nor does it satisfy the triangle inequality.

Consistency of approximate maximum likelihood estimators, under suitable conditions, does follow from Theorem 3.3.13, and assumption (A-3), and (A-4) for the true θ_0, will follow from Theorem 3.3.15 rather than having to be assumed:

3.3.16 Theorem. Assume (A-1) holds in the log likelihood case, for a measurable family $\{P_\theta,\ \theta \in \Theta\}$ dominated by a σ-finite measure v, with $(dP_\theta/dv)(x) \;=\; f(\theta,x)$, so that $h(\theta,x) \;:=\; -\log f(\theta,x)$. Also suppose $P = P_{\theta_0}$ for some $\theta_0 \in \Theta$ and $P_{\theta_0} \ne P_\theta$ for any $\theta \ne \theta_0$. Then (A-3) holds and (A-4) holds for the given θ_0. Assume T_n are approximate maximum likelihood estimators, i.e. approximate M-estimators in this case. If (A-2) and (A-5) also hold, or (A-2′) and (3.3.11), then the T_n are consistent.

Proof. If (A-1) through (A-5) hold then (A-2′) and (3.3.11) hold by Lemmas 3.3.9 and 3.3.10, and then Theorem 3.3.13 applies. So just (A-3) and (A-4) need to be proved. Let $a(x) \;:=\; -\log f(\theta_0,x)$. We have $0 < f(\theta_0,x) < \infty$ a.s. for P, and so $-\infty < \log f(\theta_0,x) < \infty$. Thus $h(\theta,x) - a(x)$ is well-defined a.s. and equals

$$-\log(f(\theta,x)/f(\theta_0,x)) \;=\; -\log R_{P_\theta/P_{\theta_0}}$$

as shown in Appendix A of the 18.466 OCW notes. Thus for all θ,

$$\gamma(\theta) \;:=\; E[h(\theta,x) - a(x)] \;=\; I(P_{\theta_0}, P_\theta) \;\ge\; 0 \;>\; -\infty$$

by Theorem 3.3.15 and $\gamma(\theta_0) = 0$, so (A-3) holds. Also by Theorem 3.3.15, $\gamma(\theta) = 0$ only for $\theta = \theta_0$, so (A-4) also holds. □

PROBLEMS

1. Let $h(\theta,x) = (x-\theta)^2$ for $x, \theta \in \mathbb{R}$.
   (a) Show that h is adjustable for a law P if and only if $\int |x| dP(x) < \infty$.
   (b) Show that then (A-4) holds and evaluate θ_0.
   (c) Show that for some $a(\cdot)$, (A-5) holds in this case for $b(\theta) = \theta^2 + 1$.
2. Recall that for a law P on $\mathbb{R}$, a point m is a *median* of P iff both $P((-\infty, x]) \ge 1/2$ and $P([x, +\infty)) \ge 1/2$. Thus if P is a continuous distribution without atoms, m is a median if and only if $P((-\infty, m]) = 1/2$. If P is any law on $\mathbb{R}$ having a unique median θ_0 and $h(\theta,x) \;:=\; |x-\theta|$, show that conditions (A-1) through (A-5) hold for some $a(\cdot)$ and $b(\cdot)$ (suggested in the text).

NOTES

An early result relating to consistency of maximum likelihood estimators was given by Cramér (1946), §33.3, namely, that under some hypotheses, there exist roots of the likelihood equation(s) converging in probability to the true value θ_0. If there are multiple roots, it was not clear how to select roots that would converge, but in case there was a unique root and it gave a maximum of the likelihood (as with exponential families), Cramér's theorem gave consistency of maximum likelihood estimates under his conditions.

Wald (1949) proved consistency of maximum likelihood estimates under some conditions. The present forms of the theorems and proofs through 3.3.13 are essentially as in Huber (1967). Dudley (1998) gave an extension, replacing the local compactness assumption by a uniform law of large numbers assumption. Kullback and Leibler (1951) defined their information and gave Theorem 3.3.15. Kullback (1983) gives an update.

REFERENCES

Cramér, Harald (1945). *Mathematical Methods of Statistics*. Almqvist & Wicksells, Uppsala, Sweden; Princeton University Press, 1946; 10th printing 1963.

Dudley, R. M. (1998). Consistency of M-estimators and one-sided bracketing. In *High Dimensional Probability*, *Progress in Probability* **43**, Birkhäuser, Basel.

Haughton, D. M.-A. (1983). On the choice of a model to fit data from an exponential family. Ph. D. dissertation, Mathematics, M.I.T.

Haughton, D. M.-A. (1988). On the choice of a model to fit data from an exponential family. *Ann. Statist.* **16**, 342-355.

Hoffman, K. (1975). *Analysis in Euclidean Space*. Prentice-Hall, Englewood Cliffs, NJ.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** (Univ. of Calif. Press, Berkeley and Los Angeles), 221-233.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Kullback, S. (1983). Kullback information. In *Encyclopedia of Statistical Sciences* **4**, pp. 421-425, Eds. S. Kotz, N. L. Johnson. Wiley, New York.

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79-86.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.* **20**, 595-601.