The Annals of Probability
1979, Vol. 7, No. 5, 882–886

ESTIMATION OF A CONVEX REAL PARAMETER OF AN UNKNOWN INFORMATION SOURCE

BY JOHN C. KIEFFER

University of Missouri-Rolla

Let $\mathcal{P}$ be the family of all stationary information sources with alphabet A. Let $F : \mathcal{P} \to (-\infty, \infty)$ be convex and upper semicontinuous in the weak topology. It is shown that for $n = 1, 2, \cdots$, there is an estimator $Y_n : A^n \to (-\infty, \infty)$, such that if $\mu \in \mathcal{P}$ is ergodic and the process $(X_1, X_2, \cdots)$ has distribution μ, then $Y_n(X_1, \cdots, X_n) \to F(\mu)$ in L^1 mean.

Let A be a finite set with a elements. Let $\mathcal{A}$ denote the power set of A. Let $(A^\infty, \mathcal{A}^\infty)$ be the measurable space consisting of A^∞, the set of all sequences $(x_1, x_2, \cdots)$ from A, and $\mathcal{A}^\infty$, the usual product σ-field. For each $i = 1, 2, \cdots$, let $X_i : A^\infty \to A$ be the coordinate map such that $X_i(x_1, x_2, \cdots) = x_i$. For $n = 1, 2, \cdots$, let $X^n : A^\infty \to A^n$ be the map $X^n = (X_1, \cdots, X_n)$. Let $T : A^\infty \to A^\infty$ be the shift transformation. Let R be the set of real numbers. Let $\mathcal{P}$ be the family of all T-stationary probability measures on $\mathcal{A}^\infty$. In information-theoretic terms, the elements of $\mathcal{P}$ are called stationary information sources. We topologize $\mathcal{P}$ with the weak topology. Let $\mathcal{P}_e$ be the set of T-ergodic measures in $\mathcal{P}$.

Let $F : \mathcal{P} \to R$ be given. We will call F *estimable in the mean* if for each $n = 1, 2, \cdots$, there exists $Y_n : A^n \to R$ such that for every $\mu \in \mathcal{P}_e$, $Y_n(X^n) \to F(\mu)$ in $L^1(\mu)$ mean. We will call F *estimable almost surely* if the estimators $\{Y_n\}_{n=1}^\infty$ can be found so that $Y_n(X^n) \to F(\mu)$ a.s. $[\mu]$, for any $\mu \in \mathcal{P}_e$.

We state the two main results to be proved in this paper.

THEOREM 1. *Let $F : \mathcal{P} \to R$ be continuous. Then F is estimable almost surely.*

THEOREM 2. *Let $F : \mathcal{P} \to R$ be upper semicontinuous and convex. Then F is estimable in the mean. Moreover, the estimators $\{Y_n\}$ may be chosen so that*

(a) $\limsup_{n\to\infty} Y_n(X^n) = F(\mu)$ a.s. $[\mu]$, $\mu \in \mathcal{P}_e$.

(b) $\int Y_n(X^n)d\mu \geqslant F(\mu)$ *for all n and all $\mu \in \mathcal{P}$.*

If in addition we assume that F is affine, the estimators $\{Y_n\}$ can be chosen so that (a), (b) *and the following hold*:

(c) $\int Y_n(X^n)d\mu \downarrow F(\mu)$, $\mu \in \mathcal{P}$.

Theorems 1 and 2 may be used to give a solution to the variable-rate variable-distortion universal source coding problem of information theory. As pointed out in Kieffer (1978), anytime the desired distortion levels (or rate levels) of the sources can be estimated in the mean, then variable-rate variable-distortion universal coding is possible.

Received February 1, 1978.

AMS 1970 *subject classifications.* Primary 94A15, 60G10; secondary 28A65.

Key words and phrases. Ergodic information source, upper-semicontinuous and convex function of a source, sequence of estimators, weak topology.

As another application, note that the entropy function $H : \mathcal{P} \to R$ is upper semicontinuous and affine. Thus the entropy function is estimable in the mean. Bailey (1976) has shown in addition that H is estimable almost surely. It would be interesting to know whether in Theorem 2 estimation almost surely is possible.

Before proceeding with the proofs of our main results we give a couple of interesting corollaries of Theorem 2.

COROLLARY 1. *Let* $\mathcal{K} \subset \mathcal{P}$ *satisfy*:
(a) $\mathcal{K}$ *is closed*;
(b) *if* $\mu, \nu \in \mathcal{P}$ *and* $0 < \alpha < 1$ *and* $\alpha\mu + (1 - \alpha)\nu \in \mathcal{K}$, *then* $\mu, \nu \in \mathcal{K}$.
Then $I_{\mathcal{K}}$, *the indicator function of* $\mathcal{K}$, *is estimable in the mean.*

One can thus determine whether a given $\mu \in \mathcal{P}_e$ is in $\mathcal{K}$ or not. For one can construct estimators $\{Y_n\}$ so that $Y_n \to 1$ in $L^1(\mu)$ mean if $\mu \in \mathcal{K}$ and $Y_n \to 0$ in $L^1(\mu)$-mean if $\mu \notin \mathcal{K}$.

EXAMPLES. 1. Take $\mathcal{K} \subset \mathcal{P}_e$, $\mathcal{K}$ closed. Then assumption (b) of Corollary 1 is automatically satisfied, since $\mathcal{P}_e$ is the set of extreme points of the convex set $\mathcal{P}$.

2. Take $\mathcal{K}$ to be the family of all memoryless sources in $\mathcal{P}$. Then $\mathcal{K}$ is closed and $\mathcal{K} \subset \mathcal{P}_e$.

3. For $N = 1, 2, \cdots$, take $\mathcal{K}$ to be the family of all N-Markovian sources in $\mathcal{P}$. Then $\mathcal{K}$ is closed. If $\alpha\mu + (1 - \alpha)\nu \in \mathcal{K}$, then all the ergodic components of μ, ν must be N-Markovian with the same transition probabilities. Thus $\mu, \nu \in \mathcal{K}$.

4. Take $\mathcal{K}$ to be the family of all periodic sources in $\mathcal{P}$ with period N. (μ is periodic with period N if $\mu\{\omega : T^N\omega = \omega\} = 1$.)

We remark that for Examples 2 and 3 Bailey (1976) obtained the stronger result that $I_{\mathcal{K}}$ is estimable almost surely.

DEFINITION. Let $(\Lambda, \mathcal{F})$ be a measurable space. A family $\{\mu_\theta : \theta \in \Lambda\} \subset \mathcal{P}_e$ is said to be *regular* if for each $E \in \mathcal{A}^\infty$, the map $\theta \to \mu_\theta(E)$ from Λ to R is $\mathcal{F}$-measurable.

COROLLARY 2. *Let* $(\Lambda, \mathcal{F}, \lambda)$ *be a probability space. Let* $\{\mu_\theta : \theta \in \Lambda\}$ *be a regular family. Let* $\mu \in \mathcal{P}$ *be the measure such that*
(a) $\mu(E) = \int \mu_\theta(E) d\lambda(\theta)$, $E \in \mathcal{A}^\infty$.
Let $F : \mathcal{P} \to R$ *be upper semicontinuous, bounded, and affine. Then* $F(\mu) = \int_\Lambda F(\mu_\theta) d\lambda(\theta)$.

PROOF. Find $\{Y_n\}$ which estimate F in the mean and satisfy (a)–(c) of Theorem 2. Since F is bounded we can assume the Y_n's are uniformly bounded. For each $\theta \in \Lambda$, $\int Y_n d\mu_\theta \to F(\mu_\theta)$ and the convergence is bounded, so $\int_\Lambda [\int Y_n d\mu_\theta] d\lambda(\theta) \to \int_\Lambda F(\mu_\theta) d\lambda(\theta)$. But $\int_\Lambda [\int Y_n d\mu_\theta] d\lambda(\theta) = \int Y_n d\mu$. (This follows from (a) above.) But $\int Y_n d\mu \to F(\mu)$ by (c) of Theorem 2. The result follows.

Corollary 2 is a result of Jacobs (1962). It has important information-theoretic applications. See Kieffer (1975) and Gray and Davisson (1974).

The rest of the paper will consist of the proofs of Theorems 1 and 2.

DEFINITIONS. If $\omega \in A^n$ and $n > m$ let $p(\omega, m)$ be the probability measure on A^m such that the $p(\omega, m)$-probability of a block $b \in A^m$ is the frequency with which that block appears in ω. If $\mu \in \mathcal{P}$, and $n = 1, 2, \cdots$, let μ_n be the probability measure on A^n such that $\mu_n(b) = \mu(X^n = b)$, $b \in A^n$. Note that by the ergodic theorem, if $\mu \in \mathcal{P}_e$ and m is fixed, then $p(X^n, m) \to \mu_m$ a.s. $[\mu]$. Define a probability measure p on A^n to be *invariant* if for every $1 \leqslant m < n$ and $b \in A^m$, $\Sigma_{b' \in A^{n-m}} p(b, b') = \Sigma_{b' \in A^{n-m}} p(b', b)$. It is well known that p is invariant if and only if there exists $\mu \in \mathcal{P}$ with $\mu_n = p$. For $n = 1, 2, \cdots$, let $\mathcal{P}_n$ be the set of all invariant probability measures on A^n. If $\omega \in A^n$ and $n > 2m$, define $\hat{p}(\omega, m)$ to be the measure on A^m obtained as follows: replace the last $m - 1$ elements of ω by the first $m - 1$ elements of ω, getting a sequence $\omega' \in A^n$. Define $\hat{p}(\omega, m) = p(\omega', m)$. It is easily checked that $\hat{p}(\omega, m) \in \mathcal{P}_m$. If q is a probability measure on A^m and $k < m$, let $[q]_k$ denote the probability measure on A^k such that $[q]_k(b) = \Sigma_{b' \in A^{m-k}} q(b, b')$. If p, q are probability measures on A^n, define the distance between them to be $|p - q| = \Sigma_{b \in A^n} |p(b) - q(b)|$.

We omit the easy proof of the following lemma.

LEMMA 1. (a) $|p(\omega, m) - \hat{p}(\omega, m)| \leqslant ma^m/(n - m + 1)$, $n > 2m$, $\omega \in A^n$.
(b) $|[p(\omega, m)]_k - p(\omega, k)| \leqslant 2a^k(m - k)/(n - k + 1)$, $n > m > k$, $\omega \in A^n$.

PROOF OF THEOREM 1. Let $\mathcal{P}^*$ be the set of all probability measures on $\mathcal{A}^\infty$, with the weak topology. For $n = 1, 2, \cdots$, let $\mathcal{P}_n^*$ be the set of all probability measures on A^n. We are given a continuous $F : \mathcal{P} \to R$. Let $F^* : \mathcal{P}^* \to R$ be any continuous extension of F. Let $C(\mathcal{P}^*)$ denote the vector space of all continuous real-valued functions defined on $\mathcal{P}^*$, with the supremum norm. We will call $G \in C(\mathcal{P}^*)$ finite-dimensional if for some n there exists a continuous $G^* : \mathcal{P}_n^* \to R$ such that $G(\mu) = G^*(\mu_n)$, $\mu \in 6\mathrm{p}^*$. By the Stone-Weierstrass theorem, the collection of all finite-dimensional functions in $C(\mathcal{P}^*)$ is uniformly dense in $C(\mathcal{P}^*)$. Hence, we may write $F^* = \Sigma_{i=1}^\infty 2^{-i} G_i$, where each G_i is finite dimensional and the G_i's are uniformly bounded by a number B. For each G_i, find an integer n_i and $G_i^* : \mathcal{P}_{n_i}^* \to R$ such that $G_i(\mu) = G_i^*(\mu_{n_i})$, $\mu \in \mathcal{P}^*$. For each i, define a sequence $\{f_i^{(n)}\}_{n=1}^\infty$ of functions from $A^\infty \to R$ as follows:

$$f_i^{(j)} = G_i^*\big(p(X^j, n_i)\big), \ |G_i^*\big(p(X^i, n_i)\big)| < B + 1, j \geqslant n_i;$$
$$= B, \text{ otherwise.}$$

Define the nth estimator Y_n so that $Y_n(X^n) = \Sigma_{i=1}^\infty 2^{-i} f_i^{(n)}$.

DEFINITION. If K is a convex subset of some vector space and $f : K \to R$, we say f is *ε-convex* if for any choice of finitely many $x_1, \cdots, x_j \in K$ and nonnegative numbers $\alpha_1, \cdots, \alpha_j$ adding to one, we have $f(\Sigma_{i=1}^j \alpha_i x_i) \leqslant \Sigma_{i=1}^j \alpha_i f(x_i) + \varepsilon$.

The following gives a generalization of Jensen's inequality for ε-convex functions. The proof is obtained by making the obvious modifications in a proof of Jensen's inequality for convex functions.

LEMMA 2. *Let K be a compact convex subset of R^n. Let $g \in C(K)$ be ε-convex. Let X be an n-dimensional random variable with* $\Pr[X \in K] = 1$. *Then* $E[g(X)] \geqslant g[E(X)] - \varepsilon$.

LEMMA 3. *Let $F : \mathcal{P} \to R$ be convex. Let $G : \mathcal{P} \to R$ satisfy* $\sup_{\mu \in \mathcal{P}} |G(\mu) - F(\mu)| \leqslant \varepsilon$. *Then G is 2ε-convex. Furthermore, if there exists $G^* : \mathcal{P}_n \to R$ such that $G(\mu) \equiv G^*(\mu_n)$, then G^* is 2ε-convex.*

PROOF. Easy.

LEMMA 4. *Let $G : \mathcal{P}_N \to R$ be ε-convex and continuous. Define $G^* : \mathcal{P} \to R$ so that $G^*(\mu) \equiv G(\mu_N)$. There exists $M > 2N$ such that for $n > M$ and all $\mu \in \mathcal{P}$, $E_\mu[G(\hat{p}(X^n, N))] \geqslant G^*(\mu) - 2\varepsilon$.*

PROOF. By Lemma 1, $|p(X^n, N) - \hat{p}(X^n, N)| \leqslant Na^N/(n - N + 1)$, so for all $\mu \in \mathcal{P}$, $|E_\mu[p(X^n, N)] - E_\mu[\hat{p}(X^n, N)]| \leqslant Na^N/(n - N + 1)$. By Lemma 2, $E_\mu[G(\hat{p}(X^n, N))] \geqslant G(E_\mu[\hat{p}(X^n, N)]) - \varepsilon$. Now for $\mu \in \mathcal{P}$, $E_\mu[p(X^n, N)] = \mu_N$. The result follows using the uniform continuity of G.

LEMMA 5. *Let $n_2 > n_1$. Let $G_i : \mathcal{P}_{n_i} \to R$ be continuous, $i = 1, 2$. For $i = 1, 2$, define $G_i^* : \mathcal{P} \to R$ so that $G_i^*(\mu) \equiv G_i(\mu_{n_i})$. Suppose $G_2^* \leqslant G_1^*$. Given $\varepsilon > 0$, there exists $N > 2n_2$ such that for $n > N$, $G_2[\hat{p}(X^n, n_2)] \leqslant G_1[\hat{p}(X^n, n_1)] + \varepsilon$.*

PROOF. $G_2^* \leqslant G_1^*$ implies that $G_2(\hat{p}(X^n, n_2)) \leqslant G_1([\hat{p}(X^n, n_2)]_{n_1})$. We have $|\hat{p}(X^n, n_1) - [\hat{p}(X^n, n_2)]_{n_1}| \leqslant |\hat{p}(X^n, n_1) - p(X^n, n_1)| + |p(X^n, n_1) - [p(X^n, n_2)]_{n_1}| + |[p(X^n, n_2)]_{n_1} - [\hat{p}(X^n, n_2)]_{n_1}| \leqslant n_1 a^{n_1}/(n - n_1 + 1) + 2a^{n_1}(n_2 - n_1)/(n - n_1 + 1) + n_2 a^{n_2}/(n - n_2 + 1)$, by Lemma 1. Now apply the uniform continuity of G_1.

LEMMA 6. *Let E be a separable metrizable locally convex space. Let K be a compact convex subset of E. Let $f : K \to R$ be upper semicontinuous and convex. Then f is the pointwise limit of a decreasing sequence of continuous convex functions from $K \to R$. If in addition f is affine, then f is the pointwise limit of a decreasing sequence of continuous affine functions from $K \to R$.*

PROOF. Meyer (1966), pages 222–223, proves this result for decreasing nets instead of sequences, in a general locally convex space. If one assumes separability and metrizability, an easy modification of Meyer's arguments allows one to replace nets by sequences.

PROOF OF THEOREM 2. Let $F : \mathcal{P} \to R$ be a given upper semicontinuous function, either convex or affine. By Lemma 6, if F is convex (affine), F is the pointwise limit of a decreasing sequence of continuous convex (continuous affine) functions. Call $G \in C(\mathcal{P})$ finite-dimensional if there exists N and a continuous $G^* : \mathcal{P}_N \to R$ such that $G(\mu) \equiv G^*(\mu_N)$. Using the Stone-Weierstrass theorem, it follows that F is a pointwise limit of a decreasing sequence of finite-dimensional functions from $C(\mathcal{P})$; by Lemma 3, if F is convex (affine), each term of the sequence can be chosen to be as nearly convex (affine) as desired. Using Lemmas 4 and 5, we thus

may find increasing sequences $\{N_i\}_{i=1}^{\infty}$, $\{M_i\}_{i=1}^{\infty}$ of positive integers, and sequences of functions $\{G_i\}_{i=1}^{\infty}$, $\{G_i^*\}_{i=1}^{\infty}$ such that:

(a) For each i, $G_i^* \in C(\mathcal{P})$, $G_i \in C(\mathcal{P}_{N_i})$ and $G_i^*(\mu) \equiv G_i(\mu_{N_i})$.

(b) $G_i^* \downarrow F$.

(c) $M_i > 2N_i$.

(d) If F is convex, then for all $\mu \in \mathcal{P}$, $E_\mu G_i(\hat{p}(X^n, N_i)) \geqslant F(\mu)$, $n \geqslant M_i$, $i = 1, 2, \cdots$.

(e) If F is affine, then for all $\mu \in \mathcal{P}$, $G_{i+2}^*(\mu) \leqslant E_\mu G_{i+1}(\hat{p}(X^n, N_{i+1})) \leqslant G_i^*(\mu)$, $n \geqslant M_{i+1}$, $i = 1, 2, \cdots$.

(f) $G_{i+1}(\hat{p}(X^n, N_{i+1})) \leqslant G_i(\hat{p}(X^n, N_i))$, $n \geqslant M_{i+1}$, $i = 1, 2, \cdots$.

Define a sequence $\{Z_i\}_{i=1}^{\infty}$ of functions on A^∞ so that $Z_i = G_{2i}(\hat{p}(X^{M_{2i}}, N_{2i}))$, $i = 1, 2, \cdots$. If F is convex, $E_\mu Z_i \geqslant F(\mu)$ for all $\mu \in \mathcal{P}$. If F is affine, $E_\mu Z_i \downarrow F(\mu)$ for all $\mu \in \mathcal{P}$. For fixed j, $Z_i \leqslant G_j(\hat{p}(X^{M_{2i}}, N_j))$ for i sufficiently large. Thus if $\mu \in \mathcal{P}_e$, $\limsup_{i\to\infty} Z_i \leqslant G_j^*(\mu)$ a.s. $[\mu]$, $j = 1, 2, \cdots$. Letting $j \to \infty$, we get $\limsup_{i\to\infty} Z_i \leqslant F(\mu)$ a.s. $[\mu]$. By Fatou's lemma, $E_\mu[\limsup_{i\to\infty} Z_i] \geqslant \limsup_{i\to\infty} E_\mu Z_i \geqslant \liminf_{i\to\infty} E_\mu Z_i \geqslant F(\mu)$. We conclude two things from this: $\limsup_{i\to\infty} Z_i = F(\mu)$ a.s. $[\mu]$, and $\lim_{i\to\infty} E_\mu[Z_i - F(\mu)] = 0$. To show $\lim_{i\to\infty} E_\mu[|Z_i - F(\mu)|] = 0$, it thus suffices to show $\lim_{i\to\infty} E_\mu[(Z_i - F(\mu))^+] = 0$. Now, $E_\mu[(Z_i - F(\mu))^+] \leqslant E_\mu[(Z_i - G_j(\hat{p}(X^{M_{2i}}, N_j)))^+] + E_\mu[(G_j(\hat{p}(X^{M_{2i}}, N_j)) - F(\mu))^+]$, for each fixed j. Letting $i \to \infty$ and then $j \to \infty$, we get $\lim_{i\to\infty} E_\mu[(Z_i - F(\mu))^+] = 0$. If B is an upper bound for G_1^*, define the estimators $\{Y_n\}$ so that $Y_n(X^n) = B$ for $n < M_2$ and $Y_n(X^n) = Z_i$ for $M_{2i} \leqslant n < M_{2i+2}$.

REFERENCES

[1] BAILEY, D. H. (1976). Sequential schemes for classifying and predicting ergodic processes. Ph.D. dissertation, Dept. of Mathematics, Stanford Univ.

[2] GRAY, R. M. and DAVISSON, L. D. (1974). Source coding theorems without the ergodic assumption. *IEEE Trans. Information Theory* **20** 502–516.

[3] JACOBS, K. (1962). Über die Struktur der mittleren Entropie. *Math. Z.* **78** 33–43.

[4] KIEFFER, J. C. (1975). On the optimum average distortion attainable by fixed-rate coding of a nonergodic source. *IEEE Trans. Information Theory* **21** 190–193.

[5] KIEFFER, J. C. (1978). A unified approach to weak universal source coding. *IEEE Trans. Information Theory* **24** 674–682.

[6] MEYER, P. A. (1966). *Probability and Potentials*. Blaisdell, Waltham, Mass.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MISSOURI
ROLLA, MISSOURI 65401