A New Approach to Filtering and Adaptive Control: Stability Results

Leigh Tesfatsion

Department of Economics
University of Southern California
Los Angeles, California 90007

Transmitted by R. Kalaba

ABSTRACT

A new approach to adaptive control is proposed. The principal distinguishing feature is the direct estimation and updating of the criterion function by means of a filtering operation on a vector of transitional pseudo-return functions. The data storage and computational problems often associated with explicit probability distribution updating via Bayes' rule are thus avoided. Convergence properties are established for a simple linear criterion function filter designed for a class of adaptive control problems typified by a well-known two-armed bandit problem. Optimality properties are established for the filter in a companion paper.

1. INTRODUCTION

Currently available adaptive-control methods focus on state-distribution estimation as an intermediate step towards optimal control-variable selection. As is well known, the need to evaluate and store updated state probabilities following each new datum often renders these methods infeasible for even moderately complex problems.

On the other hand, for certain types of adaptive-control problems the estimation of state distributions has no value in and of itself. Ultimate interest focuses on the criterion function, which is indirectly estimated by means of the state distribution estimate. By shifting attention from the state as random process to the criterion function, as a random function of the state, it becomes intuitively clear that one ought to be able to estimate and update the criterion function directly, without recourse to prior state-space specification, updated state probabilities, and Bayes's rule.

In Sec. 2 below, a new approach to adaptive control is proposed, based on these observations. The principal distinguishing feature is the direct con-

sistent estimation and updating of the criterion function by means of a filtering operation on a vector of return functions associated with previous state realizations. Potentially significant reductions in information and computation requirements are thus obtained.

In Sec. 3 the stability properties of a simple linear criterion-function filter are established for a class of adaptive-control problems with control-dependent states. Briefly, it is shown that control-variable sequences selected in accordance with the criterion-function estimates converge to a local maximum of the "true" criterion function under plausible restrictions. Results are illustrated in terms of a two-armed bandit problem. Sufficient conditions for convergence of control variable sequences to a global maximum of the true criterion function are established in [5]. Preliminary versions of the results established in the present paper and its companion paper [5] are discussed in [6].

The work of R. E. Kalaba and D. M. Detchmendy [3] is similar in spirit to the state-bypass approach developed here. An interesting procedure is proposed for converting output observations on a process directly into control signals via a single performance index which, at each time instant, weights the deviations from the observational history along with a measure of the future cost of control.

2. DIRECT ESTIMATION AND UPDATING OF CRITERION FUNCTIONS FOR ADAPTIVE CONTROL

A general method for selecting policies in accordance with directly updated criterion function estimates will be proposed below for the following class of adaptive-control problems:

PROBLEM 2.1. A decision maker must select policies (control variables) $\theta_1, \theta_2, \ldots$ at equally spaced times $t_1 < t_2 < \cdots$ from a policy choice set Θ. In each period $n \equiv [t_n, t_{n+1})$, $n \geqslant 1$, the decision maker observes the state ω_n of his problem environment subsequent to his policy selection θ_n. The objective of the decision maker in period n is to maximize his realized real-valued return $W(\omega_n, \theta_n)$. However, from the decision maker's myopic viewpoint the unfolding state sequence (ω_n) is a possibly policy-dependent realization for a random process whose characteristics are only vaguely discernable in period 1.

EXAMPLE 2.2 (Two-armed bandit: cf. [1]). In each period n a gambler places a bet b_n on one of two gambling devices. If the chosen bet–gambling-device policy is $\theta_n \equiv (b_n, \alpha)$, then the gambler receives b_n with unknown

probability p_a and $-b_n$ with probability $1-p_a$. Letting ω denote a generic element of the state set $\Omega^G \equiv \{1:(\text{gambler wins}), -1:(\text{gambler loses})\}$, the returns $W^G(\omega_n, \theta_n)$ of the gambler in period n are $\omega_n b_n$.

The general policy-selection method proposed for Problem 2.1 is as follows:

PROCEDURE 2.3.

Period 1.

Specify a prior criterion function $W_1^\wedge : \Theta \to R$ for first-period preobservation evaluation of the policies $\theta \in \Theta$.
Control: Select $\theta_1^* \in \Theta$ to maximize $W_1^\wedge$.
Observe the first-period state ω_1^*.

Period n $(n \geqslant 2)$.

Filter: For each policy $\theta \in \Theta$, estimate nth-period expected returns by

$$W_n^\wedge(\theta) = H(0,\theta,n) W_1^\wedge(\theta) + \sum_{j=1}^{n-1} H(j,\theta,n) W(\omega_j^*, \theta).$$

Control: Select $\theta_n^* \in \Theta$ to maximize $W_n^\wedge$.
Observe the nth-period state ω_n^*.

Thus for each policy $\theta \in \Theta$ and each period $n \geqslant 1$ the updated criterion function "output" $W_n^\wedge(\theta)$ is the end result of a linear filtering operation on the "input returns process" $\left(W_1^\wedge(\theta), W(\omega_1^*, \theta), \ldots, W(\omega_{n-1}^*, \theta)\right)$. Presumably the filter weights $H(j,\theta,n)$ could be selected to meet various optimality criteria important for the problem at hand. For some problems it might even be advantageous to use a nonlinear filtering operation. Discrete time periods are of course not essential.

ILLUSTRATION 2.4 (One-armed bandit). Consider the gambler in Example 2.2. Suppose there is only one distinct gambling device, with winning probability p. The policy choice set for the gambler in each period n is then

$$\Theta^G = \{b, \ldots\},$$

the range of his possible bets; and the true expected return associated with

each possible bet b is

$$W_\infty^G(b) \equiv pW^G(1,b) + [1-p]W^G(-1,p)$$
$$= pb + [1-p][-b].$$

Let $W_1^{\wedge G}: \Theta^G \to R$ denote a prior criterion function specified by the gambler for first period preobservation evaluation of the bets $b \in \Theta^G$. For the problem at hand it is reasonable to assume that the gambler knows his state set $\Omega^G \equiv \{1, -1\}$; thus one plausible specification for $W_1^{\wedge G}$ might be the barycentric prior $W_1^{\wedge G}(b) \equiv p^{\wedge}b + [1-p^{\wedge}][-b]$, where $p^{\wedge}$ is a prior estimate for p. In general, however, a decision maker with only a vague conception of possible states might be unable or unwilling to specify $W_1^{\wedge}$ in the form of a barycentric weighting of possible returns.

For each $n \geqslant 1$, $b \in \Theta^G$, and $j \in \{0, \ldots, n-1\}$, let the filter weight be $H^G(j,b,n) \equiv 1/n$. The policy selection Procedure 2.3 proposed for the one-armed bandit problem is then as follows:

PROCEDURE 2.5.

Period n $(n \geqslant 1)$.

Filter: For each bet $b \in \Theta^G$ estimate nth-period expected returns by

$$W_n^{\wedge G}(b) \equiv \begin{cases} W_1^{\wedge G}(b) & \text{if } n = 1, \\ \dfrac{1}{n}\left[W_1^{\wedge G}(b) + \displaystyle\sum_{j=1}^{n-1} W^G(\omega_j^*, b) \right] & \text{if } n > 1, \end{cases}$$

where ω_j^* is the state realized in period j, and $W^G(\omega_j^*, b) \equiv \omega_j^* b$ denotes the returns that would have accrued to the gambler for the realized state ω_j^* had he selected bet b in period j.
Control: Select a bet $b_n^* \in \Theta^G$ to maximize $W_n^{\wedge}$.
Observe the nth-period state ω_n^*.

The procedure (2.5) for updating the prior criterion function $W_1^{\wedge G}$ is formally analogous to Bayes's rule for updating prior probabilities. For uniform (constant) criterion priors $W_1^{\wedge G}$, selection of bets in accordance with Procedure 2.5 reduces to the following simple opportunity-cost maxim: in each period n select a bet b_n^* which would have yielded maximum average returns for the realized win-loss states $\omega_1^*, \ldots, \omega_{n-1}^*$. (The analogy with the maximum-likelihood principle is obvious.)

As will be seen below (Theorems 3.7 and 3.8 with $\text{card} A^* = 1$), the criterion-function estimates $W_n^{\wedge G}$ are consistent, and the gambler's bets b_n^* converge almost surely to a global maximum for the true criterion function W_∞^G if Θ^G is compact, $W_1^{\wedge G}$ is continuous, and successive win-loss observations are independent.

The information requirements of the policy-selection Procedure 2.3 are minimal. In contrast, Bayesian adaptive control methods generally require the motion of the state to be expressible as a system of difference or differential equations driven by random disturbances, with all uncertainty concerning functional forms and disturbance covariances reducible to uncertainty about one or more parameters lying in known parameter spaces.

It is also obvious that the Procedure 2.3 is computationally tractable. The filter weights $H(j,\theta,n)$ can be specified on line. In addition, for certain filter weights (e.g., those used in Sec. 3) the updated criterion function $W_n^\wedge$ has the recursive form

$$W_n^\wedge(\theta) = a(n,\theta) W_{n-1}^\wedge(\theta) + b(n,\theta) W(\omega_{n-1}^*, \theta).$$

The computational feasibility of Bayesian methods is limited by the need to evaluate and store updated probability assessments following each new observation.

Clearly, however, the increased tractability of Procedure 2.3 is at the cost of myopia with regard to future possibilities. The exact nature of this trade-off has not yet been fully clarified.

3. STABILITY

Stability properties will now be established for a simple linear criterion-function filter designed to handle dependence between state distribution and current policy choice. Sufficient conditions will be given for both asymptotic and finite-time convergence of policy selections.

The following standard definitions will be used.

DEFINITION 3.1. Let μ be a finitely additive set function defined on a field $\mathcal{F}$ of subsets of a set X. Then for every $F \in \mathcal{F}$ the *total variation* $v(\mu, F)$ *of* μ *on* F is defined as

$$v(\mu, F) \equiv \sup \sum_{i=1}^{n} |\mu(F_i)|,$$

where the supremum is taken over all finite sequences (F_i) of disjoint sets in

$\mathcal{F}$ with $F_i \subseteq F$. If μ is bounded, then the *positive* and *negative variations* μ^+ and μ^- of μ are defined on $\mathcal{F}$ by

$$\mu^+(F) \equiv \frac{v(\mu, F) + \mu(F)}{2},$$

$$\mu^-(F) \equiv \frac{v(\mu, F) - \mu(F)}{2}.$$

Both μ^+ and μ^- are bounded finitely additive set functions on $\mathcal{F}$. (See Dunford and Schwartz [2, Chapter III].)

Consider the adaptive-control Problem 2.1. If the distribution of states depends on current policy choice, then clearly the policy-selection Procedure 2.3 with simple filter weights $H(j, \theta, n) \equiv 1/n$ (cf. Illustration 2.4) is inappropriate. Intuitively, the knowledge that a state ω_j^* obtained following the choice of policy θ_j^* in period j cannot be directly used to update the evaluations of a policy θ' unless θ_j^* is exchangeable with θ' in the sense that

$$\text{Prob}(\omega = \omega_j^* | \theta_j^*) = \text{Prob}(\omega = \omega_j^* | \theta').$$

The policy choice set Θ must somehow be partitioned into sets of policies which are mutually exchangeable; and for each policy θ' currently being evaluated, the impulse response function H should give positive weight only to terms $W(\omega_j^*, \theta')$ corresponding to periods j in which the chosen policies θ_j^* are exchangeable with θ'. These statements will now be formalized.

Assume the set Ω of possible states ω in each period n is a topological space, and let $\mathcal{S}$ denote the σ-algebra of all Borel sets $S \subseteq \Omega$. Let Ω_∞ denote the set of all infinite state sequences $\omega_\infty = (\omega_1, \omega_2, \ldots)$, $\omega_j \in \Omega$, $j \geq 1$, and let $\mathcal{S}_\infty$ denote the σ-algebra of Ω_∞ generated by all cylinder sets of the form $\{\omega_\infty \in \Omega_\infty | \omega_{i_1} \in S_1, \ldots, \omega_{i_n} \in S_n\}$, where $S_j \in \mathcal{S}$, $j \in \{1, \ldots, n\}$, $n \geq 1$. Let Θ_∞ denote the set of all infinite policy sequences $\theta_\infty \equiv (\theta_1, \theta_2, \ldots)$, $\theta_j \in \Theta$, $j \geq 1$. Finally, let $P(\cdot | \theta_\infty)$, $\theta_\infty \in \Theta_\infty$, be a family of parametrized probability measures on $(\Omega_\infty, \mathcal{S}_\infty)$ with the following interpretation. If the decision maker in Problem 2.1 selects the policy sequence $\theta_\infty \equiv (\theta_n)_{n \geq 1} \in \Theta_\infty$, then $P(\cdot | \theta_\infty)$ is the probability distribution of the state sequence $\omega_\infty \equiv (\omega_n)_{n \geq 1}$.

If a partition $D(A) \equiv \{\Theta_\alpha | \alpha \in A\}$ of the policy choice set Θ satisfies either of the stable frequency conditions (C) or (CV) below, then the policies in each partition set Θ_α, $\alpha \in A$, are mutually exchangeable in a weak-frequency-limit sense.

(C) There exist probability measures $\sigma(\cdot|\alpha): \mathcal{S} \to [0,1]$, $\alpha \in A$, with the following property. If $\theta'_\infty \equiv (\theta'_n)_{n\geq 1} \in \Theta_\infty$ is selected by the decision maker in Problem 2.1, then for $P(\cdot|\theta'_\infty)$-a.e. state sequence $\omega'_\infty \equiv (\omega'_n)_{n\geq 1} \in \Omega_\infty$,

$$\lim_n \frac{\sum_{j \in I_n(\alpha)} 1_S^j(\omega'_\infty)}{\operatorname{card} I_n(\alpha)} = \sigma(S|\alpha), \qquad S \in \mathcal{S},$$

for all $\alpha \in A$ such that $\lim_n \operatorname{card} I_n(\alpha) = \infty$, where

$$1_S^j(\omega'_\infty) \equiv \begin{cases} 1 & \text{if} \quad \omega'_j \in S \\ 0 & \text{if} \quad \omega'_j \notin S; \end{cases}$$

and for each $\alpha \in A$ and $n > 1$,

$I_1(\alpha) \equiv \varnothing$,
$I_n(\alpha) \equiv$ the possibly empty index set comprising all periods $j \in \{1, \ldots, n-1\}$ for which $\theta'_j \in \Theta_\alpha$,
$\operatorname{card} I_n(\alpha) \equiv$ the cardinality of $I_n(\alpha)$.

(CV) There exist probability measures $\sigma(\cdot|\alpha): \mathcal{S} \to [0,1]$, $\alpha \in A$, with the following property. If $\theta'_\infty \in \Theta_\infty$ is selected by the decision maker in Problem 2.1, then for $P(\cdot|\theta'_\infty)$-a.e. state sequence $\omega'_\infty \in \Omega_\infty$,

$$\lim_n v\left(\frac{\sum_{j \in I_n(\alpha)} 1_{(\cdot)}^j(\omega'_\infty)}{\operatorname{card} I_n(\alpha)} - \sigma(\cdot|\alpha), S \right) = 0, \qquad S \in \mathcal{S},$$

for all $\alpha \in A$ such that $\lim_n \operatorname{card} I_n(\alpha) = \infty$, where the random variables 1_S^j and index sets $I_n(\alpha)$ are as defined in (C).

Intuitively, for each $\theta_\infty \in \Theta_\infty$ and $\alpha \in A$, condition (C) requires the state variables $(\omega_j | j \in \cup_{n\geq 1} I_n(\alpha))$ to resemble independent drawings from a fixed distribution $\sigma(\cdot|\alpha)$ on $(\Omega, \mathcal{S})$. Condition (CV) strengthens (C) by further restricting the allowable variation in the "samples" $(\omega_j | j \in \cup_{n\geq 1} I_n(\alpha))$, $\alpha \in A$. Sufficient conditions for (C) and (CV) to hold for a partition $D(A)$ of Θ are as follows:

THEOREM 3.2. *Let $D(A) = \{\Theta_\alpha | \alpha \in A\}$ be a partition of the policy choice set Θ.*

(i) *Suppose* $\mathbb{S}$ *and* A *are countable, and for each* $S \in \mathbb{S}$ *and* $\theta'_\infty = (\theta'_n)_{n \geq 1} \in \Theta_\infty$ *the random variables* 1^j_S, $j \geq 1$, *are independent on* $(\Omega_\infty, \mathbb{S}_\infty, P(\cdot|\theta'_\infty))$, *with expectation*

$$\mathrm{E}\, 1^j_S \equiv \sigma(S|\alpha) \qquad \text{if} \quad \theta'_j \in \Theta_\alpha, \quad \alpha \in A.$$

Then condition (C) *holds for* $D(A)$.

(ii) *If condition* (C) *holds for* $D(A)$ *and* $\mathbb{S}$ *is finite, then condition* (CV) *holds for* $D(A)$.

PROOF. Assertion (i) follows immediately from the well-known strong law for independent and uniformly bounded random variables with identical finite expectation (see Révész [4, Theorem 2.7.1, p. 59]), the assumed countable additivity of $P(\cdot|\theta_\infty)$, $\theta_\infty \in \Theta_\infty$, and the nullity of a countable union of null sets.

Assertion (ii) follows immediately from a lemma by Dunford and Schwartz [2, Lemma 5, p. 97], which asserts that for any real (or complex) bounded finitely additive set function μ defined on a field $\mathcal{F}$ of subsets of a set X,

$$v(\mu, X) \leq 4 \sup_{F \in \mathcal{F}} |\mu(F)|. \qquad \blacksquare$$

ILLUSTRATION 3.3 (Two-armed bandit). Consider the two-armed bandit Example 2.2 with winning probabilities P_α and P_β for two distinct gambling devices α and β. In each period n the gambler must choose both a bet and a gambling device. His policy choice set is thus of the form

$$\Theta^G \equiv \Theta_\alpha \cup \Theta_\beta,$$

where

$$\Theta_\alpha \equiv B_\alpha \times \{\alpha\}, \qquad \Theta_\beta \equiv B_\beta \times \{\beta\},$$

and B_α and B_β are ranges of possible bets. Let $\mathbb{S}^G$ denote the power set of the gambler's win-loss state set $\Omega^G \equiv \{1, -1\}$. If for any sequence of plays on the gambling devices α and β the gambler's win-loss observations are independent, then condition (CV) holds for the gambler's problem with $A \equiv \{\alpha, \beta\}$, $D(A) \equiv \{\Theta_\alpha, \Theta_\beta\}$, $\Omega \equiv \Omega^G$, $\mathbb{S} = \mathbb{S}^G$, $\sigma(1|\alpha) \equiv P_\alpha$, and $\sigma(1|\beta) \equiv P_\beta$.

As will be seen below, the following filter-control procedure proposed for Problem 2.1 has reasonable stability properties if condition (C) (finite Θ) or (CV) (infinite Θ) holds for the partition $D(A^*)$ specified by the decision maker in period 1.

PROCEDURE 3.4.

Period 1.

Specify a prior criterion function $W_1^\wedge : \Theta \to R$ for first-period preobservation evaluation of the policies $\theta \in \Theta$.
Control: Select $\theta_1^* \in \Theta$ to maximize $W_1^\wedge$.
Observe the first period state ω_1^*.
Specify a partitioning $D(A^*) = \{\Theta_\alpha | \alpha \in A^*\}$ of the policy choice set Θ into sets Θ_α of policies judged to be mutually exchangeable with respect to their effect on state distribution.

Period n ($n \geqslant 2$).

Filter: For each policy $\theta \in \Theta$ estimate nth-period expected returns by

$$W_n^\wedge(\theta) = \frac{W_1^\wedge(\theta) + \sum_{j \in I_n(\alpha)} W(\omega_j^*, \theta)}{1 + \operatorname{card} I_n(\alpha)}, \qquad \theta \in \Theta_\alpha, \quad \alpha \in A^*,$$

where for each $\alpha \in A^*$

$I_n(\alpha) \equiv$ the possibly empty index set comprising all periods $j \in \{1, \ldots, n-1\}$ for which $\theta_j^* \in \Theta_\alpha$;
$\operatorname{card} I_n(\alpha) \equiv$ the cardinality of $I_n(\alpha)$.

Control: Select $\theta_n^* \in \Theta$ to maximize $W_n^\wedge$.
Observe the nth-period state ω_n^*.

All subsequent theorems, definitions, and remarks will concern the components $(\Theta, W_1^\wedge, D(A^*), W, W_n^\wedge)$ for the filter-control Procedure 3.4.

REMARK. If in period 1 the decision maker is able to specify policy-conditioned proabability priors $\sigma_0(\cdot|\theta)$, $\theta \in \Theta$, on the state space $(\Omega, \mathcal{S})$, reflecting his judgment concerning the likelihood of events $S \in \mathcal{S}$, then a plausible specification for $D(A^*)$ would be the collection of all $\sim$-equivalence classes $\Theta_\alpha \subseteq \Theta$, where

$$\theta' \sim \theta'' \quad \Leftrightarrow \quad \sigma_0(\cdot|\theta') = \sigma_0(\cdot|\theta'').$$

Suppose condition (C) holds for the partition $D(A^*)$. If the return function sections $W(\cdot,\theta)$, $\theta\in\Theta$, are bounded and continuous over the state space $(\Omega,\mathcal{S})$, then the map $W_\infty:\Theta\to R$ given by

$$W_\infty(\theta)\equiv\int_\Omega W(\omega,\theta)\sigma(d\omega|\alpha),\qquad \theta\in\Theta_\alpha,\quad \alpha\in A^*,$$

is well defined and may appropriately be referred to as the *true criterion function* for Procedure 3.4. As the following theorem demonstrates, under the above assumptions the updated criterion function $W_n^\wedge$ is a consistent estimator for W_∞ over partition sets $\Theta_\alpha\in D(A^*)$ selected infinitely often by the decision maker.

LEMMA 3.5 (Dunford and Schwartz [2, Theorem 15, p. 316]). *Let X be a topological space, and let $\mathcal{F}$ be a field of subsets $F\subseteq X$ containing the open sets of X. Let μ, μ_n, $n=1,2,\ldots$ be a bounded sequence of finitely additive measures defined on $\mathcal{F}$. If*

$$\lim_n \mu_n(F)=\mu(F)$$

for every open set $F\in\mathcal{F}$ satisfying $\mu(F)=\mu(\text{closure}(F))$, then

$$\lim_n \int_X f(x)\,\mu_n(dx)=\int_X f(x)\,\mu(dx)$$

for every real bounded continuous function f on X.

THEOREM 3.6. *Assume condition* (C) *holds for $D(A^*)$, and the return function sections $W(\cdot,\theta)$, $\theta\in\Theta$, are bounded and continuous over $(\Omega,\mathcal{S})$. If $\theta^*_\infty\in\Theta_\infty$ is selected in accordance with Procedure* 3.4, *then $P(\cdot|\theta^*_\infty)$*-a.s.

$$\lim_n \text{card}\, I_n(\alpha)=\infty \quad \text{for some } \alpha\in A^*$$

$$\Rightarrow \quad \lim_n W_n^\wedge(\theta)=W_\infty(\theta),\quad \theta\in\Theta_\alpha.$$

PROOF. The proof follows immediately from condition (C), the definition of $W_n^\wedge$, and Lemma 3.5. ■

By replacing condition (C) with condition (CV) and strengthening remaining assumptions, uniform convergence of the estimator $W_n^{\wedge}$ can be guaranteed.

THEOREM 3.7. *Assume*:

1. *Condition* (CV) *holds for* $D(A^*)=\{\Theta_\alpha|\alpha\in A^*\}$.
2. (Θ,τ) *is a topological space, and the partition sets* $\Theta_\alpha\in D(A^*)$ *are* τ*-compact.*
3. *The prior criterion function* $W_1^{\wedge}$ *and the return function sections* $W(\omega,\cdot)$, $\omega\in\Omega$, *are continuous over* (Θ,τ).
4. *For each* $\alpha\in A^*$ *the function*

$$W_\alpha(\cdot)\equiv\sup_{\theta\in\Theta_\alpha}|W(\cdot,\theta)|$$

is bounded and continuous over $(\Omega,\mathcal{S})$.

If $\theta_\infty^*\in\Theta_\infty$ *is selected in accordance with Procedure* 3.4, *then* $P(\cdot|\theta_\infty^*)$*-a.s.*

$$\lim_n \operatorname{card} I_n(\alpha)=\infty \quad \text{for some } \alpha\in A^*$$

$$\Rightarrow \quad \lim_n \sup_{\theta\in\Theta_\alpha}|W_n^{\wedge}(\theta)-W_\infty(\theta)|=0.$$

PROOF. Let $\theta_\infty^*\in\Theta_\infty$ be selected in accordance with Procedure 3.4. Suppose $\lim_n \operatorname{card} I_n(\alpha)=\infty$ for some $\alpha\in A^*$. For each $n\geqslant 2$ and $S\in\mathcal{S}$ define

$$\sigma_n(S|\alpha)\equiv\sum_{j\in I_n(\alpha)}\frac{1_S(\omega_j^*)}{1+\operatorname{card} I_n(\alpha)},$$

$$\mu_n(S)\equiv\sigma_n(S|\alpha)-\sigma(S|\alpha),$$

where ω_j^* denotes the state observed in period j, 1_S denotes the indicator function for S, and $\sigma(\cdot|\alpha)$ denotes the frequency-limit measure whose existence is guaranteed by (CV). Letting μ_n^+, μ_n^-, and $v(\mu_n,\cdot)$ be as defined in Definition 3.1, it follows from condition (CV) that $P(\cdot|\theta_\infty^*)$-a.s.

$$\lim_n v(\mu_n,S)=\lim_n \mu_n(S)=\lim_n \mu_n^+(S)=\lim_n \mu_n^-(S)=0$$

for every $S \in \mathcal{S}$. Thus by condition 4 and Lemma 3.5,

$$\lim_n \int_\Omega W_\alpha(\omega)\,\mu_n^+(d\omega) = \lim_n \int_\Omega W_\alpha(\omega)\,\mu_n^-(d\omega) = 0 \qquad P(\cdot\,|\theta_\infty^*)\text{-a.s.}$$

By conditions 2 and 3, $W_1^\wedge$ is bounded over Θ_α; and by condition 4 the functions $W(\cdot,\theta)$ and $|W(\cdot,\theta)|$, $\theta \in \Theta$, are bounded and continuous over $(\Omega, \mathcal{S})$, and hence integrable with respect to μ_n^+, μ_n^-, and μ_n for each $n \geqslant 1$. Thus

$$0 \leqslant \lim_n \sup_{\theta \in \Theta_\alpha} |W_n^\wedge(\theta) - W_\infty(\theta)|$$

$$= \lim_n \sup_{\theta \in \Theta_\alpha} \left| \frac{W_1^\wedge(\theta)}{1 + \operatorname{card} I_n(\alpha)} + \int_\Omega W(\omega,\theta)\,\mu_n(d\omega) \right|$$

$$\leqslant \lim_n \sup_{\theta \in \Theta_\alpha} \left[O\left(\frac{1}{\operatorname{card} I_n(\alpha)}\right) + \int_\Omega |W(\omega,\theta)|\,\mu_n^+(d\omega) + \int_\Omega |W(\omega,\theta)|\,\mu_n^-(d\omega) \right]$$

$$\leqslant \lim_n \left[O\left(\frac{1}{\operatorname{card} I_n(\alpha)}\right) + \int_\Omega W_\alpha(\omega)\,\mu_n^+(d\omega) + \int_\Omega W_\alpha(\omega)\,\mu_n^-(d\omega) \right]$$

$$= 0 \qquad P(\cdot\,|\theta_\infty^*)\text{-a.s.}$$

It is clear from condition (CV) that the required $P(\cdot\,|\theta_\infty^*)$-null set may be chosen independently of α. ■

If in addition to conditions 1–4 the policy choice set Θ is metrizable and A^* is finite, then policy sequences selected in accordance with Procedure 3.4 converge a.s. to the set of local maxima for the true criterion function W_∞. Formally

THEOREM 3.8. *Let conditions* 1–4 *in Theorem* 3.7 *hold. Suppose* (Θ, τ) *is metrizable with metric* d, *and the index set* A^* *is finite. If* $\theta_\infty^* \equiv (\theta_n^*)_{n \geqslant 1} \in \Theta_\infty$ *is selected in accordance with Procedure* 3.4, *then* $P(\cdot\,|\theta_\infty^*)$*-a.s.*

$$\lim_n d(\theta_n^*, M) = 0, \tag{3.1}$$

where

$$M \equiv \bigcup_{\alpha \in A^*} \mathcal{P}(\alpha),$$

and for each $\alpha \in A^*$

$$\mathcal{P}(\alpha) \equiv \{\theta \in \Theta_\alpha | \theta \text{ maximizes } W_\infty \text{ over } \Theta_\alpha\}.$$

PROOF. By conditions 2–4 and Lebesgue's dominated-convergence theorem, W_∞ is continuous over each compact partition set Θ_α, $\alpha \in A^*$. Thus M is nonempty.

Suppose $\theta'_\infty \equiv (\theta'_n)_{n \geqslant 1} \in \Theta_\infty$ is selected in accordance with Procedure 3.4, and suppose the observed state sequence $\omega'_\infty \notin N(\theta'_\infty)$, where $N(\theta'_\infty)$ is the $P(\cdot|\theta'_\infty)$-null set for which convergence in Theorem 3.7 fails to hold for θ'_∞. To prove Theorem 3.8 it suffices to prove that (3.1) holds for θ'_∞.

Suppose to the contrary that (3.1) does not hold for θ'_∞. Then there exists $\delta > 0$ such that given any integer m there is an integer $m^* \equiv N(m, \delta) \geqslant m$ satisfying

$$d(\theta'_{m^*}, M) \geqslant \delta. \tag{3.2}$$

Since by assumption $\Theta \equiv \cup_{\alpha \in A^*} \Theta_\alpha$ is compact metric, and hence sequentially compact, it may be assumed without loss of generality that

$$\lim_m d(\theta'_{m^*}, \theta') = 0 \tag{3.3}$$

for some $\theta' \in \Theta$. Thus by (3.2), (3.3), and the continuity of $d(\cdot, M)$ over Θ,

$$d(\theta', M) \geqslant \delta. \tag{3.4}$$

Let $\Theta_{\alpha'}$ denote the partition set of $D(A^*)$ which contains θ'. Since by assumption the partition sets Θ_α, $\alpha \in A^*$, are compact and disjoint, it follows by (3.3) that $\theta'_{m^*} \in \Theta_{\alpha'}$ for all sufficiently large m. Thus by definition of θ'_{m^*} (a $W^{\wedge}_{m^*}$-maximizing policy selection) and uniform convergence of $W^{\wedge}_n$ to W_∞ over $\Theta_{\alpha'}$ (Theorem 3.7),

$$\lim_m W^{\wedge}_{m^*}(\theta'_{m^*}) = \lim_m \left[\max_{\theta \in \Theta_{\alpha'}} W^{\wedge}_{m^*}(\theta) \right] = \max_{\theta \in \Theta_{\alpha'}} W_\infty(\theta), \tag{3.5}$$

$$\lim_m \left| W^{\wedge}_{m^*}(\theta'_{m^*}) - W_\infty(\theta'_{m^*}) \right| = 0. \tag{3.6}$$

By the continuity of W_∞, Eqs. (3.3), (3.5), and (3.6) imply

$$\max_{\theta \in \Theta_{\alpha'}} W_\infty(\theta) = \lim_m W_{m^*}^{\wedge}(\theta'_{m^*}) = \lim_m W_\infty(\theta'_{m^*}) = W_\infty(\theta'),$$

which in turn implies

$$\theta' \in \mathcal{P}(\alpha') \subseteq M. \tag{3.7}$$

Since (3.7) contradicts (3.4), the supposition that (3.1) does not hold for θ'_∞ cannot be maintained. ■

THEOREM 3.9. *Suppose in addition to the hypotheses of Theorem* 3.8 *the following condition holds*:

(R) *The values* $[\max_{\theta \in \Theta_\alpha} W_\infty(\theta)]$, $\alpha \in A^*$, *are distinct.*

If $\theta^*_\infty \equiv (\theta^*_n)_{n \geq 1} \in \Theta_\infty$ *is selected in accordance with Procedure* 3.4, *then* $P(\cdot|\theta^*_\infty)$*-a.s. there exists* $\alpha^* \in A^*$ *such that*

(i) $\theta^*_n \in \Theta_{\alpha^*}$ *for all sufficiently large* n;
(ii) $\lim_n d(\theta^*_n, \mathcal{P}(\alpha^*)) = 0$.

REMARK. In terms of the two-armed bandit example (Illustration 3.3) with $D(A^*) \equiv \{\Theta_\alpha, \Theta_\beta\}$, conclusion (i) asserts that one of the two gambling devices $\{\alpha, \beta\}$ will be selected all but finitely many times; and conclusion (ii) asserts that the gambler's bets will converge to the set of optimal bets for that gambling device.

PROOF. Suppose $\theta^0_\infty = (\theta^0_n)_{n \geq 1} \in \Theta_\infty$ is selected in accordance with Procedure 3.4, and the observed state sequence $\omega^0_\infty \notin N(\theta^0_\infty)$, where $N(\theta^0_\infty)$ is the $P(\cdot|\theta^0_\infty)$-null set for which convergence in Theorem 3.7 fails to hold for θ^0_∞. To prove Theorem 3.9 it suffices to prove that (i) and (ii) hold for θ^0_∞.

Suppose no partition set contains θ^0_n for all sufficiently large n. Since $\text{card} A < \infty$, there must then exist distinct partition sets $\Theta_{\alpha'}$ and $\Theta_{\alpha''}$ which each contain an infinite number of the policies θ^0_n. As established in Theorem 3.8, the functions W_∞ and $W_n^{\wedge}$, $n \geq 1$, are continuous over the compact sets $\Theta_{\alpha'}$ and $\Theta_{\alpha''}$ and thus attain finite maxima on these sets. It follows by

Theorem 3.7 that

$$\lim_n \max_{\theta \in \Theta_{\alpha'}} W_n^{\wedge}(\theta) = \max_{\theta \in \Theta_{\alpha'}} W_\infty(\theta), \tag{3.8}$$

$$\lim_n \max_{\theta \in \Theta_{\alpha''}} W_n^{\wedge}(\theta) = \max_{\theta \in \Theta_{\alpha''}} W_\infty(\theta). \tag{3.9}$$

By condition (R) it may be assumed without loss of generality that $\max_{\theta \in \Theta_{\alpha'}} W_\infty(\theta) > \max_{\theta \in \Theta_{\alpha''}} W_\infty(\theta)$. Then, by (3.8) and (3.9), for some $\varepsilon > 0$ and integer $\bar{n}$,

$$\max_{\theta \in \Theta_{\alpha'}} W_n^{\wedge}(\theta) \geqslant \max_{\theta \in \Theta_{\alpha''}} W_n^{\wedge}(\theta) + \varepsilon, \qquad n \geqslant \bar{n}. \tag{3.10}$$

On the other hand, by definition of (θ_n^0) and by supposition,

$$W_n^{\wedge}(\theta_n^0) = \max_{\theta \in \Theta} W_n^{\wedge}(\theta) \geqslant \max_{\theta \in \Theta_{\alpha'}} W_n^{\wedge}(\theta), \qquad n \geqslant 1, \tag{3.11}$$

$$W_n^{\wedge}(\theta_n^0) = \max_{\theta \in \Theta_{\alpha''}} W_n^{\wedge}(\theta) \qquad \text{for infinitely many } n. \tag{3.12}$$

Clearly (3.11) and (3.12) contradict (3.10).

Hence for some $\alpha^0 \in A^*$, we have $\theta_n^0 \in \Theta_{\alpha^0}$ for all sufficiently large n, so that assertion (i) holds for θ_∞^0. It now follows immediately from Theorem 3.8 with Θ_{α_0} in place of Θ, $\mathcal{P}(\alpha^0)$ in place of M, and $A^* \equiv \{\alpha^0\}$, that assertion (ii) also holds for θ_∞^0. ■

The hypotheses of Theorem 3.9 can be weakened and the conclusion strengthened if the policy choice set Θ is finite. Formally,

THEOREM 3.10. *Assume*

1. *Condition* (C) *holds for* $D(A^*) = \{\Theta_\alpha | \alpha \in A^*\}$.
2. *The return function sections* $W(\cdot, \theta)$, $\theta \in \Theta$, *are bounded and continuous over* $(\Omega, \mathcal{S})$.
3. Θ *is finite.*

If $\theta_\infty^* \equiv (\theta_n^*)_{n \geqslant 1} \in \Theta_\infty$ *is selected in accordance with Procedure* 3.4, *then* $P(\cdot | \theta_\infty^*)$*-a.s. there exists an integer* n' *such that*

$$\theta_n^* \in M \equiv \bigcup_{\alpha \in A^*} \mathcal{P}(\alpha), \qquad n \geqslant n'. \tag{3.13}$$

Suppose in addition the following condition holds:

4. *The values* $[\max_{\theta \in \Theta_\alpha} W_\infty(\theta)]$, $\alpha \in A^*$, *are distinct.*

Then for some $\alpha^* \in A^*$ *and integer* n'',

$$\theta_n^* \in \mathcal{P}(\alpha^*), \qquad n \geqslant n''. \tag{3.14}$$

PROOF. Suppose $\theta_\infty^0 \equiv (\theta_n^0)_{n \geqslant 1} \in \Theta_\infty$ is selected in accordance with Procedure 3.4, and suppose the observed state sequence $\omega_\infty^0 \notin N^\wedge(\theta_\infty^0)$, where $N^\wedge(\theta_\infty^0)$ is the $P(\cdot|\theta_\infty^0)$-null set for which convergence in Theorem 3.6 fails to hold for θ_∞^0. To prove Theorem 3.10 it suffices to prove (3.13) and (3.14) hold for θ_∞^0 under the stated conditions.

Assume conditions 1–3 hold. Since Θ is finite, M is nonempty. Suppose θ_n^0 does not lie in M for all sufficiently large n. Then there must exist a policy $\theta' \notin M$ such that $\theta_n^0 = \theta'$ i.o. Let $\Theta_{\alpha'}$ denote the partition set containing θ'. Then

$$\max_{\theta \in \Theta} W_n^\wedge(\theta) = W_n^\wedge(\theta_n^0) = W_n^\wedge(\theta') \quad \text{i.o.}$$

implies

$$\max_{\theta \in \Theta_{\alpha'}} W_n^\wedge(\theta) = W_n^\wedge(\theta') \quad \text{i.o.} \tag{3.15}$$

On the other hand, by Theorem 3.6 and finiteness of Θ,

$$\lim_n \left[\max_{\theta \in \Theta_{\alpha'}} W_n^\wedge(\theta) \right] = \max_{\theta \in \Theta_{\alpha'}} W_\infty(\theta), \tag{3.16}$$

$$\lim_n W_n^\wedge(\theta') = W_\infty(\theta'). \tag{3.17}$$

Clearly Eqs. (3.15), (3.16), and (3.17) imply

$$\max_{\theta \in \Theta_{\alpha'}} W_\infty(\theta) = W_\infty(\theta'),$$

i.e.,

$$\theta' \in \mathcal{P}(\alpha') \subseteq M,$$

a contradiction. Thus $\theta_n^0 \in M$ for all sufficiently large n, i.e., (3.13) holds for θ_∞^0.

Now assume condition 4 holds in addition to conditions 1–3. Since $\theta_n^0 \in M$ for all sufficiently large n, either there exists α^0 such that $\theta_n^0 \in \mathcal{P}(\alpha^0)$ for all sufficiently large n, or there exist $\alpha' \neq \alpha''$ such that $\theta_n^0 \in \mathcal{P}(\alpha')$ i.o. and $\theta_n^0 \in \mathcal{P}(\alpha'')$ i.o. Suppose the latter holds. Then by Theorem 3.6 and finiteness of Θ,

$$\lim_n \left[\max_{\theta \in \Theta_{\alpha'}} W_n^{\wedge}(\theta) \right] = \max_{\theta \in \Theta_{\alpha'}} W_\infty(\theta); \tag{3.18}$$

$$\lim_n \left[\max_{\theta \in \Theta_{\alpha''}} W_n^{\wedge}(\theta) \right] = \max_{\theta \in \Theta_{\alpha''}} W_\infty(\theta). \tag{3.19}$$

By condition 4 it may be assumed without loss of generality that $\max_{\theta \in \Theta_{\alpha'}} W_\infty(\theta) > \max_{\theta \in \Theta_{\alpha''}} W_\infty(\theta)$. Then by (3.18) and (3.19), there exist $\varepsilon > 0$ and an integer $\bar{n}$ such that

$$\max_{\theta \in \Theta_{\alpha'}} W_n^{\wedge}(\theta) \geqslant \max_{\theta \in \Theta_{\alpha''}} W_n^{\wedge}(\theta) + \varepsilon, \qquad n \geqslant \bar{n}. \tag{3.20}$$

On the other hand, by the definition of (θ_n^0) and by supposition,

$$W_n^{\wedge}(\theta_n^0) = \max_{\theta \in \Theta} W_n^{\wedge}(\theta) \geqslant \max_{\theta \in \Theta_{\alpha'}} W_n^{\wedge}(\theta), \qquad n \geqslant 1; \tag{3.21}$$

$$W_n^{\wedge}(\theta_n^0) = \max_{\theta \in \Theta_{\alpha''}} W_n^{\wedge}(\theta) \qquad \text{for infinitely many } n. \tag{3.22}$$

Since (3.21) and (3.22) contradict (3.20), there must exist $\alpha^0 \in A^*$ such that $\theta_n^0 \in \mathcal{P}(\alpha^0)$ for all sufficiently large n, i.e., (3.14) holds for θ_∞^0. ■

4. CONCLUSION

A simple distribution-free method has been proposed for direct estimation and updating of a criterion function without recourse to prior state-space specification, updated state probabilities, and Bayes's rule. Sufficient conditions have been given for the consistency of criterion-function estimates, and for both the asymptotic and finite-time convergence of selected policy (control-variable) sequences to a local maximum of the true criterion function, using a linear criterion-function filtering scheme designed to correct for state dependence on current policy choice. Sufficient conditions for convergence to a global maximum are established in [5].

REFERENCES

1 R. Bellman, A problem in the sequential design of experiments, *Sankhyā* **16** (1956), 221–229.
2 N. Dunford and J. T. Schwartz, *Linear Operators, Part I: General Theory*, Interscience, New York, 1957.
3 R. E. Kalaba and D. M. Detchmendy, Direct conversion of observational histories into control signals, *Inf. Sci.* 1 (1968), 1–5.
4 P. Révész, *The Laws of Large Numbers*, Academic, New York, 1967.
5 L. Tesfatsion, A new approach to filtering and adaptive control: optimality results, *J. Cybernetics*, to be published.
6 L. Tesfatsion, A new approach to filtering and adaptive control, *J. Optimization Theory Appl.*, to be published.