A CONTINUOUS KIEFER-WOLFOWITZ PROCEDURE FOR RANDOM PROCESSES[1]

By David J. Sakrison[2]

Massachusetts Institute of Technology

1. Introduction. The Kiefer-Wolfowitz procedure as previously described [1], [8], [10], was concerned with solving the following problem: given a random variable $Y = Y(x_1, x_2, \cdots, x_k)$ depending on k real valued parameters $x_1, x_2, \cdots, x_k$, determine the values of these parameters which minimize

$$M(x_1, x_2, \cdots, x_k) = E\{Y(x_1, x_2, \cdots, x_k)\}$$

by making a sequence of independent observations of the random variable Y at different values of the parameters. In the field of communications we are usually more interested in the case in which Y is an ergodic random process Y_t. Here we consider this situation and study a continuous version of the Kiefer-Wolfowitz procedure. The advantage of using a continuous version of the procedure when Y_t is a continuous time-parameter process (as opposed to periodically sampling Y_t and applying the original procedure to the samples) lies in the fact that it may be mechanized with simple analog computation components.

Our analysis considers a straightforward generalization of the original procedure and, to a certain extent, follows the pattern of Dupač's analysis [6] of the original procedure. The hypotheses of the theorems which we present are chosen from the standpoint of applicability to certain communication and data processing problems rather than from the standpoint of mathematical generality. Other continuous stochastic approximation methods have received treatment [4], [5], [7], but none seem appropriate for data processing applications. For the one-dimensional case Driml and Nedoma [5] consider a continuous version of a generalized Robbins-Munro procedure and obtain almost sure convergence under more liberal assumptions than are made here: unfortunately, their analysis cannot be extended to the multidimensional case. Neither of the procedures considered by Driml and Hanš [4] and Hanš and Špaček [7] seem particularly well suited for analog computation.

2. Notation and description of the approximation procedure. We regard the k parameters $x_1, x_2, \cdots x_k$ as the components of a k-dimensional vector $\mathbf{x}$. The basis for the space will be the unit vectors $\mathbf{e}_1, \mathbf{e}_2, \cdots \mathbf{e}_k$, $\mathbf{e}_i$ denoting a unit value of x_i and zero values for the other $k - 1$ parameters. We will denote the usual Euclidean norm and inner product by $\|\mathbf{x}\|$ and $(\mathbf{x}, \mathbf{y})$ respectively. We

Received 26 March 1963.

[1] This work was supported in part by the U.S. Army Signal Corps, the Air Force Office of Scientific Research and the Office of Naval Research, through the Research Laboratory of Electronics, Massachusetts Institute of Technology.

[2] Now at University of California, Berkeley.

denote the regression function by

$$M(\mathbf{x}) = E\{Y_t(\mathbf{x})\} \tag{1}$$

and denote by $\boldsymbol{\theta}$ the vector parameter value for which M is a minimum. Let

$$Y_{i,t}[\mathbf{x}, c(t)] = Y_t[\mathbf{x} + c(t)\mathbf{e}_i] - Y_t[\mathbf{x} - c(t)\mathbf{e}_i] \tag{2}$$

in which $c(t)$ is a positive function whose properties will be described later. The minimum-seeking approximation procedure is then defined by

$$(d/dt)X_{i,t} = -a(t)I_{i,t}c^{-1}(t)Y_{i,t}[X_t, c(t)] \tag{3}$$

and

$$X_{i,0} = x_i(0) \qquad i = 1, 2, \cdots, k$$

in which $x_i(0)$ is the initial value of the ith parameter and

$$I_{i,t} = 1 - G_i^+[X_{i,t}]\,F_i^+[Y_{i,t}] - G_i^-[X_{i,t}]\,F_i^-[Y_{i,t}] \tag{4}$$

$$\begin{aligned} G_i^+(x) &= 0, \quad x \leqq b_i - \delta & G_i^-(x) &= 0, \quad x \geqq a_i + \delta \\ &= 1, \quad x = b_i & &= 1, \quad x = a_i \\ &= \text{monotone and of bounded derivative on } [b_i - \delta, b_i] & &= \text{monotone and of bounded derivative on } [a_i, a_i + \delta] \end{aligned} \tag{5}$$

and

$$F_i^+(y) = 1 - y/\epsilon_y c(t), \qquad F_i^-(y) = 1 + y/\epsilon_y c(t) \tag{6}$$

in which ϵ_y is a positive constant; we will later place a suitable bound on ϵ_y. For convenience in the sequel we will denote by $\mathbf{Y}_t$ the vector whose ith component is $Y_{i,t}$ and by $\mathbf{Z}_t$ the vector whose ith component is $I_{i,t}Y_{i,t}$. We also define

$$\mathbf{M}_{c(t)}(\mathbf{x}) = E\{\mathbf{Y}_t(\mathbf{x})\}c^{-1}(t) \tag{7}$$

and

$$\mathbf{Q}_{c(t)}(\mathbf{x}) = E\{\mathbf{Z}_t(\mathbf{x})\}\,c^{-1}(t). \tag{8}$$

Note that these quantities are defined in terms of the parameter $\mathbf{x}$ and not the random variable $\mathbf{X}_t$.

The relation between this continuous procedure and the original Kiefer-Wolfowitz procedure should be apparent. The differential Equation (3) can simply be regarded as the limiting case of the difference equation governing the original procedure. The only real conceptual difference is that in the original procedure the different observations used sequentially were assumed to be conditionally independent; here such an assumption is not meaningful and must be replaced by an alternate condition. One other difference is the appearance of the term I_i in Equation (3): its purpose is merely to constrain the parameter x_i to the interval $[a_i, b_i]$. In the sequel we will denote the set $x_i \varepsilon [a_i, b_i]$, $i = 1, 2, \cdots, k$, by A.

3. Convergence theorems. We now make the following assumptions:

(i) $$Y_t(\mathbf{x}) = \sum_{j=1}^{N} g_j(\mathbf{x}) V_{j,t} \qquad N < \infty$$

in which the $V_{j,t}$ are ergodic random processes which are bounded in magnitude w.p. 1 and the g_j are functions whose second partial derivatives with respect to the x_i are bounded for all $\mathbf{x} \varepsilon A$

(ii) $$(\text{grad } M(\mathbf{z})|_{\mathbf{z}=\mathbf{x}}, \mathbf{x} - \boldsymbol{\theta}) \geqq K_0 \|\mathbf{x} - \boldsymbol{\theta}\|^2$$

$$\|\text{grad } M(\mathbf{x})\|^2 \leqq K_1 \|\mathbf{x} - \boldsymbol{\theta}\|^2$$

$$\text{all } \mathbf{x} \varepsilon A, \qquad 0 < K_0 \leqq K_1 < \infty$$

$$|\partial^3 M/\partial x_i^3| \leqq P \qquad \text{all } \mathbf{x} \varepsilon A,\ i = 1, 2, \cdots, k$$

$$a_i + 2\delta \leqq \theta_i \leqq b_i - 2\delta \qquad i = 1, 2, \cdots, k.$$

(iii) Let $D_{t+\rho}$ be any one of the random processes $V_{l,t+\rho}V_{m,t+\rho}$, $l, m = 1, 2, \cdots, N$ or $V_{l,t+\rho}$, $l = 1, 2, \cdots, N$ and let F_t be any bounded functional on the processes $V_{l,\tau}$, $l = 1, 2, \cdots, N$, $\tau \leqq t$ and $R_{FD}(\rho) = E\{(F_t - E\{F_t\})(D_{t+\rho} - E\{D_{t+\rho}\})\}$ then we require for all $\rho \geqq \rho_0$, $\rho_0 < \infty$

$$|R_{FD}(\rho)| \leqq \sigma_F \sigma_D (K_2/\rho^2), \qquad K_2 < \infty$$

(iv) $$\int_0^\infty a(t)\, dt = \infty, \qquad \int_0^\infty a(t)c^2(t)\, dt < \infty$$

$$\int_0^\infty a(t)a(\tfrac{1}{2}t)\, dt < \infty, \qquad \text{and} \qquad \int_1^\infty a(t)t^{-1}\, dt < \infty.$$

Under the above conditions we can make the following statements:

THEOREM 1. *Conditions* (i)–(iv) *imply*

$$\lim_{t\to\infty} E\{\|\mathbf{X}_t - \boldsymbol{\theta}\|^2\} = 0 \qquad \text{for all } \mathbf{x}(0) \varepsilon A.$$

In order to realize the approximation procedure we must be able to generate functions $a(t)$ and $c(t)$ satisfying Condition (iv). It will be convenient to use functions which are functionally simple; for this reason we consider

(9) $$a(t) = a/(t + 1)^\alpha \qquad c(t) = c/(t + 1)^\gamma.$$

In order that Condition (iv) be satisfied we require $\frac{1}{2} < \alpha \leqq 1$, $\gamma > \frac{1}{2}(1 - \alpha)$. If $\alpha = 1$ we will require in addition that

(10) $$a > 4K_0^{-1}.$$

Relative to this class of functions we can state:

THEOREM 2. *Conditions* (i)–(iv) *and the choice* $\alpha = 1$, $\gamma \geqq \frac{1}{4}$ *imply*

$$E\{\|\mathbf{X}_t - \boldsymbol{\theta}\|^2\} \leqq K_3/(t + 1) \qquad \text{all } \mathbf{x}(0) \varepsilon A, K_3 < \infty;$$

moreover, no choice of α and γ will yield a faster rate of convergence for all situations satisfying Conditions (i)–(iv).

Before giving the proofs of the above Theorems, let us comment briefly on the hypotheses. Conditions (ii) and (iv) are of the usual variety and (i) delineates the form of processes to be considered; only (iii) requires comment. It would appear at first that (iii) is rather restrictive in that it holds uniformly for all bounded functionals F_t; however, this turns out to be merely a requirement that the processes $V_{j,t}$ not be too "predictable." Consider trying to predict $D_{t+\rho}$ with some bounded operation, say F_t, on the past of the processes. If we use F_t in a linear manner to predict $D_{t+\rho}$

$$\hat{D}_{t+\rho} = E\{D_{t+\rho}\} + R_{FD}(\rho)(\sigma_F^2)^{-1}(F_t - E\{F_t\})$$

the mean square prediction error is $\epsilon(\rho) = (\sigma_F^2)^{-1}(\sigma_F^2\sigma_D^2 - R_{FD}^2(\rho))$ and thus our requirement is simply a condition that the minimum mean square prediction error approach its asymptotic value as fast as the inverse of the fourth power of the prediction time. Another way of viewing this requirement is as follows: if we were to represent $V_{j,t}$ in the form $V_{j,t} = \int_0^\infty h(\tau)N_{j,t-\tau}\,d\tau$ in which $N_{j,t}$ was a process bounded in magnitude w.p. 1 and such that $N_{j,t}$ and $N_{j,t+\tau}$ were statistically independent for $\tau \geqq \tau_0 > 0$ (e.g. $N_{j,t}$ might be a "shot noise" process) then Condition (iii) would be satisfied if for all $\tau \geqq T > 0$, [12],

$$|h(\tau)| \leqq K/\tau^3 \qquad K < \infty.$$

Although this rate of decrease of the magnitude of the impulse response (resolvent kernel) of the linear transformation is more rapid than that required for a general stable linear transformation, it is considerably slower than that of most physical transformations. It should be noted that we require the rate in Condition (iii) to be ρ^2 only to establish Theorem 2; Theorem 1 remains true if this is weakened to $\rho^{1+\epsilon}$, $\epsilon > 0$. To require that the processes in question be bounded in magnitude w.p. 1 is no real restriction since this will be true for any physically observable random process. We now proceed with the proofs of the Theorems.

Proof of Theorem 1. From Equation (3) we have

$$(d/dt)\|\mathbf{X}_t - \boldsymbol{\theta}\|^2 = -2a(t)(\mathbf{X}_t - \boldsymbol{\theta}, c^{-1}(t)\mathbf{Z}_t(\mathbf{X}_t)). \tag{11}$$

The right hand side of this equation is bounded in magnitude w.p. 1 for all t by Condition (i) and so is $\|\mathbf{X}_t - \boldsymbol{\theta}\|^2$; thus by a Theorem of Kolmogoroff [9]

$$E\{(d/dt)\|\mathbf{X}_t - \boldsymbol{\theta}\|^2\} = (d/dt)E\{\|\mathbf{X}_t - \boldsymbol{\theta}\|^2\}. \tag{12}$$

For brevity we denote

$$b(t) = E\{\|\mathbf{X}_t - \boldsymbol{\theta}\|^2\}. \tag{13}$$

Adding and subtracting a term from the right hand side of Equation (11) and

taking expected values yields

$$(14)\qquad \begin{aligned}(d/dt)b(t) &= 2a(t)E\{(\mathbf{X}_t - \boldsymbol{\theta}, -\mathbf{Q}_c(\mathbf{X}_t))\} \\ &\quad + 2a(t)E\{(\mathbf{X}_t - \boldsymbol{\theta}, \mathbf{Q}_c(X_t) - c^{-1}(t)\mathbf{Z}_t(\mathbf{X}_t))\}.\end{aligned}$$

We now develop suitable bounds for the two terms on the right hand side of this equation. First consider $\mathbf{M}_c(\mathbf{x})$. By means of a Taylor's series we can express the ith component of this vector as

$$(15)\qquad M_{c,i}(\mathbf{x}) = 2[(\partial M(\mathbf{z})/\partial x_i)|_{\mathbf{z}=\mathbf{x}} + \tfrac{1}{6}c^2R], \qquad |R| \leqq P.$$

Thus, using Condition (ii)

$$(16)\qquad (\mathbf{x} - \boldsymbol{\theta}, -\mathbf{M}_c(\mathbf{x})) \leqq -2K_0\|\mathbf{x} - \boldsymbol{\theta}\|^2 + k^{\frac{1}{3}}P\tfrac{1}{3}c^2\|\mathbf{x} - \boldsymbol{\theta}\|.$$

Next consider $\mathbf{Q}_c(\mathbf{x})$. If $a_i + \delta \leqq x_i \leqq b_i - \delta$ for $i = 1, 2, \cdots, k$, then $\mathbf{Q}_c(\mathbf{x}) = \mathbf{M}_c(\mathbf{x})$ and if $b_i - \delta \leqq x_i \leqq b_i$, the ith component of $\mathbf{Q}_x(\mathbf{x})$ is

$$(17)\quad Q_{c,i}(\mathbf{x}) = M_{c,i}(\mathbf{x})(1 - G_i^+(x_i)) + \epsilon_y^{-1}G_i^+(x_i)[M_{c,i}^2(\mathbf{x}) + c^{-2}(t)\sigma_{Y_i}^2].$$

Now for x_i in the interval assumed $\partial M/\partial x_i$ is positive and bounded away from zero. We now assume that ϵ_y has been taken small enough that

$$(18)\qquad \epsilon_y \leqq \inf_{\substack{\mathbf{x}\varepsilon A, a_i \leqq x_i \leqq a_i+\delta \\ b_i-\delta \leqq x_i \leqq b_i, i=1,2,\cdots,k}} \partial M/\partial x_i\,, \qquad \epsilon_y > 0.$$

Now substituting Equation (15) into Equation (17), using inequality (18), and weakening the resulting inequality by omitting a negative term, we have

$$(19)\qquad -Q_{c,i}(\mathbf{x}) \leqq -2\partial M(z)/\partial x_i\,|_{\mathbf{z}=\mathbf{x}} + c^2K_4/k^{\frac{1}{3}}, \qquad b_i - \delta \leqq x_i \leqq b_i$$

$$(20)\qquad K_4/k^{\frac{1}{3}} = \frac{1}{3}\left[P + \frac{4}{\epsilon_y}P \sup_{\substack{\mathbf{x}\varepsilon A, a_i \leqq x_i \leqq a_i+\delta \\ b_i-\delta \leqq x_i \leqq b_i, i=1,2,\cdots,k}} \frac{\partial M}{\partial x_i} - \frac{1}{\epsilon_y}\frac{c^2(0)}{3}P^2\right].$$

A similar argument applies to $a_i \leqq x_i \leqq a_i + \delta$, hence for either $x_i \varepsilon [a_i\,, a_i + \delta]$ or $[b_i - \delta, b_i]$ we have

$$(21)\quad -(x_i - \theta_i)Q_{c,i}(\mathbf{x}) \leqq -2(x_i - \theta_i)\partial M(\mathbf{z})/\partial x_i\,|_{\mathbf{z}=\mathbf{x}} + c^2K_4(x_i - \theta_i)/k^{\frac{1}{3}}.$$

The first term on the right hand side of the inequality being negative. Thus, combining inequalities (16) and (21), we have for any $\mathbf{x}\,\varepsilon\,A$

$$(22)\qquad (\mathbf{x} - \boldsymbol{\theta}, -\mathbf{Q}_c(\mathbf{x})) \leqq -2K_0\|\mathbf{x} - \boldsymbol{\theta}\|^2 + c^2K_4\|\mathbf{x} - \boldsymbol{\theta}\|.$$

Since this holds for any $\mathbf{x}\,\varepsilon\,A$ it also holds for any X_t generated by the approximation procedure up to time t; thus, substituting X_t into inequality (22) and taking expected values of both sides we see that we can bound the first term on the right hand side of Equation (14) by

$$(23)\qquad B_1(t) = -4a(t)K_0b(t) + 2a(t)c^2(t)K_4E\{\|\mathbf{X}_t - \boldsymbol{\theta}\|\}.$$

We now consider the second term on the right hand side of Equation (14).

The ith term in this inner product is

$$T_i = 2a(t)E\{(X_{i,t} - \theta_i)[Q_{c,i}(\mathbf{X}_t) - c^{-1}(t)Z_{i,t}(\mathbf{X}_t)]\} \tag{24}$$

which, by using Equations (2)–(6) and Condition (i), may be expressed in the form

$$T_i = 2a(t)\sum_{j=1}^{N}\sum_{k=1}^{N} E\{f_{jk}(\mathbf{X}_t, c(t))[V_{j,t}V_{k,t} - E\{V_{j,t}V_{k,t}\}]\} + \sum_{j=1}^{N} E\{f_j(\mathbf{X}_t, c(t))[V_{j,t} - E\{V_{j,t}\}]\}. \tag{25}$$

By Condition (i) all of the $f_{jk}(\mathbf{x})$ and $f_j(\mathbf{x})$ appearing in this expression are bounded and possess bounded first partial derivatives with respect to all of the x_i for all $\mathbf{x} \varepsilon A$ and all $t \geqq 0$. Consider the l, mth term in this expression and for brevity denote $V_{l,t}V_{m,t}$ by D_t. Then this term can be written

$$T_{iml} = -2a(t)\sum_{q=1}^{k}\int_0^t a(\tau)E\left\{\frac{\partial f_{lm}}{\partial X_q}(\mathbf{X}_\tau, c(\tau))\frac{Z_{q,\tau}}{c(\tau)}(D_t - E\{D_t\})\right\}d\tau \tag{26}$$

in which the order of integration and expectation has been exchanged since the process in the integrand is bounded in magnitude w.p. 1 [9]. Now, since the processes inside the expectation are bounded w.p. 1 for all l, m, and i, we can use Condition (iii) to bound the expectation in the integrand in magnitude by

$$\begin{aligned} L &= K_5 && t - 1 \leqq \tau \leqq t \\ &= K_{5/(t-\tau)^2} && 0 \leqq \tau \leqq t - 1. \end{aligned} \tag{27}$$

Substituting this bound into Equation (26), splitting the integral up into integrals over $[0, \frac{1}{2}t]$, $[\frac{1}{2}t, t - 1]$, and $[t - 1, t]$, and overbounding each integral, yields

$$|T_{iml}| \leqq 2a(t)[K_6 a(\tfrac{1}{2}t) + K_7\mu(t - 1)t^{-1}]/kN(N + 1) \qquad K_6, K_7 < \infty, \tag{28}$$

in which

$$\begin{aligned} \mu(t) &= 0, && t < 0 \\ &= 1, && t \geqq 0. \end{aligned}$$

Thus the second term on the right hand side of Equation (14) may be bounded in magnitude by

$$B_2(t) = 2a(t)[K_6 a(\tfrac{1}{2}t) + K_7\mu(t - 1)t^{-1}] \tag{29}$$

combining this bound with that of Equation (23) and inserting in Equation (14) yields

$$\frac{db(t)}{dt} \leqq -4a(t)K_0 b(t) + 2a(t)c^2(t)K_4 E\{\|\mathbf{X}_t - \boldsymbol{\theta}\|\} + 2a(t)[K_6 a(\tfrac{1}{2}t) + K_7\mu(t - 1)t^{-1}]. \tag{30}$$

Now, for any $\epsilon_t > 0$

$$E\{\|\mathbf{X}_t - \boldsymbol{\theta}\|\} \leqq \epsilon_t + \epsilon_t^{-1} E\{\|\mathbf{X}_t - \boldsymbol{\theta}\|^2\} = \epsilon_t + b(t)\epsilon_t^{-1}. \tag{31}$$

Applying inequality (31) to inequality (30) for the choice

$$\epsilon_t = 2K_4 c^2(t)(K_0\epsilon)^{-1} \qquad 0 < \epsilon < 4$$

yields

$$(d/dt)b(t) + p(t)b(t) \leqq q(t) \tag{32}$$

in which $p(t) = (4 - \epsilon)K_0 a(t) > 0$ and

$$q(t) = 2a(t)[K_6 a(t/2) + K_7\mu(t-1)t^{-1}] + 4a(t)c^4(t)K_4 2(\epsilon K_0)^{-1} \geqq 0.$$

Integrating both sides of Equation (32) from 0 to t yields

$$b(t) + \int_0^t p(\tau)b(\tau)\,d\tau \leqq \int_0^t q(\tau)\,d\tau + b(0) \tag{33}$$

$$b(0) = \|\mathbf{x}(0) - \boldsymbol{\theta}\|^2.$$

Now consider the integral equation

$$b_0(t) + \int_0^t p(\tau)b_0(\tau)\,d\tau = \int_0^t q(\tau)\,d\tau + b(0) \tag{34}$$

with the solution

$$b_0(t) = b_0 \exp\left[-\int_0^t p(\tau)\,d\tau\right] + \int_0^\infty f_{[0,t]}(\tau)\,d\tau \tag{35}$$

in which

$$f_{[0,t]}(\tau) = \exp\left[-\int_\tau^t p(\xi)\,d\xi\right] q(\tau) \qquad 0 \leqq \tau \leqq t \tag{36}$$

$$= 0 \qquad \text{elsewhere.}$$

Now the non-negativeness of $p(t)$ and $g(t)$ and the continuity of $b(t)$ and $b_0(t)$ guarantee that, for any function $b(t)$ which satisfies inequality (33)

$$b(t) \leqq b_0(t) \qquad \text{all } t > 0 \tag{37}$$

(this is easily shown by assuming the contrary and reaching a contradiction). Thus we focus our attention on bounding $b_0(t)$. Now $0 \leqq f_{[0,t]}(\tau) \leqq q(\tau)$ for all t and τ greater than 0 and by Condition (iv) $q(\tau)$ is integrable, thus by the general convergence theorem of Lebesgue and Condition (iv)

$$\lim_{t\to\infty} b_0(t) = \lim_{t\to\infty} b(0) \exp\left[-\int_0^t p(\tau)\,d\tau\right] \tag{38}$$

$$+ \int_0^\infty \lim_{t\to\infty} f_{[0,t]}(\tau)\,d\tau = 0$$

which completes the proof of Theorem 1.

To complete the proof of Theorem 2, we merely note that for $a(t)$ and $c(t)$ as specified in the statement of the theorem

$$(39) \qquad \begin{aligned} p(t) &= (4 - \epsilon)K_0 a(t + 1)^{-1}, & t > 0 \\ q(t) &\leqq K_8(t + 1)^{-2}, & t > 0. \end{aligned}$$

Substituting these expressions in Equation (35) and carrying out the integration, we obtain

$$(40) \qquad \begin{aligned} b_0(t) \leqq\ & b(0)(t + 1)^{-(4-\epsilon)K_0 a} \\ & + K_8\{[(4 - \epsilon)K_0 a - 1](t + 1)^{-1} - (t + 1)^{-(4-\epsilon)K_0 a}\}, \quad 0 < \epsilon < 4 \end{aligned}$$

thus for $a > 4K_0^{-1}$ we reach the positive side of Theorem 2.

The "minimax" property of the choice $\alpha = 1$, $\gamma \geqq \frac{1}{4}$ is most easily established by a one-dimensional example $Y_t(x) = [D_t - xV_t]^2$ in which

$$D_t = \sum_{j=-\infty}^{\infty} D_j p(t - jT - \theta), \qquad V_t = \sum_{j=-\infty}^{\infty} V_j p(t - jT - \theta)$$

the quantity θ being a random variable uniformly distributed between 0 and T and the random variables D_j and V_j are statistically independent of the random variables D_m and V_m, $m \neq j$. The function $p(t)$ is zero outside the interval $[0, T]$. This simple situation can be analyzed in terms of an equivalent discrete time-parameter process. In this case the process is independent of c, and is in fact simply a Robbins-Munro process on the derivative of $M(x) = E\{Y_t(x)\}$. This process has been studied extensively; and, that the rate of convergence cannot exceed n^{-1} for a broad class of situations, is established by the work of Schmetterer [13], Chung [3], or Sacks [10].

4. An application of the theorems. The most obvious application of the methods discussed here to the field of communications and data processing is the optimum filter or predictor problem. Here we state the filtering (or prediction) problem in an appropriate setting and list a set of restrictions which are sufficient to guarantee that the conditions of the theorems are met. From the standpoint of physical applications, these restrictions seem to admit almost all situations of practical interest.

The form of filter (or predictor) to be considered is shown in Figure 1. The process V_t is the one observed and the process S_t is the one we desire to estimate. This form is general in that any filter which operates on only a finite interval of the input process can be approximated arbitrarily closely by such a form [2]. The parameters $x_1, x_2, \cdots, x_k$ are to be adjusted by the method of Section 2 in order to minimize $M(\mathbf{x}) = E\{W[S_t - Q_t(\mathbf{x})]\}$ in which W is some appropriate weighting function on the error, $S - Q$. For a discussion of how this procedure can be mechanized by analog simulation and of the restrictions involved, the reader is referred to Sakrison [11].

The following restrictions on the processes involved and the error weighting

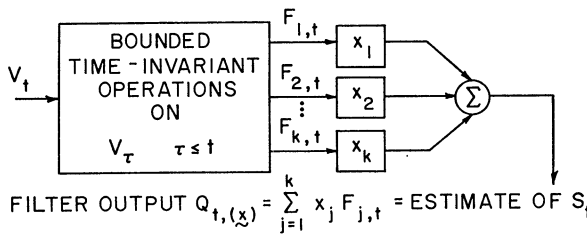

FIG. 1. Form of filter to be designed.

function are sufficient to guarantee that Conditions (i)–(iii) of the theorems are satisfied:

(a) The processes $F_{i,t}$ $i = 1, 2, \cdots, k$ and S_t are jointly ergodic and bounded in magnitude w.p. 1.

(b) The correlation coefficient between any one of the F_i and any linear combination of the remaining F_j is unequal to ± 1.

(c) The function $W[e]$ is assumed to be a polynomial of degree N, $N < \infty$, and required to be "strictly" convex in the sense that

$$W[\alpha a + (1 - \alpha)b] \leqq \alpha W[a] + (1 - \alpha)W[b] - E\alpha^2|a - b|^2$$

$$0 \leqq \alpha \leqq 1, \qquad E \geqq \epsilon > 0 \quad \text{for} \quad 0 \leqq \alpha \leqq \epsilon_0 > 0.$$

(d) Condition (iii) is assumed to hold for any random variable D_t of the form

$$D_t = (S_t)^{q_0} \prod_{j=1}^{k} (F_{j,t})^{q_j}, \qquad \sum_{j=0}^{k} q_j \leqq 2N.$$

With the exception of restriction (c) these restrictions and their relation to the conditions of the Theorems are quite straightforward. Restriction (c) merely states that the error weighting polynomial W must consist of a convex function plus a positive quadratic term. For the proof that these conditions are sufficient to satisfy the assumptions of the theorems see [12].

Acknowledgment. The author wishes to gratefully acknowledge several helpful discussions with Mr. Lee Gardner of Lincoln Laboratories and Professor Gian Carlo Rota of Massachusetts Institute of Technology.

REFERENCES

[1] BLUM, J. R. (1954). Multidimensional stochastic approximation procedures. *Ann. Math. Statist.* **25** 737–744.

[2] CAMERON, R. H. and MARTIN, W. T. (1947). The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals. *Ann. of Math.* **48** 385–392.

[3] CHUNG, K. L. (1954). On a stochastic approximation method. *Ann. Math. Statist.* **25** 463–483.

[4] DRIML, M. and HANŠ, O. (1960). Continuous stochastic approximations. *Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes*. Czechloslovak Academy of Sciences, Prague 113–122.

[5] DRIML, M. and NEDOMA, J. (1960). Stochastic approximations for continuous random processes. Same Vol. as [4]. 145–158.

[6] DUPAČ, V. (1957). On the Kiefer-Wolfowitz approximation method. *Časopis Pěst. Mat.* **82** 47–75. (An English Translation exists as report no. 22G-0008, Lincoln Laboratory, Lexington, Mass. 24 Feb., 1960. M. D. Friedman, translator, and L. A. Gardner and E. J. Magee Editors.)

[7] HANŠ, O. and ŠPAČEK, A. (1960). Random fixed point approximation by differentiable trajectories. *Transactions of the Second Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes.* Czechloslovak Academy of Sciences, Prague 203–213.

[8] KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23** 462–466.

[9] KOLMOGOROFF, A. M. (1956). *Foundations of the Theory of Probability.* Chelsea, New York.

[10] SACKS, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.* **29** 373–405.

[11] SAKRISON, D. (1963). Iterative design of optimum filters for non-mean-square error criteria. *Transactions of the I.E.E.E. Professional Group on Information Theory.* **9** No. 3 161–167.

[12] SAKRISON, D. (1962). Design of filters for non-mean-square error performance criteria by means of a continuous adjustment procedure. Quarterly Progress Report No. 66 189–201 and No. 67 119–126, Research Laboratory of Electronics, Massachusetts Institute of Technology.

[13] SCHMETTERER, L. (1953). Bermerkungen Zum Verfahren der Stochastischen Iteration. *Österreich. Ingenieur-Archiv.* VII 111–117.