UNIVERSITY OF HOUSTON · Founded 1927 ·

DEPARTMENT OF MATHEMATICS

UNIVERSITY OF HOUSTON HOUSTON, TEXAS

NASA CR-147522

(NASA-CR-147522) MULTIDIMENSIONAL STOCHASTIC APPROXIMATION USING LOCALLY CONTRACTIVE FUNCTIONS (Houston Univ.) 13 p HC $3.50 CSCL 12A

N76-20913

Unclas

G3/65 21494

MULTIDIMEN. STOCHASTIC APPROX.
USING LOCALLY CONTRACTIVE FNS.
W. M. LAWTON, INST FOR ADV STUDY
PRINCETON, NEW JERSEY
REPORT #46 AUGUST 1975

PREPARED FOR
EARTH OBSERVATION DIVISION, JSC
UNDER
CONTRACT NAS-9-12777

3801 CULLEN BLVD.
HOUSTON, TEXAS 77004

Multidimensional Stochastic Approximation Using Locally Contractive Functions

August, 1975

by

Wayne M. Lawton
Institute for Advanced Study
Princeton, New Jersey

Report #46
NAS-9-12777

Bibliography

1. C.C. Blaydon, K.S. Fu, and R.L. Kashyap, "Stochastic approximation", Adaptive, Learning and Pattern Recognition Systems, Academic Press, New York and London, 1970, Edited by J.M. Mendel and K.S. Fu.

2. A.S. Householder, Theory of Matrices and Numerical Analysis, Blaisdell Publishing Co., New York, 1964.

3. B. Charles Peters, Jr. and Homer F. Walker, "An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture or normal distributions", to appear.

Multidimensional Stochastic Approximation Using Locally Contractive Functions

1. Summary. A Robbins-Monro type multidimensional stochastic approximation algorithm which converges in mean square and with probability one to the fixed point of a locally contractive regression function is developed. The algorithm is applied to obtain maximum likelihood estimates of the parameters for a mixture of multivariate normal distributions.

2. Introduction. Let E_k be real k-dimensional Euclidean space with inner product denoted by $< , >$ and norm denoted by $\|\cdot\|$. Corresponding to every positive definite real $k\times k$ matrix B we define the B-inner product $<x,y>_B = <x,By>$ and the B-norm $\|x\|_B = <x,Bx>^{1/2}$, for $x,y \in E_k$. A function $\overline{F}:D \longrightarrow E_k$, where D is an open subset of E_k, is locally contractive at a point $\theta^\circ \in D$ if there exists a B-norm on E_k and a number λ, $0 \leq \lambda < 1$ such that

$$\|\theta^\circ - \overline{F}(\theta)\|_B \leq \lambda \|\theta - \theta\|_B \tag{2.1}$$

whenever θ is sufficiently near θ°. If the above inequality holds for every θ in some neighborhood W of θ°, we say $\overline{F}$ is λ-locally contractive at θ° throughout W. Clearly, θ° will be a unique fixed point in W for $\overline{F}$. For amy $k\times k$ matrix A, let the spectral radius of A be denoted by $\rho(A) = \sup\{|\lambda| : \lambda$ is an eigenvalue of $A\}$. The Frechet derivative of $\overline{F}$, if it exists, will be denoted by $\nabla\overline{F}$. The following result, a consequence of Taylor's theorem and the theory in [2; section 2.3], will be used in part

4 of this paper.

(2.2) Lemma If $\nabla\overline{F}$ exists and is continuous in a neighborhood of θ^o, a necessary and sufficient condition for $\overline{F}$ to be locally contractive at θ^o is that $\overline{F}(\theta^o) = \theta^o$ and $\rho(\nabla\overline{F}(\theta^o)) < 1$.

Let $\{\overline{Y}(\theta):\theta \in D\}$ be a family of random variables with values in E_k satisfying the following conditions

(2.3) $\sup\limits_{\theta\in D} E(\ \overline{Y}(\theta)\ ^2) < \infty$ (E denotes conditional expectation with θ fixed)

(2.4) the regression function of $\{\overline{Y}(\theta)\}$, denoted by $\overline{M}(\theta) = E(\overline{Y}(\theta))$, can be expressed as $\overline{M}(\theta) = \theta - \overline{F}(\theta)$ where $\overline{F}(\theta)$ is locally contractive at $\theta^o \in D$.

In part 3 of this paper we develop an algorithm which, given the conditions above and given a sufficiently close approximation to θ^o, yields a sequence of recursively defined random variables with values in E_k which converges to θ^o in mean square and with probability one.

In part 4 of this paper, this algorithm is used to formulate a stochastic analog of the iterative procedure developed by Peters and Walker in [3] for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions.

(3) Derivation of the Algorithm

(3.1) Lemma Let $\rho_1 = (\rho(B^{-1}))^{-1}$ and $\rho_2 = \rho(B)$ where B is a positive definite $k\times k$ matrix, and let r_1, r_2 be positive real numbers such that

$\frac{r_1}{r_2} \le \frac{\rho_1}{\rho_2}$. Let θ°, $\theta \in E_k$ such that $\|\theta^\circ - \bar{\theta}\| \le r_1$ and let $S = \{\theta \in E_k : \|\theta - \bar{\theta}\| \le r_2\}$. Define a function $\varphi : E_k \longrightarrow S$ as follows

$$\varphi(\theta) = \begin{cases} \theta \quad \text{if} \quad \theta \in S \\ (1-t)\bar{\theta} + t\theta \quad \text{where} \quad t = \dfrac{r_2}{\|\bar{\theta}-\theta\|} \quad \text{if} \quad \theta \notin S \end{cases}$$

Then the following inequality holds for every $\theta \in E_k$

$$\|\theta^\circ - \quad (\theta)\|_B^2 \le \|\theta^\circ - \theta\|_B$$

Proof. We may assume $t = \dfrac{r_2}{\|\bar{\theta}-\theta\|} < 1$ for otherwise $\theta \in S$ implying $\varphi(\theta) = \theta$. By the fundamental theorem of calculus:

$$\|\theta^\circ - \theta\|_B^2 = \|\theta^\circ - \varphi(\theta)\|_B^2 + \int_{s=t}^{s=1} \frac{\partial}{\partial s}[\|\theta^\circ - (1-s)\bar{\theta} - s\theta\|_B^2]ds$$

It suffices to prove that the integrand is nonnegative for all s, $t \le S \le 1$. Consider

$$\frac{\partial}{\partial s}[\|\theta - (1-s)\bar{\theta} - s\theta\|_B^2] = 2<\bar{\theta} - \theta, \theta^\circ - (1-s)\bar{\theta} - s\theta>_B$$

$$= 2s<\bar{\theta} - \theta, \bar{\theta} - \theta>_B + 2<\bar{\theta} - \theta, \theta^\circ - \bar{\theta}>_B.$$

Since $t \le s \le 1$, by the principle axis theorem for real symmetric matrices, the left side of the above expression is bounded below by $2\ r_2\|\bar{\theta}-\theta\|\ \rho_1$. Similarly, the right hand side of the above expression is bounded, in absolute value, by $2\ r_1\|\theta-\theta\|\rho_2$. The result follows from the hypothesis concerning r_1, r_2, ρ_1 and ρ_2.

For the remainder of part 3 we adopt the notation and assumptions stated previously. Furthermore, we assume that $\bar{F}$ is locally contractive at $\theta^\bullet$ throughout S and that $S \subseteq D$. Define a family $\{Y(\theta):\theta \in E_k\}$ of random variables by

$$(3.2) \qquad Y(\theta) = \bar{Y}(\varphi(\theta)) - \mathscr{G}(\theta) + \theta$$

Then the following inequality is valid for every $\varepsilon > 0$.

(3.3) <u>Corollary</u> $\inf_{\varepsilon\leq \|\theta-\theta^\circ\|_B} E(<\theta-\theta^\circ, Y(\theta)>_B) > 0$

<u>Proof</u>. Since $E(Y(\theta)) = \mathscr{G}(\theta) - \bar{F}(\mathscr{G}(\theta)) - \varphi(\theta) + \theta$, the expression above $=$ $\inf_{\varepsilon\leq \|\theta-\theta^\circ\|_B} \{<\theta-\theta^\circ, \theta-\theta^\circ>_B + <\theta-\theta^\circ, \theta^\circ - \bar{F}(\varphi(\theta))>_B\}$. The first term above $= \|\theta-\theta^\circ\|^2_B$, and the second term is bounded above in absolute value by

$$\|\theta-\theta^\circ\|_B\ \|\bar{F}(\varphi(\theta)) - \theta^\circ\|_B$$

$$\leq \lambda\|\theta-\theta^\circ\|_B\ \|\mathscr{G}(\theta) - \theta^\circ\|_B \qquad \text{(by (2.1))}$$

$$\leq \lambda\|\theta-\theta^\circ\|^2_B \qquad \text{(by (3.1))}$$

Therefore $\inf_{\varepsilon\leq\|\theta-\theta^\circ\|_B} E(<\theta-\theta^\circ, Y(\theta)>_B) \geq (1-\lambda)\varepsilon^2 > 0.$

(3.4) Definition. A gain sequence is a sequence $\{a_\ell\}$ of positive numbers satisfying $\sum_{\ell=1}^{\infty} a_\ell = \infty$ and $\sum_{\ell=1}^{\infty} a_\ell^2 < \infty$.

(3.5) Remark. For any $C > 0$, $\{a_\ell = \frac{c}{\ell}\}$ is a gain sequence since $\underset{k \to \infty}{\text{limit}} \sum_{\ell=1}^{K} \frac{c}{\ell} = \underset{k \to \infty}{\text{limit}}\ c \log K = \infty$ and $\sum_{\ell=1}^{\infty} a_\ell^2 = \frac{c\pi^2}{6}$.

(3.6) Theorem. Let $\{Y(\theta):\theta \in E_k\}$ be as in (3.2) and let $\{a_\ell\}$ be a gain sequence. Then the following sequence of recursively defined random vectors

(3.7) $\theta_{\ell+1} = \theta_\ell - a_\ell Y(\theta_\ell)$, θ_1 arbitrarily chosen, converges in mean square and with probability one to θ.

Proof. We refer the reader to the algorithm described in [1,pp 332-333] and the convergence proof given in the appendix to [1,pp 350-352]. Replacing their gain sequence $\{\rho_k\}$ with the gain sequence $\{a_\ell\}$, and replacing their norm $\|\cdot\|$ and inner product $<V,W> = V'W$ (where V' denotes the transpose of the vector V and $V'W$ denotes matrix multiplication) with the B-norm and B-inner product respectively, the theorem will follow once we verify that conditions (A1) - (A3) in [1,pp 332-333] are satisfied.

(A1) Since $E(Y(\theta)) = \theta - \overline{F}(\mathcal{J}(\theta))$, this result follows from (3.3) since B positive definite implies $\|\theta-\theta^\circ\|_B = 0$ if and only if $\theta = \theta^\circ$.

(A2) Follows from (3.3).

(A3) $E(\|Y(\theta)\|_B^2) = E(<\overline{Y}(\mathcal{J}(\theta)) - \mathcal{J}(\theta) + \theta,\ \overline{Y}(\mathcal{J}(\theta)) - \mathcal{J}(\theta) + \theta>_B)$
$\leq h(1 + \|\theta-\theta^\circ\|_B^2)$ for sufficiently large $h > 0$ since $S \subseteq D$ and by (2.3) $\sup_{\theta \in D} E(\|\overline{Y}(\theta)\|^2) < \infty$.

4. Application to Maximum Likelihood Estimation

Let $D = \{(\alpha_i, \mu_i, \Sigma_i)\}_{i=1,\dots,m}$ where each $\alpha_i > 0$ with $\sum_{i=1}^{m} \alpha_i = 1$, each $\mu_i \in R^n$ and each Σ_i is a positive definite real symmetric $n\times n$ matrix. We consider D to be realized as an open subset D of E_k where $k = \frac{m(n+1)(n+2)}{2} - 1$. For each $\theta = (\alpha_i, \mu_i, \Sigma_i)_{i=1,\dots,m}$ let $x(\theta)$ be a random variable with values in R^n and with distribution function

$$p(\theta,x) = \sum_{i=1}^{m} \alpha_i \, p_i(x) \qquad \text{for} \quad x \in R^n$$

where

$$p_i(\theta,x) = (2\pi)^{-n/2} |\Sigma_i|^{-1/2} \exp\{-\tfrac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i)\}$$

for each $i = 1,\dots,m$.

Fix $\theta^\circ \in D$ and let $\{x_k\}_{k=1,\dots,N} \subseteq R^n$ be an independent sample of observations of $x(\theta^\circ)$. A <u>maximum-likelihood estimate</u> of θ° based on $\{x_k\}$ is a choice of $\theta \in D$ which locally maximizes the <u>log-likelihood function</u>

$$L = \sum_{k=1}^{N} \log p(\theta, x_k)$$

In the appendix to [3], Peters and Walker prove there exists a sufficiently small neighborhood of θ°, such that with probability → 1 as $N \to \infty$, there exists a unique maximum-likelihood estimate of θ° in that neighborhood. Furthermore, with probability → 1 as $N \to \infty$, this estimate $\to \theta^\circ$. This estimate will be called the consistent <u>maximum-likelihood estimate</u> (which we abbreviate by c.m.l.e.)

Equating $\nabla_\theta L = 0$ and performing algebraic manipulation of the resulting equations, yields the following necessary conditions for a m.l.e. θ of θ°.

$\theta = \mathcal{I}(\theta)$ where $\mathcal{I}: D \to D$ is defined by

$$\mathcal{I}((\alpha_i,\mu_i,\Sigma_i)_{i=1,\ldots,m}) = (\alpha_i',\mu_i',\Sigma_i')_{i=1,\ldots,m}$$

where, for each $i = 1,\ldots,m$

$$\alpha_i' = \frac{\alpha_i}{N} \sum_{k=1}^{N} x_k \frac{p_i(x_k)}{p(x_k)} \tag{4.1}$$

$$\mu_i' = \{\frac{1}{N} \sum_{k=1}^{N} x_k \frac{p_i(x_k)}{p(x_k)}\} \Big/ \{\frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\} \tag{4.2}$$

$$\Sigma_i' = \{\frac{1}{N} \sum_{k=1}^{N} (x_k - {}_i)(x_k - {}_i)^T \frac{p_i(x_k)}{p(x_k)}\} \Big/ \{\frac{1}{N} \sum_{k=1}^{N} \frac{p_i(x_k)}{p(x_k)}\} \tag{4.3}$$

where each p_i and p is evaluated with respect to the parameters $\theta = (\alpha_i,\mu_i,\Sigma_i)_{i=1,\ldots,m}$. $\mathcal{I}$ will be called the <u>likelihood function</u>.

In [3], Peters and Walker develop an iterative procedure which, starting with any initial estimate θ' which is sufficiently close to θ°, yields a sequence in D converging to the c.m.l.e. of θ° based on $\{x_k\}_{k=1,\ldots,N}$. Their technique consists in proving that, for $\varepsilon < \frac{4}{m(n+1)(n+2)}$, the function

$$\mathcal{I}_\varepsilon(\theta) = (1-\varepsilon)\theta + \varepsilon\, \mathcal{I}(\theta) \tag{4.4}$$

is locally contractive (at the c.m.l.e. of θ) throughout a neighborhood of

θ^o. Thus, for any θ' in this neighborhood of θ^o, the sequence $\{\theta_\ell\}$ defined recursively by

$$(4.5) \qquad \theta_{\ell+1} = \Phi_\varepsilon(\theta_\ell) \ , \ \theta_1 = \theta'$$

converges to the c.m.l.e. of θ.

In concluding, they discuss the computational advantages of this procedure as compared to classical numerical techniques such as Newton's method or the method of scoring. In particular, the procedure satisfies the following conditions.

(4.6) At each stage of the iteration in (4.5), the constraints on the parameters in θ are satisfied.

(4.7) The 'step size' ε depends only on n and m and not on θ^o.

(4.8) The procedure does not require the inversion, at each stage of the iteration, of a $k \times k$ matrix.

We will present a stochastic approximation analog of the iterative procedure defined by (4.5). In contrast to the classical stochastic method of scoring, our procedure satisfies the conditions (4.6) and (4.7). A step size ε will not appear explicitly in our algorithm.

Fix $\theta^o \in D$ and let $\{x_k\}_{k=1,\ldots,\infty}$ be an infinite sequence of independent samples of observations of $x(\theta^o)$. For any function g from R^n to any real vector space V, let

$$E(g) = \int_{R^n} g(x)p(\theta^o,x)dx$$

be the expectation, if it exists, of $g \circ x(\theta^{o})$. ($\circ$ denotes composition of functions). Then, by the strong law of large numbers, with probability one as $N \to \infty$, the equations (4.1) - (4.3) converge to

$$(4.9) \qquad \alpha_i' = \alpha_i E(\frac{p_i}{p})$$

$$(4.10) \qquad \mu_i' = E(x \frac{p_i}{p})/E(\frac{p_i}{p})$$

$$(4.11) \qquad \Sigma_i' = E((x-\mu_i)(x-\mu_i)^T \frac{p_i}{p})/E(\frac{p_i}{p}).$$

We denote the corresponding limiting value of the likelihood function by $\mathcal{J}$. Clearly $\mathcal{J}(\theta)$ is a continuously differentiable function of θ and $\mathcal{J}(\theta^{o}) = \theta^{o}$; also by (4.4), $\mathcal{J}_{\varepsilon}(\theta) = (1-\varepsilon)\theta + \varepsilon\,\mathcal{J}(\theta)$ is locally contractive at θ^{o}. By (2.2), $\rho(\nabla \mathcal{J}_{\varepsilon}(\theta^{o})) < 1$, implying that the eigenvalues of $\nabla \mathcal{J}(\theta^{o})$ have real parts strictly less than 1. Now define a function $d:D \to D$ by

$$(4.12) \qquad d((\alpha_i,\mu_i,\Sigma_i)_{i=1,\ldots,m}) = (\alpha_i,\alpha_i\mu_i,\alpha_i\Sigma_i)_{i=1,\ldots,m}.$$

Clearly d is a differentiable function from $D \to D$ such that d^{-1} exists and is differentiable. Define a function $\mathcal{J}':D \to D$ by $\mathcal{J}'(\theta) = d \circ \mathcal{J} \circ d^{-1}(\theta)$. By the chain rule for Frechet Derivatives,

$$(4.13) \qquad \nabla \mathcal{J}'(d(\theta^{o})) = [\nabla d(\theta^{o})][\nabla \mathcal{J}(\theta^{o})][\nabla d(\theta^{o})]^{-1}$$

hence the function

$$\mathcal{J}'_{\varepsilon}(\theta) = (1-\varepsilon)\theta + \varepsilon\mathcal{J}'(\theta) \tag{4.13}$$

is locally contractive throughout some neighborhood W of $d(\theta^\circ)$.

Define a family $\{\overline{Y}(\theta):\theta \in D\}$ of random variables with values in E_k by

$$\overline{Y}(\theta) = \varepsilon(\alpha_i-\alpha_i \frac{p_i'}{p'},\ \mu_i-\alpha_i x(\theta^\circ) \frac{p_i'}{p'},\ \Sigma_i-\alpha_i(x(\theta^\circ)-\frac{\mu_i}{\alpha_i})(x(\theta^\circ)-\frac{\mu_i}{\alpha_i})^T \frac{p_i'}{p})_{i=1,\ldots,m} \tag{4.14}$$

where $p_i' = p_i(d^{-1}(\theta),x(\theta^\circ))$ and $p' = p(d^{-1}(\theta),x(\theta^\circ))$.

Then $E(\overline{Y}(\theta)) = \theta - \mathcal{J}'(\theta)$. Therefore, the family $\{\overline{Y}(\theta):\theta \in D\}$ satisfies conditions (2.3) and (2.4).

Let $\{Y(\theta):\theta \in E_k\}$ be constructed from $\{\overline{Y}(\theta):\theta \in D\}$ as in (3.2) and let $\{a_\ell\}$ be any gain sequence. Then by (3.6), the sequence in (3.7) converges in mean square and with probability one to $d(\theta^\circ)$. Since $\{\varepsilon a_\ell\}$ is a gain sequence whenever $\{a_\ell\}$ is a gain sequence, ε need not appear explicitly in the sequence in (3.7).