The Annals of Mathematical Statistics
1970, Vol. 41, No. 1, 164–171

A MAXIMIZATION TECHNIQUE OCCURRING IN THE STATISTICAL ANALYSIS OF PROBABILISTIC FUNCTIONS OF MARKOV CHAINS

BY LEONARD E. BAUM, TED PETRIE, GEORGE SOULES, AND NORMAN WEISS

Institute for Defense Analyses, California Institute of Technology and Columbia University

1. Introduction. This paper lays bare a principle which underlies the effectiveness of an iterative technique which occurs in employing the maximum likelihood method in statistical estimation for probabilistic functions of Markov chains. We exhibit a general technique for maximizing a function $P(\lambda)$ when P belongs to a large class of probabilistically defined functions.

In the earlier note [2], an application of an inequality was made to ecology. The more general approach of this paper has allowed the use of a certain transformation and inequality in maximizing likelihood function in models for stock market behavior [3] and sunspot behavior. We also expect to apply these techniques to problems in weather prediction.

Let $A = (a_{ij})$ be an $s \times s$ stochastic matrix. Let $a = (a_i)$, $i = 1, \cdots, s$ be a probability distribution. For each $i = 1, \cdots, s$ let $f_i(y)$ be a probability density: $\int f_i(y)\,dy = 1$. For the triple $A, a, f = \{f_i\}$ we define a stochastic process $\{Y_t\}$ with density

$$(1)\quad P(A,a,f)\{Y_1 = y_1, Y_2 = y_2 \cdots Y_T = y_T\} = \sum_{i_0, i_1, \cdots, i_T = 1}^{s} a_{i_0} a_{i_0 i_1} f_{i_1}(y_1) a_{i_1 i_2} f_{i_2}(y_2) \cdots a_{i_{T-1} i_T} f_{i_T}(y_T).$$

For convenience we denote this expression by $P_{y_1 \cdots y_T}(A, a, f)$.

We call the process $Y = \{Y_t\}$ a probabilistic function of the Markov process $\{X_t\}$ determined by A. If a is chosen as a stationary distribution for the matrix A then Y will be a stationary stochastic process.

Let Λ be an open subset of Euclidean n space. Suppose that to each $\lambda \in \Lambda$, we have a smooth assignment $\lambda \to (A(\lambda), a(\lambda), f(\lambda))$. Specifically each $f_i(\lambda, \cdot)$ is a density in y and for each fixed y is a smooth function in λ. Under these assumptions, for each fixed $y_1, y_2, \cdots, y_T$, $P_{y_1 \cdots y_T}(\lambda) = P_{y_1 \cdots y_T}(A(\lambda), a(\lambda), f(\lambda))$ is a smooth function of λ. Given a fixed Y-sample $y = y_1, \cdots, y_T$ we seek a parameter value λ^0 which maximizes the likelihood $P_y(\lambda) = P_{y_1 \cdots y_T}(\lambda)$ determined from $A(\lambda)$, $a(\lambda)$, $f(\lambda)$ by (1).

One might suspect from the complicated nature of the expression (1) for $P_{y_1 \cdots y_T}(A, a, f)$ and the difficult analysis of maximizing this function of λ for very special choices of f presented in [2], [8] that a simple explicit procedure for maximization for a general f would be quite difficult; however, this is not the case. There is an extremely simple feature of this function which under mild hypothesis on f enable us to define a continuous transformation $\mathcal{T}$ mapping Λ into itself with

Received March 24, 1969.

the property that $P_{y_1 \cdots y_T}(\mathcal{T}(\lambda)) > P_{y_1 \cdots y_T}(\lambda)$ unless λ is a critical point of $P_{y_1 \cdots y_T}(\lambda)$.

Theorem 2.1 is the simple but important principle which is applied in defining the transformation $\mathcal{T}$. The convergence properties of $\mathcal{T}^k(\lambda)$ as $k \to \infty$ are discussed in Proposition 2.2 and following. In Section 3, we transform Theorem 2.1 into a form suitable for applications, viz., Theorem 3.1, and show that the method is applicable to the Normal, Poisson, Binomial and Gamma distributions but not to the Cauchy distribution. A feature of the first three examples is that the transformation $\mathcal{T}(\lambda)$ is *explicitly* presented as a function of the observations $y_1 \cdots y_T$ and the parameters. In each case, the transformation and its iterates are *practically* computable. (See [3] for a detailed illustration of the case in which $f_i(\lambda, \cdot)$ is the normal density.) In Section 4, we prove that the method is applicable to location and scale parameters for general strictly log concave density functions (Theorem 4.1) and to the coordinates of the stochastic matrix defining the Markov process.

2. An inequality. Let $X = \{1, \cdots, s\}^T$ and $x = \{x_1, \cdots, x_T\} \in X$. Then the function (1) of λ is of the form: $P(\lambda) = \sum_{x \in X} p(x, \lambda)$, with

$$p(x, \lambda) = a_{x_0}(\lambda) \prod_{t=1}^{T} a_{x_{t-1}x_t}(\lambda) f_{x_t}(\lambda, y_t).$$

More generally let X be a totally finite measure space with measure μ. Let $p(x, \lambda)$ be a positive real-valued function on $X \times \Lambda$, where Λ is a subset of Euclidean space, which is measurable and integrable in x for fixed λ. Let $P(\lambda) = \int_X p(x, \lambda)\, d\mu(x)$ and $Q(\lambda, \lambda') = \int_X p(x, \lambda) \log p(x, \lambda')\, d\mu(x)$.

THEOREM 2.1. *If $Q(\lambda, \bar{\lambda}) \geqq Q(\lambda, \lambda)$ then $P(\bar{\lambda}) \geqq P(\lambda)$. The inequality is strict unless $p(x, \lambda) = p(x, \bar{\lambda})$ almost everywhere $d\mu(x)$.*

PROOF. $\log t$ is strictly concave for $t > 0$ since $d^2/dt^2(\log t) = -t^{-2} < 0$. Hence

$$\begin{aligned}
\log P(\bar{\lambda})/P(\lambda) &= \log \textstyle\int_X p(x, \bar{\lambda})\, d\mu(x)/P(\lambda) \\
&= \log \textstyle\int_X [p(x, \lambda)\, d\mu(x)/P(\lambda)] p(x, \bar{\lambda})/p(x, \lambda) \\
&\geqq \textstyle\int_X [p(x, \lambda)\, d\mu(x)/P(\lambda)] \log [p(x, \bar{\lambda})/p(x, \lambda)] \\
&= (P(\lambda))^{-1} [Q(\lambda, \bar{\lambda}) - Q(\lambda, \lambda)] \geqq 0
\end{aligned}$$

by hypothesis. For the first inequality we have used Jensen's inequality for the probability measure $dv_\lambda(x) = p(x, \lambda)\, d\mu(x)/P(\lambda)$. This inequality is strict unless $p(x, \bar{\lambda})/p(x, \lambda)$ is constant almost everywhere $dv_\lambda(x)$ hence unless $p(x, \bar{\lambda}) = p(x, \lambda)$ almost everywhere $d\mu(x)$.

PROPOSITION 2.1. *Let $p(x, \lambda)$ be continuously differentiable in λ for almost all x. Let $\mathcal{T}(\lambda)$ be a continuous map of $\Lambda \to \Lambda$ such that for each fixed λ, $\mathcal{T}(\lambda)$ is a critical point of $Q(\lambda, \lambda')$ as a function of λ'. Then all fixed points of $\mathcal{T}$ are critical points of P and if moreover $P(\mathcal{T}(\lambda)) > P(\lambda)$ unless $\mathcal{T}(\lambda) = \lambda$, all limit points of $\mathcal{T}^n(\lambda_0)$ are fixed points of $\mathcal{T}$ for any $\lambda_0 \in \Lambda$.*

PROOF. $\partial P(\lambda)/\partial \lambda_i \big|_\lambda = \partial Q(\lambda, \lambda')/\partial \lambda_i' \big|_{\lambda' = \lambda}$. Thus λ is a critical point of P iff it is a critical point of $Q(\lambda, \lambda')$ as a function of λ'. Hence by the hypothesis on $\mathcal{T}$, if

$\mathscr{T}(\lambda) = \lambda$ then λ is a critical point of P. Suppose that $\mathscr{T}^{n_i}(\lambda_0)$ converges to λ. Then $\mathscr{T}^{n_i+1}(\lambda_0)$ converges to $\mathscr{T}(\lambda)$; so $P(\mathscr{T}^{n_i}(\lambda_0)) \leqq P(\mathscr{T}^{n_i+1}(\lambda_0)) \leqq P(\mathscr{T}^{n_i+1}(\lambda_0))$ and $P(\lambda) \leqq P(\mathscr{T}(\lambda)) \leqq P(\lambda)$. Thus $P(\lambda) = P(\mathscr{T}(\lambda))$ and since this equality can only hold if $\mathscr{T}(\lambda) = \lambda$, this shows that: all limit points of the sequence $\{\mathscr{T}^n(\lambda_0)\}$ are fixed points of $\mathscr{T}$.

3. The transformation $\mathscr{T}$ and applications. The remainder of this paper will consist of applications of Theorem 2.1 in various situations of interest. To a given $P(\lambda) = \int_X p(x, \lambda)\, d\mu(x)$ there is naturally associated the auxiliary function: $Q(\lambda, \lambda') = \int_X p(x, \lambda) \log p(x, \lambda')\, d\mu(x)$. If we define $\mathscr{T}: \Lambda \to \Lambda$ by $\mathscr{T}(\lambda) = \{\bar{\lambda} \in \Lambda \mid \max_{\lambda' \in \Lambda} Q(\lambda, \lambda') = Q(\lambda, \bar{\lambda})\}$ then $Q(\lambda, \mathscr{T}(\lambda)) \geqq Q(\lambda, \lambda)$ so Theorem 2.1 guarantees $P(\mathscr{T}(\lambda)) \geqq P(\lambda)$. Under natural hypotheses on P we will see that

(i) $\mathscr{T}$ exists and is single valued;
(ii) $\mathscr{T}$ is continuous;
(iii) $\mathscr{T}$ is effectively computable;
(iv) $P(\mathscr{T}(\lambda)) > P(\lambda)$ save when λ is a critical point of P in which case λ is a fixed point of $\mathscr{T}$.

If P has finitely many critical points, for example, it follows from (iv) that for each λ not a critical point, $\mathscr{T}^n(\lambda)$ approaches a critical point of P which will be a local maximum (save for starting points λ in a lower dimensional manifold). We cannot rule out limit cycle behavior of the iterates $\mathscr{T}^n$ without some hypothesis on the critical point set of P as above.

Let $\Lambda = R^s$. For almost all x let $\log p(x, \lambda) = \sum_{i=1}^s \log p_i(x, \lambda_i)$ where for each i and almost all x $\log p_i(x, \lambda_i)$ is strictly concave in λ_i and $\lim_{|\lambda_i| \to \infty} \log p_i(x, \lambda_i) = -\infty$. Define $Q_i(\lambda, \lambda_i')$ by

$$Q_i(\lambda, \lambda_i') = \int_X p(x, \lambda) \log p_i(x, \lambda_i')\, d\mu(x).$$

Then for λ fixed, $Q_i(\lambda, \lambda_i')$ is a strictly concave function of λ_i' which $\to -\infty$ at $\pm\infty$ and hence has a unique global maximum $\bar{\lambda}_i$, which is a critical point of $Q_i(\lambda, \lambda_i')$. Define $\mathscr{T}: \lambda \to \bar{\lambda} = \{\bar{\lambda}_i\}$.

THEOREM 3.1. *Under the above assumptions, for all $\lambda \in \Lambda$ $P(\mathscr{T}(\lambda)) \geqq P(\lambda)$ with equality if and only if λ is a critical point of P or equivalently is a fixed point of $\mathscr{T}$.*

PROOF.

$$Q(\lambda, \bar{\lambda}) = \sum_{i=1}^s Q_i(\lambda, \bar{\lambda}_i) \geqq \sum_{i=1}^s Q_i(\lambda, \lambda_i) = Q(\lambda, \lambda)$$

so Theorem 2.1 implies $P(\bar{\lambda}) \geqq P(\lambda)$. Since for each i $Q_i(\lambda, \lambda_i)$ is strictly concave in λ_i the inequality $Q(\lambda, \bar{\lambda}) \geqq Q(\lambda, \lambda)$ (and hence the inequality $P(\bar{\lambda}) \geqq P(\lambda)$) will be strict unless $\bar{\lambda}_i = \lambda_i$ for each i. This latter is the case iff $\partial Q_i / \partial \lambda_i' \big|_{\lambda' = \lambda} = 0, i = 1, \cdots, s$; i.e., $\partial P / \partial \lambda_i \big|_{\lambda} = 0, i = 1, \cdots, s$.

Here is a sample of the inequalities that can be analyzed via our technique.

PROPOSITION 3.1. *If y_t, $t = 1, \cdots, T$ is a sequence of real numbers and $c_x = c_{x_1 \cdots x_T}$ then*

$$P(\lambda) = \sum_x c_x \prod_{t=1}^{T} \exp(-|\lambda_{x_t} - y_t|)$$

assumes its global maximum at a point $\bar{\lambda}$ where each $\bar{\lambda}_i$ is some y_{t_i}.

PROOF. We have $Q(\lambda, \lambda') = \Sigma_i Q_i(\lambda, \lambda_i')$ where $Q_i(\lambda, \lambda_i') = -\Sigma \gamma_t(i)|\lambda_i' - y_t|$ with $\gamma_t(i) \geqq 0$. Then $Q_i(\lambda, \lambda_i') \to -\infty$ as $|\lambda_i| \to \infty$.

$$\frac{\partial}{\partial \lambda_i'} Q_i(\lambda, \lambda_i') = -\sum \gamma_i(t) \operatorname{sgn}(\lambda_i' - y_t), \qquad \lambda_i' \neq \text{any } y_t,$$

is monotonic decreasing from ∞ at $\lambda_i' = -\infty$ to $-\infty$ at $\lambda_i' = \infty$, constant in intervals between y_t, and changes sign at some $y_t = \bar{\lambda}_i$. We then have $Q_i(\lambda, \bar{\lambda}_i) = \max_{\lambda_i'} Q_i(\lambda, \lambda_i') \geqq Q_i(\lambda, \lambda_i)$. By Theorem 2.1 $P(\bar{\lambda}) \geqq P(\lambda)$. Now if λ_i is any parameter which appears in $P(\lambda)$ with a non-zero coefficient c_x then $P(\lambda) \to -\infty$ as $|\lambda_i| \to \infty$ so that $P(\lambda)$ assumes its maximum at some point $\hat{\lambda}$. But $\mathscr{T}(\hat{\lambda}) = \bar{\lambda}$ is a point with each $\bar{\lambda}_i$ a y_{t_i} and $P(\bar{\lambda}) \geqq P(\hat{\lambda}) = \max P(\lambda)$.

Applications of Theorem 3.1 *to familiar probability distributions.* Let $X = \{1, \cdots, s\}^T$ and $x = \{x_1, \cdots, x_T\} \in X$. For $i = 1, \cdots, s$ let $f_i(r)$ be a strictly log concave probability density function. For each i we obtain a one parameter family of density functions by introducing λ_i as a location parameter: $f_i(r - \lambda_i)$. For a given fixed real sequence $y_1, \cdots, y_T$ let $p(x, \lambda) = \prod_{t=1}^{T} f_{x_t}(y_t - \lambda_{x_t})$ and $P(\lambda) = \sum_{x \in X} p(x, \lambda)\mu(x)$, $\mu(x) > 0$. Then $\log p(x, \lambda) = \sum_{i=1}^{s} \log p_i(x, \lambda_i)$ where

$$\log p_i(x, \lambda_i) = \sum_{S_{xi}} \log f_i(y_t - \lambda_i)$$

with S_{xi} denoting the set of t such that $x_t = i$. Also $\log f_i(y_t - \lambda_i)$ is strictly concave in λ_i for each y_t and $\to -\infty$ as $|\lambda| \to \infty$ since f_i is a probability density. Hence, Theorem 3.1 is applicable.

Theorem 3.1 applies straight away to these three examples: the density $f_i(\lambda, y) = c \exp(-k|\lambda_i - y|^{\varepsilon})$, $\varepsilon > 1$, $k > 0$ and $-\infty < \lambda_i < \infty$, the Poisson distribution on the non-negative integers $f_i(\lambda, y) = e^{-\lambda_i} \lambda_i^{y}/y!$ and $0 \leqq \lambda_i < \infty$ and the binomial distribution on the integers $0, 1, \cdots, N$, $f_i(\lambda, y) = \binom{N}{y} \lambda_i^{y} (1 - \lambda_i)^{N-y}$ and $0 \leqq \lambda_i \leqq 1$. In each case the method is the same. We determine the transformation $\mathscr{T}$ by solving explicitly for the unique λ_i' zero of $\partial/\partial \lambda_i' Q_i(\lambda, \lambda') = 0$. For example, in the first case $\mathscr{T}(\lambda)_i = \bar{\lambda}_i$ is the unique zero of

$$\sum_{x \in X} \prod_{t=1}^{T} \exp(-|\lambda_{x_t} - y_t|^{\varepsilon}) \mu(x) \sum_{S_{xi}} |\lambda_i' - y_t|^{\varepsilon - 1}.$$

For $\varepsilon = 2$ we have the important case of the normal density.

In each of these three cases $\mathscr{T}(\lambda)$ has the explicit form

$$\mathscr{T}(\lambda)_i = \left[\sum_{x \in X} \mu(x) \prod_{t=1}^{T} v_i(x) f_{x_t}(\lambda_{x_t}, y_t)\right] \cdot \left[\sum_{x \in X} \mu(x) \prod_{t=1}^{T} n_i(x) f_{x_t}(\lambda_{x_t}, y_t)\right]^{-1}$$

where $v_i(x) = \sum_{S_{xi}} y_t$ and $n_i(x)$ is the number of elements in S_{xi}. In the most important case where $\mu(x) = a_{x_1} a_{x_1 x_2} \cdots a_{x_{T-1} x_T}$,

$$\mathscr{T}(\lambda)_i = \left[\sum_{t=1}^{T} \gamma_t(i) y_t\right] \left[\sum_{t=1}^{T} \gamma_t(i)\right]^{-1}$$

where $\gamma_t(i)$ may be obtained through backward and forward inductive computations in t. We have $\gamma_t(i) = \alpha_t(i)\beta_t(i)$ where

$$\alpha_t(j) = \sum_{i=1}^{s} \alpha_{t-1}(i) a_{ij} b_j(\lambda_j, y_t),$$

$j = 1, \cdots, s, t = 2, \cdots, T,$

$$\beta_t(i) = \sum_{j=1}^{s} \beta_{t+1}(j) a_{ij} b_j(\lambda_j, y_{t+1}),$$

$i = 1, \cdots, s, t = T-1, \cdots, 1$. $\gamma_t(i)$ has a probabilistic interpretation as the λ likelihood of the event $\{Y_1 = y_1, Y_2 = y_2, \cdots, Y_T = y_T, X_t = i\}$. See [6].

The analysis of the case of the gamma density is somewhat more complicated. In particular we do not have an explicit expression for $\mathscr{T}(\lambda)$. Nonetheless Theorem 3.1 still applies. Here is a discussion.

The expression $f_{\alpha,\nu}(r) = (\Gamma(\nu))^{-1}\alpha^{\nu} r^{\nu-1} e^{-\alpha r}$, $\alpha > 0$, $\nu > 0$ defines a two parameter family of densities for $r \geqq 0$. Here $\Gamma(\nu) = \int_0^{\infty} t^{\nu-1} e^{-t}\, dt$ is the gamma function. Let $\{\alpha_i, \nu_i : i = 1, \cdots, s\}$ be unknown parameters and let

$$P(\alpha, \nu) = \sum_x \mu(x) \prod_{t=1}^{T} f_{\alpha_{x_t}, \nu_{x_t}}(y_t) = \sum_x p(x, \alpha, \nu)\mu(x) \qquad \text{and}$$

$$Q(\alpha, \nu, \alpha', \nu') = \sum_x p(x, \alpha, \nu)\mu(x) \sum_{i=1}^{s} \sum_{S_{xi}} \log f_{\alpha_i' \nu_i'}(y_t),$$

$$= \sum_{i=1}^{s} Q_i(\alpha, \nu, \alpha_i', \nu_i').$$

Let α, ν be fixed throughout. $Q_i(\alpha, \nu, \alpha_i', \nu_i') \to -\infty$ as $\alpha_i', \nu_i' \to 0$ or ∞. We write

$$Q_i(\alpha, \nu, \alpha_i', \nu_i') = \sum_{t=1}^{T} \gamma_t(i) \log f_{\alpha_i' \nu_i'}(y_t).$$

For α, ν fixed, one proves using the infiite expansion for $\log \Gamma(\nu)$ [1] page 16, that the Hessian of the Q_i function with respect to the prime coordinates is negative definite. (The Hessian of a function $F(x_1, \cdots, x_n)$ is the matrix: $(\partial^2 F/\partial x_i \partial x_j)$.) Hence $Q_i(\alpha, \nu, \alpha_i', \nu_i')$ is strictly concave as a function of the two variables α_i', ν_i'. Since $Q_i(\alpha, \nu, \alpha_i', \nu_i') \to -\infty$ as ν_i', α_i' go to the boundary, we conclude that $Q_i(\alpha, \nu, \alpha_i', \nu_i')$ has a unique critical point $\bar{\alpha}_i, \bar{\nu}_i$ which is a global maximum and is obtained as a solution of $\partial Q_i/\partial \alpha_i' = 0$, $\partial Q_i/\partial \nu_i' = 0$. As usual $P(\{\bar{\alpha}_i, \bar{\nu}_i\}) > P(\alpha, \nu)$ save at critical points (α, ν) of P.

The Cauchy density yields a counterexample which shows the difficulties in applying Theorem 3.1 to a wider class of density functions. If $f(r) = \pi^{-1}(1+r^2)^{-1}$ then $d^2 f/dr^2 = -2\pi^{-1}(1-3r^2)(1+r^2)^{-3}$ is only concave for $3r^2 \leqq 1$ so we cannot proceed as in the preceding examples for the one parameter family of Cauchy densities with location parameter: $f(\lambda, r) = \pi^{-1}(1+(r-\lambda)^2)^{-1}$. On the contrary it is easy to see that the function

$$\frac{\partial}{\partial \lambda_i'} Q_i(\lambda, \lambda_i') = \frac{2}{\pi} \sum_{t=1}^{T} \gamma_t(i) \frac{y_t - \lambda_i'}{1+(y_t-\lambda_i')^2}$$

need not have a unique zero for suitable $\{y_t\}$. For λ fixed, one of those critical points of $Q_i(\lambda, \lambda_i')$ will provide a global maximum of $Q_i(\lambda, \lambda_i')$ and hence a suitable $\bar{\lambda}_i$ but some further procedure is needed for finding such a global maximum.

4. Some general situations where the transformation $\mathcal{T}$ is applicable. With a given probability density function $f(u)$ we can obtain a two parameter family $f_{m,\sigma}(u)$ by introducing location and scale parameters m and $\sigma > 0$ respectively: $f_{m,\sigma}(u) = \sigma^{-1} f((u-m)/\sigma)$. Given density functions $f_1, \cdots, f_s$ consider

$$P(m, \sigma) = P(\{m_i, \sigma_i\}) = \sum_{x \in X} \mu(x) \prod_{t=1}^{T} \frac{1}{\sigma_{x_t}} f_{x_t}\left(\frac{y_t - m_{x_t}}{\sigma_{x_t}}\right)$$

where $\mu(x) \geqq 0$, and again $X = \{1, \cdots, s\}^T$. Let $v_x(m, \sigma)$ be the summand of P; then the auxiliary functions Q_i are

$$Q_i(m, \sigma, m_i', \sigma_i') = \sum_x v_x(m, \sigma) \sum_{s_{xi}} [\log f_i((y_t - m_i')/\sigma_i') - \log \sigma_i']$$

and $Q(m, \sigma, m', \sigma') = \sum_{i=1}^{s} Q_i(m, \sigma, m_i', \sigma_i')$.

The next theorem shows that our procedure carries over to such functions P if we require (essentially) only the natural hypothesis that each f_i is strictly log concave.

THEOREM 4.1. *For $i = 1, \cdots, s$ let $f_i(u)$ be a strictly* log *concave density function with its unique local maximum at $u = 0$, and assume for every state i the following non-pathology condition holds: $\exists$ times $t_n = t_n(i)$, $n = 1, 2$ such that $y_{t_1} \neq y_{t_2}$ and $\sum_{x: x_{t_n} = i} v_x(m, \sigma) > 0$, $n = 1, 2$. Then for fixed m and σ, the system of equation $\partial Q/\partial m_i' = 0$, $\partial Q/\partial \sigma_i' = 0$ has a unique solution $\{\bar{m}_i', \bar{\sigma}_i'\}$ which is the global maximum of $Q(m, \sigma, m', \sigma')$ as a function of m' and σ'. Defining $\mathcal{T}(m, \sigma) = \{\bar{m}_i', \bar{\sigma}_i'\}$ we then have $P(\mathcal{T}(m, \sigma)) \geqq P(m, \sigma)$ with strict inequality unless (m, σ) is a critical point of P or equivalently $\mathcal{T}(m, \sigma) = (m, \sigma)$.*

PROOF. Since $\partial Q/\partial m_i' = \partial Q_i/\partial m_i'$ and $\partial Q/\partial \sigma_i' = \partial Q_i/\partial \sigma_i'$, we may consider the functions Q_i independently. For notational convenience we suppress the dependence of Q_i on $m = \{m_i\}$ and $\sigma = \{\sigma_i\}$, then delete the subscripts i and the primes as well. In such notation the theorem is proved by showing $Q(m, \sigma) \to -\infty$ as (m, σ) approaches the boundary $\delta\Omega$ of $\Omega(|m| < \infty, 0 < \sigma < \infty)$, and that Q has a unique critical point in Ω which is a local and hence global maximum. Now

$$Q(m, \sigma) = \sum_x v_x \cdot \sum_{s_{xi}} [\log f(z_t) - \log \sigma]$$

where $z_t = (y_t - m)/\sigma$. If $\gamma_t = \gamma_t(i) = \sum_{s_{xi}} v_x$, then

$$Q(m, \sigma) = \sum_{t=1}^{T} \gamma_t [\log f(z_t) - \log \sigma].$$

Now $Q(m, \sigma) \to -\infty$ iff $\prod_{t=1}^{T} f(z_t)^{\gamma_t}/\sigma^{\Sigma\gamma_t} \to 0$, and that this is indeed the case as $(m, \sigma) \to \delta\Omega$ rests on two facts. Namely, a strictly log concave density function $f(u)$ is always less than $e^{-a|u|}$ if $a > 0$ is sufficiently small and $|u|$ is sufficiently large, so that $\lim_{|u| \to \infty} u^n f(u) = 0$ for any n; also it is precisely the above non-pathology condition which guarantees for every m the existence of a $t = t(m)$ such that $z_t \neq 0$ and $\gamma_t > 0$, so that as $\sigma \to 0$ the desired behavior of Q occurs. The lengthy but straightforward details are omitted.

We finish by showing each critical point of Q in Ω is a local maximum, for it is then obvious topologically (and follows rigorously from Morse theory) that Q has

only one local maximum, which is then a unique global maximum. Since this is true for every i, an application of Theorem 2.1 completes the proof. Let $g = \log f$. Then

$$\frac{\partial^2 Q}{\partial m^2} = \frac{1}{\sigma^2}\sum_t \gamma_t g''(z_t),$$

$$\frac{\partial^2 Q}{\partial \sigma^2} = \frac{-1}{\sigma}\frac{\partial Q}{\partial \sigma} + \frac{1}{\sigma^2}\sum_t \gamma_t z_t^2 g''(z_t) + \frac{1}{\sigma^2}\sum_t \gamma_t z_t g'(z_t),$$

and

$$\frac{\partial^2 Q}{\partial m\,\partial \sigma} = \frac{-1}{\sigma}\frac{\partial Q}{\partial m} + \frac{1}{\sigma^2}\sum_t \gamma_t z_t g''(z_t).$$

Now $\partial^2 Q/\partial m^2 < 0$ since $g'' < 0$. At a critical point $\partial Q/\partial m = \partial Q/\partial \sigma = 0$, and the determinant $\det(H)$ of the Hessian H of $Q(m, \sigma)$ is given by

$$\sigma^4 \det(H) = \sum_t \gamma_t g''(z_t) \sum_t \gamma_t z_t^2 g''(z_t) - \left(\sum_t \gamma_t z_t g''(z_t)\right)^2 + \sum_t \gamma_t g''(z_t) \sum_t \gamma_t z_t g'(z_t).$$

Now $ug'(u) < 0$ since $g'(0) = 0$, so the third term above is positive. Since the function u^2 is convex, the sum of the first two terms is also positive, which proves H is negative definite at a critical point and the proof of the theorem is complete.

It is to be noted that H is not in general a negative definite form.

COROLLARY 4.1. *Let* $f(u) = \exp(-c|u|^\alpha) \alpha > 1$. *Then* $\log f(u) = -c|u|^\alpha$ *is strictly concave so the previous theorem applies.* The special case $\alpha = 2$, the normal distribution is of great utility. The unique critical point $\{\bar{m}_i, \bar{\sigma}_i\}$ is then simply computed from:

$$\bar{m}_i = \left[\sum_t \gamma_t(i) y_t\right]\left[\sum_t \gamma_t(i)\right]^{-1}$$

$$\bar{\sigma}_i^2 = \left[\sum_t \gamma_t(i) y_t^2\right]\left[\sum_t \gamma_t(i)\right]^{-1} - \bar{m}_i^2.$$

Since $\gamma_t(i)$ is proportional to the *a posteriori* probability of being in the state i at time t, the reestimation $\{\bar{m}_i, \bar{\sigma}_i\}$ has an obvious probabilistic interpretation which led to its original use.

Application. We finally carry out the transformation $\mathcal{T}$ on all the coordinates mentioned in Section 1; i.e., we will include the stochastic matrix coordinates $A = (a_{ij})$ and starting probabilities $a = (a_i)$.

We consider the situation in which each $y_t \in (1, 2, \cdots, m)$ and each f_j is a probability distribution in $(1, 2, \cdots, m)$ rather than a probability density on the reals since this was the case originally dealt with [2], [6]. The following *mutatis mutandis* applies to the case where each f_j is a centered log concave density function as in Theorem 4.1.

Let $P(a, A, f) = \sum_{x \in X} p(x, a, A, f)$ where

$$p(x, a, A, f) = a_{x_0} \prod_{t=1}^{T} a_{x_{t-1}x_t} f_{x_t}(y_t)$$

and $y_t \in \{1, 2, \cdots, m\}$. Then

$$Q(a, A, f, a', A', f') = \sum_x p(x, a, A, f)\left\{\log a'_{x_0} + \sum_t \log a'_{x_{t-1}x_t} + \sum_t \log f'_{x_t}(y_t)\right\}$$

is an extremely simple function of the variables a', A' and f'. In fact an elementary computation shows that there is a unique $\bar{a}'$, $\bar{A}'$, $\bar{f}'$ which maximizes Q as a function of the primed coordinates. We find:

$$\bar{a}_i' = [\textstyle\sum_{x_0=i} p(x,a,A,f)][\sum_x p(x,a,A,f)]^{-1}$$

$$\bar{a}_{ij}' = [\textstyle\sum_x p(x,a,A,f) \sum_{x_{t-1}=i,x_t=j} 1][\sum_x p(x,a,A,f) \sum_{x_{t-1}=i,t\geqq 1} 1]^{-1}$$

$$f_j'(k) = [\textstyle\sum_x p(x,a,A,f) \sum_{x_t=j,y_t=k} 1][\sum_x p(x,a,A,f) \sum_{x_t=j} 1]^{-1}$$

Hence, the transformation $\mathcal{T}: \{a, A, f\} \to \{\bar{a}, \bar{A}, \bar{f}\}$ increases the function $P(a, A, f)$. Strict inequality holds save at fixed points of $\mathcal{T}$ or critical points of P suitably interpreted.

REFERENCES

[1] ARTIN, EMIL (1964). *The Gamma Function.* Holt, Rinehart and Winston, New York.

[2] BAUM, LEONARD E. and EAGON, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73** 360–363.

[3] BAUM, LEONARD E.; GAINES, STOCKTON; PETRIE, TED; and SIMONS, JAMES. Probabilistic models for stock market behavior. To appear.

[4] BAUM, LEONARD E. and PETRIE, TED (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **37** 1554–1563.

[5] BAUM, LEONARD E. and SELL, GEORGE R. Growth transformation for functions on manifolds. To appear *Pacific J. Math.*

[6] BAUM, LEONARD E. and WELCH, LLOYD. A statistical estimation procedure for probabilistic functions of finite Markov processes. Submitted for publication *Proc. Nat. Acad. Sci. USA.*

[7] PETRIE, TED. Theory of probabilistic functions of finite state Markov chains. To appear.

[8] ROTHAUS, OSCAR. An inequality. Unpublished manuscript.