An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes

LEONARD E. BAUM

Institute for Defense Analyses, Princeton, New Jersey

We say that $\{Y_t\}$ is a probabilistic function of the Markov process $\{X_t\}$ if

$$P(X_{t+1} = j \mid X_t = i, \quad X_{t-1}, \ldots, Y_t, \ldots) = a_{ij}, \qquad i, j = 1, \ldots, s;$$

$$P(Y_{t+1} = k \mid X_{t+1} = j, \quad X_t = i, \quad X_{t-1}, \ldots, Y_t, Y_{t-1}, \ldots) = b_{ij}(k),$$

$$i, j = 1, \ldots, s, \quad k = 1, \ldots, r.$$

We assume that $\{a_{ij}\}$, $\{b_{ij}(k)\}$ are unknown and restricted to be in the manifold M

$$a_{ij} \geqslant 0, \qquad \sum_{j=1}^{s} a_{ij} = 1, \qquad i = 1, \ldots, s,$$

$$b_{ij}(k) \geqslant 0, \qquad \sum_{k=1}^{r} b_{ij}(k) = 1, \qquad i, j = 1, \ldots, s.$$

We see a Y sample $\{Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T\}$ but not an X sample and desire to estimate $\{a_{ij}, b_{ij}(k)\}$.

We would like to choose maximum likelihood parameter values, i.e., $\{a_{ij}, b_{ij}(k)\}$ which maximize the probability of the observed sample $\{y_t\}$

$$P_{\{y_t\}}(\{a_{ij}, b_{ij}(k)\}) = P(\{a_{ij}, b_{ij}(k)\})$$

$$= \sum_{i_0, i_1, \ldots, i_T = 1}^{s} a_{i_0} a_{i_0 i_1} b_{i_0 i_1}(y_1)\, a_{i_1 i_2} b_{i_1 i_2}(y_2) \cdots a_{i_{T-1} i_T} b_{i_T}(y_T) \quad (1)$$

where a_i are initial probabilities for the Markov process. For this purpose we define a transformation $\tau\{a_{ij}, b_{ij}(k)\} = \{\bar{a}_{ij}, \bar{b}_{ij}(k)\}$ of M into itself where

$$\begin{aligned}
\bar{a}_{ij} &= \frac{\sum_t P(X_t = i, \quad X_{t+1} = j \mid \{y_t\}, \quad \{a_{ij}, b_{ij}(k)\})}{\sum_t P(X_t = i \mid \{y_t\}, \quad \{a_{ij}, b_{ij}(k)\})} \\
&= \frac{\sum_t \alpha_t(i)\, \beta_{t+1}(j)\, a_{ij} b_{ij}(y_{t+1})}{\sum_t \alpha_t(i)\, \beta_t(i)} \\
&= \frac{a_{ij}\, \partial P/\partial a_{ij}}{\sum_j a_{ij}\, \partial P/\partial a_{ij}}, \qquad \text{(2a)}
\end{aligned}$$

$$\begin{aligned}
\bar{b}_{ij}(k) &= \frac{\sum_{Y_t = k} P(X_t = i, \quad X_{t+1} = j \mid \{y_t\}, \quad \{a_{ij}, b_{ij}(k)\})}{\sum_t P(X_t = i, \quad X_{t+1} = j \mid \{y_t\}, \quad \{a_{ij}, b_{ij}(k)\})} \\
&= \frac{\sum_{Y_t = k} \alpha_t(i)\, \beta_{t+1}(j)\, a_{ij} b_{ij}(y_{t+1})}{\sum_t \alpha_t(i)\, \beta_{t+1}(j)\, a_{ij} b_{ij}(y_{t+1})} \\
&= \frac{b_{ij}(k)\, \partial P/\partial b_{ij}(k)}{\sum_k b_{ij}(k)\, \partial P/\partial b_{ij}(k)}. \qquad \text{(2b)}
\end{aligned}$$

The second of the equivalent forms in Eqs. (2) contains quantities $\alpha_t(i)$, $\beta_t(j)$ which are defined inductively forwards and backwards, respectively, in t by

$$\alpha_{t+1}(j) = \sum_{i=1}^{s} \alpha_t(i)\, a_{ij} b_{ij}(y_{t+1}), \qquad j = 1,..., s, \quad t = 0, 1,..., T-1,$$

$$\beta_t(i) = \sum_{j=1}^{s} \beta_{t+1}(j)\, a_{ij} b_{ij}(y_{t+1}), \qquad i = 1,..., s, \quad t = T-1, T-2,..., 0. \qquad (3)$$

Note that the $\alpha_t(i)$, $\beta_t(i)$, $i = 1,..., s$, $t = 0,..., T$ can all be computed with $4s^2T$ multiplications. Hence

$$P(\{a_{ij}, b_{ij}(k)\}) = \sum_{i=1}^{s} \alpha_t(i)\, \beta_t(i)$$

(identically in t) can be computed with $4s^2T$ multiplications rather than the $2Ts^{T+1}$ multiplications indicated in the defining formula (1). Similarly, the partial derivatives of P needed for defining the image in (2) are computed from the α's and β's with a work factor linear in T, not exponential in T.

There are three ways of rationalizing the use of this transformation, defined in (2):

(a) Bayesian *a posteriori* reestimation suggested the transformation τ originally and is embodied in the first expressions for $\bar{a}_{ij}$ and $\bar{b}_{ij}(k)$.

(b) An attempt to solve the likelihood equation obtained by setting the partial derivatives of P with respect to the a_{ij} and $b_{ij}(k) = 0$, taking due account of the restraints, is indicated in the third expressions for $\bar{a}_{ij}$ and $\bar{b}_{ij}(k)$ since the likelihood equations can be put into the form

$$a_{ij} = \frac{a_{ij}\, \partial P/\partial a_{ij}}{\sum_j a_{ij}\, \partial P/\partial a_{ij}},$$

$$b_{ij}(k) = \frac{b_{ij}(k)\, \partial P/\partial b_{ij}(k)}{\sum_k b_{ij}(k)\, \partial P/\partial b_{ij}(k)}.$$

THEOREM 1. [*1*] *$P(\tau\{a_{ij}, b_{ij}(k)\}) > P(\{a_{ij}, b_{ij}(k)\})$ unless $\tau\{a_{ij}, b_{ij}(k)\} = \{a_{ij}, b_{ij}(k)\}$ which is true if and only if $\{a_{ij}, b_{ij}(k)\}$ is a critical point of P, i.e., a solution of the likelihood equations.*

Note that τ depends only on the first derivatives of P. Now if one moves a sufficiently small distance in the gradient direction, one is guaranteed to increase P, but how small a distance depends on the second partials. It is somewhat unexpected to find that it is possible to specify a point at which P increases, without any mention of higher derivatives.

Eagon and the author [*1*] originally observed that $P(\{a_{ij}, b_{ij}(k)\})$ is a homogeneous polynomial of degree $2T+1$ in a_i, a_{ij}, $b_{ij}(k)$ and obtained the result as an application of the following theorem.

THEOREM 2. [*1*] *Let*

$$P(z_1,..., z_n) = \sum_{\mu_1, \mu_2, ..., \mu_n} c_{\mu_1, \mu_2, ..., \mu_n} z_1^{\mu_1} z_2^{\mu_2} \cdots z_n^{\mu_n} \quad \text{where} \quad c_{\mu_1, \mu_2, ..., \mu_n} \geqslant 0$$

and $\mu_1 + \cdots + \mu_n = d$. Then

$$\tau : \{z_i\} \to \left\{ \frac{z_i\, \partial P/\partial z_i}{\sum_j z_j\, \partial P/\partial z_j} \right\}$$

maps $D : z_i \geqslant 0$, $\sum z_i = 1$ into itself and satisfies $P(\tau\{z_i\}) \geqslant P\{z_i\}$. In fact, strict inequality holds unless $\{z_i\}$ is a critical point of P in D.

For the proof, the partial derivatives were evaluated as

$$z_i\, \partial P/\partial z_i = \sum_{\mu_1, \mu_2, ..., \mu_n} c_{\mu_1, \mu_2, ..., \mu_n} \mu_i z_1^{\mu_1} z_2^{\mu_t} \cdots z_n^{\mu_n}$$

and substituted for the variables z_i in the expression for P. An elementary though very tricky juggling of the inequality between geometric and arithmetic means and Hölder's inequality then led to the desired result through a route which cast no light on what was actually happening. The author believes the following derivation due to Baum *et al.* [2], which greatly generalizes the applicability of the transformation τ, lays bare the essence of the situation.

We adopt a simplified notation. We write

$$P(\lambda) = \sum_{x \in X} p(x, \lambda)$$

where λ specifies an $[s - 1 + s(s-1) + s^2(r-1)]$-dimensional parameter point $\{a_i, a_{ij}, b_{ij}(k)\}$ in $[s + s^2 + s^2 r]$-dimensional space and $x = \{x_{i_0}, x_{i_1}, ..., x_{i_T}\}$ is a sequence of states of the unseen Markov process. The summation is over X, the space of all possible $T + 1$ long sequences of states, and $p(x, \lambda) = a_{i_0} a_{i_0 i_1} b_{i_0 i_1}(y_1) \cdots a_{i_{T-1} i_T} b_{i_{T-1} i_T}(y_T)$ is the probability of the Markov process following that sequence of states and producing the observed $\{y_t\}$ sample for the parameter values $\{a_i, a_{ij}, b_{ij}(k)\}$. More generally, we write

$$P(\lambda) = \int_{x \in X} p(x, \lambda)\, d\mu(\lambda)$$

where μ is a finite nonnegative measure and $p(x, \lambda)$ is positive a.e. with respect to μ. In the main application of interest μ is a counting measure: $\mu(x) = 1$ for each of the s^{T+1} points x.

We wish to define a transformation τ on the λ-space and show that $P(\tau(\lambda)) > P(\lambda)$. For this purpose we define an auxiliary function of two variables

$$Q(\lambda, \lambda') = \int_{x \in X} p(x, \lambda) \log p(x, \lambda')\, d\mu(x).$$

THEOREM 3. [2] *If* $Q(\lambda, \bar{\lambda}) \geqslant Q(\lambda, \lambda)$, *then* $P(\bar{\lambda}) > P(\lambda)$ *unless* $p(x, \bar{\lambda}) = p(x, \lambda)$ *a.e. with respect to* μ.

Proof. We shall apply Jensen's inequality to the concave function $\log x$. We wish to prove $P(\bar{\lambda}) \geqslant P(\lambda)$ or, equivalently, $\log[P(\bar{\lambda})/P(\lambda)] \geqslant 0$. Now

$$\begin{aligned}
\log \frac{P(\bar{\lambda})}{P(\lambda)} &= \log \left[\frac{1}{P(\lambda)} \int_X p(x, \bar{\lambda})\, d\mu(x)\right] \\
&= \log \int_X \left[\frac{p(x, \lambda)\, d\mu(x)}{P(\lambda)}\right] \frac{p(x, \bar{\lambda})}{p(x, \lambda)} \\
&\geqslant \int_X \left[\frac{p(x, \lambda)\, d\mu(x)}{P(\lambda)}\right] \log \frac{p(x, \bar{\lambda})}{p(x, \lambda)} \\
&= \frac{1}{P(\lambda)} [Q(\lambda, \bar{\lambda}) - Q(\lambda, \lambda)] \geqslant 0
\end{aligned}$$

by hypothesis. Jensen's inequality is applicable to the first inequality since $p(x, \lambda)\, d\mu(x)/P(x)$ is a nonnegative measure with total mass 1. Since log is strictly concave ($\log'' < 0$), equality can hold only if $p(x, \bar{\lambda})/p(x, \lambda)$ is constant a.e. with respect to $d\mu(x)$.

We now have a way of increasing $P(\lambda)$. For each λ we need only find a $\bar{\lambda}$ with $Q(\lambda, \bar{\lambda}) \geqslant Q(\lambda, \lambda)$. This may not seem any easier than directly finding a $\bar{\lambda}$ with $P(\bar{\lambda}) \geqslant P(\lambda)$. However the author shall show that under natural assumptions and in particular in the cases of interest:

(a) For fixed λ, $Q(\lambda, \lambda')$ assumes its global maximum as a function of λ' at a unique point $\tau(\lambda)$.
(b) $\tau(\lambda)$ is continuous.
(c) $\tau(\lambda)$ is effectively computable.
(d) $P(\tau(\lambda)) \geqslant P(\lambda)$ which follows from Theorem 3 and the definition of $\tau(\lambda)$ since $\lambda' = \lambda$ is one of the competitors for the global maximum of $Q(\lambda, \lambda')$ as a function of λ'.

We apply Theorem 3 to the principle case of interest. Letting $\{a, A, B\}$ denote $\{a_i, a_{ij}, b_{ij}(k)\}$, we have

$$P(a, A, B) = \sum_x p(x, a, A, B)$$

where

$$p(x, a, A, B) = a_{x_0} \prod_{t=0}^{T-1} a_{x_t x_{t+1}} \prod_{t=0}^{T-1} b_{x_t x_{t+1}}(y_{t+1}).$$

Also

$$\begin{aligned}
&Q(a, A, B; a', A', B') \\
&\quad = \sum_{x \in X} p(x, a, A, B) \left\{\log a'_{x_0} + \sum_t \log a'_{x_t x_{t+1}} + \sum_t \log b'_{x_t x_{t+1}}(y_{t+1})\right\}.
\end{aligned}$$

For fixed a, A, B we seek to maximize Q as a function of a', A', B'. We observe that for a, A, B fixed, Q is a sum of three functions—one involving only $\{a_i'\}$, the second involving only $\{a_{ij}'\}$, and the third involving only $\{b_{ij}'(k)\}$ which can be maximized separately.

We consider the second of these. Observe that

$$\sum_{x \in X} p(x, a, A, B) \sum_t \log a'_{x_t x_{t+1}} = \sum_{i=1}^{s} \left[\sum_{x \in X} p(x, a, A, B) \sum_{t : x_t = i} \log a'_{i, x_{t+1}}\right]$$

is itself a sum of s functions the ith of which involves only a'_{ij}, $j = 1, ..., s$, which can be maximized separately. If we let $n_{ij}(x)$ be the number of t's with

$x_t = i$, $x_{t+1} = j$ in the sequence of states specified by x, we can write the ith function as

$$\sum_{j=1}^{s} \sum_{x \in X} n_{ij}(x)\, p(x, a, A, B) \log a'_{ij} = \sum_{j=1}^{s} A_{ij} \log a'_{ij}$$

where $A_{ij} = \sum_{x \in X} n_{ij}(x)\, p(x, a, A, B)$. But

$$\sum_{j=1} A_{ij} \log a'_{ij}$$

as a function of $\{a'_{ij}\}$, subject to the restraints

$$\sum_{j=1}^{s} a'_{ij} = 1, \qquad a'_{ij} \geqslant 0,$$

attains a global maximum at the single point

$$\bar{a}_{ij} = A_{ij} \Big/ \sum_{j=1}^{s} A_{ij}\,.$$

This $\{\bar{a}_{ij}\}$ agrees with the first expression of (2); i.e.,

$$\sum_{t=0}^{t-1} P(X_t = i, \quad X_{t+1} = j \mid \{y_t\}, \{a_{ij}, b_{ij}(k)\}) = A_{ij}/P(\{y_t\} \mid \{a_{ij}, b_{ij}(k)\}).$$

Similarly we obtain

$$\bar{a}_i = \sum_{x_0 = i} p(x, a, A, B) \Big/ \sum_{x} p(x, a, A, B),$$

$$\bar{b}_{ij}(k) = \sum p(x, a, A, B) \sum_{x_t = i, x_{t+1} = j, y_{t+1} = k} 1 \Big/ \sum_{x} p(x, a, A, B) \sum_{x_t = i, x_{t+1} = j} 1,$$

in agreement with (1). Of course $\bar{a}_i$, $\bar{a}_{ij}$, $\bar{b}_{ij}(k)$ are computed by inductive calculations as indicated in the second expression of (2) and in (3), not as in the above formulas.

We have now shown that the transformation τ increases P in the case where the output observables Y take values in a finite state space.

We can also consider the case [2] where the output observables Y_t are real-valued. For example, imagine that

$$P(Y_t = y \mid X_t = i) = \frac{1}{(2\pi)^{1/2}\sigma_i} \exp \frac{-(y_t - m_i)^2}{2\sigma_i^2} = b(m_i, \sigma_i, y_t);$$

i.e., associated with state i of an unseen Markov process there is a normally distributed variable with an unknown mean m_i and standard deviation σ_i. Now we wish to maximize the likelihood density of an observation $y_1, \ldots, y_T$,

$$P(a, A, m, \sigma) = \sum_{x \in X} p(a, A, m, \sigma, x)$$

where

$$p(a, A, m, \sigma, x) = a_{x_0} a_{x_0 x_1} b(m_{x_1}, \sigma_{x_1}, y_1) \cdots a_{x_{T-1} x_T} b(m_{x_T}, \sigma_{x_T}, y_T).$$

With

$$Q(a, A, m, \sigma, a', A', m', \sigma') = \sum_{x \in X} p(x, a, A, m, \sigma) \log p(x, a', A', m', \sigma')$$

Theorem 3 applies since everything is nonnegative; it is sufficient to find $\bar{a}$, $\bar{A}, \bar{m}$, $\bar{\sigma}$ such that

$$Q(a, A, m, \sigma; \bar{a}, \bar{A}, \bar{m}, \bar{\sigma}) \geqslant Q(a, A, m, \sigma; a, A, m, \sigma).$$

An argument similar to one given previously shows that:

THEOREM 4. [2] *For each fixed $\{a, A, m, \sigma\}$, the function $Q(a, A, m, \sigma; a', A', m', \sigma')$ attains a global maximum at a unique point. This point $\tau(a, A, m, \sigma)$, the transform of $\{a, A, m, \sigma\}$, is given by*

$$\bar{a}_{ij} = \frac{\sum_t \alpha_t(i)\, a_{ij} \beta_{t+1}(j)\, b(m_j, \sigma_j, y_{t+1})}{\sum_{j=1}^{s} \sum_t \alpha_t(i)\, a_{ij} \beta_{t+1}(j)\, b(m_j, \sigma_j, y_{t+1})},$$

$$m_j = \frac{\sum_t \alpha_t(j)\, \beta_t(j)\, y_t}{\sum_t \alpha_t(j)\, \beta_t(j)},$$

$$\sigma_j^2 = \frac{\sum_t \alpha_t(j)\, \beta_t(j) (y_t - m_j)^2}{\sum_t \alpha_t(j)\, \beta_t(j)}.$$

The last two can be interpreted, respectively, as a posteriori means and variances.

More generally, let $b(y)$ be a strictly log concave density, i.e., $(\log b)'' < 0$. We introduce a two-parameter family involving location and scale parameters m_i, σ_i in state i by defining $b(m, \sigma, y) = b((y - m)/\sigma)$ as we did for the normal density above. The following theorem is somewhat harder to prove than the previous results for the discrete and normal output variables:

THEOREM 5. [2] *For fixed a, A, m, σ the function $Q(a, A, m, \sigma, a', A', m', \sigma')$ attains a global maximum at a single point $(\bar{a}, \bar{A}, \bar{m}, \bar{\sigma})$. The* *transformation* $\tau(a, A, m, \sigma) = (\bar{a}, \bar{A}, \bar{m}, \bar{\sigma})$ *thus defined is continuous and* $P(\tau(a, A, m, \sigma)) \geqslant P(a, A, m, \sigma)$ *with equality if and only if* $\tau(a, A, m, \sigma) = (a, A, m, \sigma)$ *which, in turn, holds if and only if* (a, A, m, σ) *is a critical point of* P.

However, the new $\bar{m}_i$, $\bar{\sigma}_i$ do not have obvious probabilistic interpretations as in the normal case above. Moreover, these $\bar{m}_i$ and $\bar{\sigma}_i$ cannot be inductively computed as in the finite and normal output cases. These facts greatly decrease the interest in the last transformation τ.

We now consider convergence properties of the iterates of the transformation τ. We have $P(\tau(\lambda)) \geqslant P(\lambda)$, equality holding if and only if $\tau(\lambda) = \lambda$ which holds if and only if λ is a critical point of P. It follows that if λ_0 is a limit point of the sequence $\tau^n(\lambda)$, then $\tau(\lambda_0) = \lambda_0$. [In fact, if $\tau^{n_i} \to \lambda_0$, then $P(\lambda_0) \leqslant P(\tau(\lambda_0)) = \lim_i P(\tau^{n_i+1}(\lambda)) \leqslant \lim_i P(\tau^{n_{i+1}}(\lambda)) = P(\lambda_0)$.] We want to conclude that $\tau^n(\lambda) \to \lambda_0$. If P has only finitely many critical points so that τ has only finitely many fixed points, this follows as an elementary point set topology exercise. However, at least theoretically, if P has infinitely many critical points, limit cycle behavior is possible.

However, τ has additional properties beyond those just used and it is possible that a theorem guaranteeing convergence to a point is provable under suitable hypotheses. For related material see References [*3*] and [*4*].

References

1. L. E. Baum and J. A. Eagon, An inequality with applications to statistical prediction for functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* **73** (1967), 360–363.

2. L. E. Baum, T. Petrie, G. Soules, and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** (1970), 164–171.

3. G. R. Blakely, Homogeneous non-negative symmetric quadratic transformations. *Bull. Amer. Math. Soc.* **70** (1964), 712–715.

4. L. E. Baum and G. R. Sell, Growth transformations for functions on manifolds. *Pacific J. Math.* **27** (1968), 211–227.