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

\[
P(X_{t+1} = j \mid X_t = i, X_{t-1}, ..., Y_t, ...) = a_{ij}, \quad i, j = 1, ..., s;
\]

\[
P(Y_{t+1} = k \mid X_{t+1} = j, X_t = i, X_{t-1}, ..., Y_t, Y_{t-1}, ...) = b_{ij}(k),
\]

\[i, j = 1, ..., s, \quad k = 1, ..., r.
\]

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

\[
a_{ij} \geq 0, \quad \sum_{j=1}^{s} a_{ij} = 1, \quad i = 1, ..., s,
\]

\[
b_{ij}(k) \geq 0, \quad \sum_{k=1}^{r} b_{ij}(k) = 1, \quad i, j = 1, ..., s.
\]

We see a \( Y \) sample \( \{Y_1 = y_1, Y_2 = y_2, ..., Y_T = y_T\} \) 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 \( \{y_t\} \)

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

\[
= \sum_{i_0, i_1, ..., i_{T-1}} a_{i_0 i_1} b_{i_0 i_1}(y_1) a_{i_1 i_2} b_{i_1 i_2}(y_2) ... a_{i_{T-1} i_T} b_{i_{T-1} i_T}(y_T)
\] (1)

where \( a_i \) are initial probabilities for the Markov process. For this purpose
we define a transformation \( r(\{a_{ij}, b_{ik}\}) = \{a_{ij}, b_{ik}\} \) of \( M \) into itself where

\[
\bar{a}_{ij} = \frac{\sum_i P(X_t = i, X_{t+1} = j \mid \{y_t\}, \{a_{ij}, b_{ik}\})}{\sum_i P(X_t = i \mid \{y_t\}, \{a_{ij}, b_{ik}\})} = \frac{\sum_i \alpha(i) \beta_{i+1}(j) a_{ij} b_{ik}(y_{t+1})}{\sum_i \alpha(i) \beta(i)} = \frac{a_{ij}^\prime}{\sum_i a_{ij} P(\bar{a}_{ij})},
\]

\( \beta_{ik}(k) = \frac{\sum_{y_{t+1} \in \mathcal{Y}} P(X_t = i, X_{t+1} = j \mid \{y_t\}, \{a_{ij}, b_{ik}\})}{\sum_i P(X_t = i \mid \{y_t\}, \{a_{ij}, b_{ik}\})} = \frac{\sum_i \beta_{i+1}(j) a_{ij} b_{ik}(y_{t+1})}{\sum_i \beta(i)} = \frac{b_{ik}(k)^\prime}{\sum_i b_{ik}(k) P(\bar{b}_{ik}(k))}.
\]

The second of the equivalent forms in Eqs. (2) contains quantities \( r_t(\bar{a}_{ij}), \beta_t(i), i = 1, \ldots, s, t = 0, 1, \ldots, T - 1 \), which are defined inductively forwards and backwards, respectively, in \( t \) by

\[
\alpha_{i+1}(j) = \sum_{i=1}^s \alpha(i) a_{ij} b_{ik}(y_{t+1}), \quad j = 1, \ldots, s, \quad t = 0, 1, \ldots, T - 1,
\]

\[
\beta_{i+1}(j) = \sum_{i=1}^s \beta_{i+1}(j) a_{ij} b_{ik}(y_{t+1}), \quad i = 1, \ldots, s, \quad t = T - 1, T - 2, \ldots, 0.
\]

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

\[
P(\{a_{ij}, b_{ik}\}) = \sum_{i=1}^s \alpha(i) \beta(i)
\]

(identically in \( r \)) can be computed with \( 4s^2T \) multiplications rather than the \( 2Ts^2+1 \) multiplications indicated in the defining formula (1). Similarly, the partial derivatives of \( P \) needed for defining the image in (2) are computed from the \( \alpha \)'s and \( \beta \)'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 \textit{a posteriori} reestimation suggested the transformation \( r \) originally and is embodied in the first expressions for \( \bar{a}_{ij} \) and \( \bar{b}_{ik}(k) \).

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

\[
\alpha_{ij} = \frac{a_{ij}^\prime}{\sum_i a_{ij} P(\bar{a}_{ij})}, \quad \beta_{ik}(k) = \frac{b_{ik}(k)^\prime}{\sum_i b_{ik}(k) P(\bar{b}_{ik}(k))}.
\]

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

Note that \( \tau \) 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 \( P(\{a_{ij}, b_{ik}(k)\}) \) is a homogeneous polynomial of degree \( 2T + 1 \) in \( a_i, a_j, b_{ik}(k) \) and obtained the result as an application of the following theorem.

THEOREM 2. [1] Let

\[
P(z_1, \ldots, z_n) = \sum_{\mu_1, \ldots, \mu_n} c_{\mu_1, \ldots, \mu_n} z_1^{\mu_1} \cdots z_n^{\mu_n} \quad \text{where} \quad c_{\mu_1, \ldots, \mu_n} \geq 0
\]

and \( \mu_1 + \cdots + \mu_n = d \). Then

\[
\tau : \{z_i\} \rightarrow \left\{ \frac{z_i^d P(\bar{z}_i)}{\sum_j z_j^d P(\bar{z}_j)} \right\}
\]

maps \( D : z_i > 0, \sum z_i = 1 \) into itself and satisfies \( P(\tau(z_i)) \geq P(z_i) \). 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

\[
z_i^d P(\bar{z}_i) = \sum_{\mu_1, \ldots, \mu_n} c_{\mu_1, \ldots, \mu_n} z_1^{\mu_1} \cdots z_n^{\mu_n}
\]

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
We wish to prove Jensen's inequality to the concave function $\log x$. We wish to prove $P(\lambda) \geq P(\lambda)$ or, equivalently, $\log[P(\lambda)/P(\lambda)] \geq 0$. Now,

$$\log \frac{P(\lambda)}{P(\lambda)} = \log \left[ \frac{1}{P(\lambda)} \int x p(x, \lambda) \, d\mu(x) \right]$$

$$= \log \int x \left[ \frac{p(x, \lambda) \, d\mu(x)}{P(\lambda)} \right] p(x, \lambda)$$

$$\geq \int x \left[ \frac{p(x, \lambda) \, d\mu(x)}{P(\lambda)} \right] \log \frac{p(x, \lambda)}{P(\lambda)}$$

$$= \frac{1}{P(\lambda)} \left[ Q(\lambda, \lambda) - Q(\lambda, \lambda) \right] \geq 0$$

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

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

(a) For fixed $\lambda$, $Q(\lambda, \lambda')$ assumes its global maximum as a function of $\lambda'$ at a unique point $\tau(\lambda)$.

(b) $\tau(\lambda)$ is continuous.

(c) $\tau(\lambda)$ is effectively computable.

(d) $P(\tau(\lambda)) \equiv P(\lambda)$ 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 $\lambda'$.

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

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

Also

$$Q(a, A, B; a', A', B')$$

$$= \sum_{x} p(x, a, A, B) \left[ \log a_{\epsilon_{0}} + \sum_{t} \log a'_{\epsilon_{t} \epsilon_{t+1}} + \sum_{t} \log b'_{\epsilon_{t} \epsilon_{t+1}}(y_{t+1}) \right].$$

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 $a'_{\epsilon}$, 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

$$\sum_{x} p(x, a, A, B) \sum_{t} \log a'_{\epsilon_{t} \epsilon_{t+1}} = \sum_{t} \left[ \sum_{x} p(x, a, A, B) \sum_{\epsilon_{t} \epsilon_{t+1}} \log a'_{\epsilon_{t} \epsilon_{t+1}} \right]$$

is itself a sum of $s$ functions the $i$th 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 $i$th function as

$$\sum_{j=1}^{s} \sum_{x \in X} n_{ij}(x) \log a_{ij} = \sum_{j=1}^{s} A_{ij} \log a_{ij}$$

where $A_{ij} = \sum_{x \in X} n_{ij}(x) \log a_{ij}$. But

$$\sum_{j=1}^{s} A_{ij} \log a_{ij}$$

as a function of $(a_{ij})$, subject to the restraints

$$\sum_{j=1}^{s} a_{ij} = 1, \quad a_{ij} \geq 0,$$

attains a global maximum at the single point

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

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

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

Similarly we obtain

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

$$\bar{b}_i(k) = \sum_{a, i} p(x, a, A, B) \sum_{a \cdot i} 1 / \sum_{x} p(x, a, A, B) \sum_{a \cdot i} 1,$$

in agreement with (1). Of course $\bar{a}_i, \bar{a}_{ij}, \bar{b}_i(k)$ 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 $T$ 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 $Y_1, \ldots, Y_T$ are real-valued. For example, imagine that

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

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 $\sigma_i$.

Now we wish to maximize the likelihood density of an observation $y_{t_1}, \ldots, y_T$,

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

where

$$p(a, A, m, \sigma, x) = a_{ij} a_{ji} b(m_{ij}, \sigma_{ij}, y_i) \cdots a_{ji} a_{ji} b(m_{ji}, \sigma_{ji}, y_T).$$

With

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

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

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

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

$$\tau_{ij} = \frac{\sum_{i} \alpha_i(i) \beta_{i+1}(j)}{\sum_{i} \alpha_i(i) \beta_{i+1}(j)} \frac{b(m_i, \sigma_i, y_{t+1})}{b(m_j, \sigma_j, y_{t+1})},$$

$$m_j = \frac{\sum_{i} \alpha_i(j) \beta_{i}(j) y_t}{\sum_{i} \alpha_i(j) \beta_{i}(j)},$$

$$\sigma_j^2 = \frac{\sum_{i} \alpha_i(j) \beta_{i}(j)(y_t - m_j)^2}{\sum_{i} \alpha_i(j) \beta_{i}(j)}.$$

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., $(\log b)^\alpha < 0$. We introduce a two-parameter family involving location and scale parameters $m_i, \sigma_i$, 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 $a, A, m, \sigma$, 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 \( \tau(a, A, m, \sigma) = (\overline{a}, \overline{A}, \overline{m}, \overline{\sigma}) \) thus defined is continuous and \( P(\tau(a, A, m, \sigma)) \geq P(a, A, m, \sigma) \) with equality if and only if \( \tau(a, A, m, \sigma) = (a, A, m, \sigma) \) which, in turn, holds if and only if \( (a, A, m, \sigma) \) is a critical point of \( P \).

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

We now consider convergence properties of the iterates of the transformation \( \tau \). We have \( P(\tau(\lambda)) \geq P(\lambda) \), equality holding if and only if \( \tau(\lambda) = \lambda \) which holds if and only if \( \lambda \) is a critical point of \( P \). It follows that if \( \lambda_0 \) 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 \( P(\lambda_0) = P(\tau(\lambda_0)) = \lim_i P(\tau^{n_i+1}(\lambda)) = \lim_i P(\tau^{n_i+1}(\lambda)) = P(\lambda_0) \).] We want to conclude that \( \tau^n(\lambda) \to \lambda_0 \). If \( P \) has only finitely many critical points so that \( \tau \) 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, \( \tau \) 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