CYCLIC I_0 PROJECTIONS AND ITS APPLICATIONS IN STATISTICS

by

N. D. Shyamalkumar
Purdue University

Technical Report #96-24

Department of Statistics
Purdue University
West Lafayette, IN USA

June 1996

CYCLIC I_0 PROJECTIONS AND ITS APPLICATIONS IN STATISTICS

N. D. Shyamalkumar

Department Of Statistics, Purdue University, U.S.A. *

Abstract

We study the behaviour of cyclic I_0 projections between two classes of probability measures. It is then shown that some convergence problems arising in Statistics can be viewed as convergence of cyclic I_0 projections. One such problem arises while using the EM algorithm for estimating the mixing distribution. We derive a basic inequality which says that the I_0 divergence distance of the true mixing distribution (or M.L.E. when we are working with a data) from the iterate is reduced by at least the distance of the true marginal from the marginal using the previous iterate at each iteration. Proof of convergence of the algorithm in the weak topology follows by exploiting this inequality in the generality of locally compact parameter spaces. This inequality also provides some intuition as to why the EM algorithm in the mixture problem gets to a neighborhood of the MLE in the first few iterations and then the speed of convergence slows down considerably, an empirical observation seen in the literature. There is some overlap of our results with that in the literature. The other problem we dealt with is the convergence of the data augmentation method of Tanner and Wong. We give a simple proof based on an inequality of the type discussed above. The convergence is proved in the total variation norm.

*Research supported by Purdue research foundation grant and partially supported by the National Science Foundation, Grants DMS-8923071 and DMS-9303556.

Key words and phrases: EM algorithm, Mixing distribution, Empirical Bayes, Data Augmentation Method.

1 Introduction

This paper studies the iterations of operators defined on the space of probability measures and establishes convergence results and provides characterization of the limit points under certain conditions. To define the operators we study in this paper we need the following notations.

Let $(\mathcal{X}, \mathcal{A})$ denote the sample space which we assume to be Polish with its associated Borel σ field. Let $(\Theta, \mathcal{B})$ denote the parameter space which also we assume to be Polish with its associated Borel sigma field. Let $P(.,.)$ be a Markov kernel, by which we mean that $P(.,.)$ is a mapping from $(\mathcal{X}, \mathcal{A}) \times \Theta$ to the unit interval such that for each fixed value of θ, $P(., \theta)$ is a probability measure on $(\mathcal{X}, \mathcal{A})$ and for each fixed set in $\mathcal{A}$ it is a measurable function in θ. We assume that the support of $P(., \theta)$ is invariant with respect to θ. Further assume that, for each θ, $P(., \theta)$ is dominated by a sigma finite measure μ and let the Radon-Nikodym derivatives be denoted by $f(x, \theta)$. For η a probability measure on $(\Theta, \mathcal{B})$, by $\eta(.,.)$ we shall denote the regular conditional probability of the measure induced by the Markov kernel $P(.,.)$ and η conditioned on the sigma field $\mathcal{A} \otimes \{\Theta, \phi\}$. Statisticians call $\eta(.,.)$ the posterior induced by the Markov kernel (model) $P(.,.)$ and the probability measure (prior) η. By the measure m_η we denote the marginal probability induced by the above Markov kernel and η. Note that in what follows we have denoted the Radon-Nikodym derivatives by the measures themselves.

Let π be a probability measures on $(\Theta, \mathcal{B})$ and m a probability measure on $(\mathcal{X}, \mathcal{A})$. We first study the behavior of the iterations of the following operator which we denote by T_m, an operator which takes probability measures on $(\Theta, \mathcal{B})$ to probability measures on $(\Theta, \mathcal{B})$. The action of T_m on π is the probability measure $\int \pi(., x)\, dm(x)$. More explicitly,

$$dT_m(\pi)(\theta) = \int \frac{f(x, \theta)}{m_\pi(x)}\, dm(x) \cdot d\pi(\theta).$$

That this is a probability measure is clear from the definitions of the quantities involved and the fact that a convex combination of probability measures is also a probability measure. The n-th iteration of T_m on π will be denoted by $T_m^n(\pi)$. Here we study convergence of the iterations in the weak topology on the space of all probability measures on $(\Theta, \mathcal{B})$. We shall denote this by simply writing $\pi_n \rightharpoonup \pi$. That this operator has infinitely many fixed points under non-trivial conditions is easy to see. So one might wonder if one can get nice results on convergence of the iterations. Such results, in particular, form part of the contents of this paper.

To define the second operator we study, we need some more notations. Let ν be a σ-finite measures on $(\Theta, \mathcal{B})$. By $\mathcal{P}_{\mathcal{X}}$ and $\mathcal{P}_{\Theta}$ we denote the set of all probability measures on $(\mathcal{X}, \mathcal{A})$ and $(\Theta, \mathcal{B})$, respectively. Let $G(.,.)$ be a Markov kernel on $(\Theta, \mathcal{B}) \otimes \mathcal{X}$. We assume that $G(., x)$ is dominated by ν for each x in $\mathcal{X}$. We shall denote by $g(., x)$ their respective Radon-Nikodym derivatives.

We define $K(.,.)$, a kernel on $\Theta \times \Theta$, by

$$K(\theta, \gamma) = \int g(\theta, x) f(x, \gamma)\, d\mu(x) \quad \forall\, \theta, \gamma \in \Theta,$$

and W is defined as the operator,

$$W(\pi)(\theta) = \int K(\theta, \gamma) \pi(\gamma) d\nu(\gamma) \quad \forall\, \theta \in \Theta.$$

We study the behavior of the iterations of the operator W on $\mathcal{P}_{\Theta}$. As observed in the earlier papers on this subject, for eg. Tanner and Wong (1987), we see that the operator W is a Markov transition operator and hence an L_1 non-expansive mapping. This is the way that the previous papers have tackled this problem. Here we adopt a different approach mentioned below.

Both the above operators can be visualized naturally as a part of an alternating projection scheme between two convex classes of probability measures, the projections in terms of the I divergences. Such a visualization gives many properties of the operator for free, in some sense. Csiszár & Tusnády (1984) prove a general theorem for alternating projection schemes under some geometric conditions on the projections. They then go on to show that these conditions are satisfied by the I-divergence projections and then study the above operator under conditions when the parameter space is a finite set. Here we take a slightly different approach. We first prove an inequality which implies that the distance between the iterate and the limit is monotonic non-increasing in the number of the iteration. The key fact that we use to prove such an inequality is a result which gives necessary and sufficient conditions for I_0 projections, very similar to results for the general f-divergences of Liese and Vajda (1987) and Rüschendorf (1984). Then by using some basic facts from the theory of f-divergences (see Liese and Vajda (1987) and Vajda (1989)) we are able to deduce all the results in Csiszár & Tusnády (1984) pertaining to the operator of our interest. Moreover, we also deal with the case when the parameter space is not a finite set. We hope that this paragraph beckons the reader to at least browse through Csiszár & Tusnády (1984) as their general theorem on alternating projections is certainly a deep and powerful result.

When the parameter space is finite the iterations of the operator T can be seen to be an EM algorithm for a suitably defined mixture problem, see Dempster, Laird, Rubin (1977). So a natural question to ask would be in what way is the approach of this paper different from that for the convergence of the EM algorithm as given in Wu (1983) (see also Boyles (1983))? Also, do we differ in conclusions? First of all the proofs for the convergence of the EM algorithm use the fact that the likelihood is non decreasing as we proceed along the iterations. From this they try to deduce the convergence of the iterations. Here we get an inequality which deals directly with the iterations and hence we are able to get stronger results in the sense that we prove that the iterations always converge and, when we start from an interior point of the simplex, we always end up with a global maximum. As a specific example, consider theorem 4.2 of Redner and Walker (1984). This theorem uses the general result for the EM algorithm as given in Wu (1983). These methods are, in turn, based on the relevant general theorem given in Zangwill (1969). To get the convergence of the iterations they assume that the matrix of the second derivatives of the likelihood is non-negative definite at all points of the simplex. This implies uniqueness of the MLE. Such conditions and uniqueness of the MLE are not required for the proofs along our lines. We do agree that this might not be highly relevant to the users of the algorithm but is nevertheless of theoretical interest. All this is not meant to be a criticism of existing methods of proof of the EM algorithm but to point out that in some cases of the applications of the EM algorithm alternative methods of proof might give stronger results.

In the next section we summarize the results from the theory of f-divergences that we shall be requiring. In the third section we give the main results of this paper relating to operator T which deal with the convergence of its iterations. We have split this section into four sub-sections. The results in the third sub-section are the ones which are useful in practice, at least currently. The results of the second sub-section generalize those of the first but require the results from Information theory given in the second section. The last sub-section points towards one of the many applications of the operator T. In the fourth section we deal with the data augmentation method.

2 I_0 Projections

We shall below define I_0 divergence and give some results from the literature about I_0 projections that we shall need in this paper. For two probability measures, P and Q, on a measure space, we shall denote by $I_0(P,Q)$, the Kullback-Leibler I_0 divergence between the two measures, defined by,

$$I_0(P,Q) = \int \log\left(\frac{dQ}{dP}\right)\ dQ$$

where the densities are with respect to some dominating measure. It is easy to see that this divergence is invariant w.r.t. to the choice of the dominating measure. Here and in the sequel we understand

$$\log 0 = -\infty, \quad \log\left(\frac{a}{0}\right) = +\infty, \quad 0\cdot(\pm\infty) = 0,$$

where $a > 0$. $I_0(P,Q)$ is always non-negative and vanishes only when $P = Q$. If $Q \not\ll P$ then $I_0(P,Q) = +\infty$. Just for the sake of general interest we also define the I_1 divergence, which also is due to Kullback & Leibler, as

$$I_1(P,Q) = \int_{\{dQ>0\}} \log\left(\frac{dP}{dQ}\right)\ dP \ + \ \infty\cdot P\{dQ=0\}$$

Note that $I_0(P,Q) = I_1(Q,P)$. It is important to note that the divergence is asymmetric and moreover does not satisfy the triangular inequality. The obvious symmetrized version, used by Jeffreys (1948) does not either gives us the triangular inequality. The triangular inequality cannot be obtained by any other reasonable modification of the divergence, see Csiszár (1964). So all this, in particular , implies that there is no reasonable way to derive a metric based on this divergence. But, never the less, we can talk about the induced topology, by which we mean the smallest topology containing the balls. In this connection it is interesting to note the following information inequality which, in particular, implies that this induced topology is stronger than the one of the total variation metric.

Let us recall that the total variation distance between two probability measures P and Q is given by

$$\|P - Q\| = 2 \cdot \sup_A |P(A) - Q(A)|$$

where the supremum is taken over all sets in the σ-field on which the probability measures are defined. The mysterious factor 2 is included to make it equal to the L_1 distance between the densities of the two measures with respect to any dominating σ-finite measure. Note that the L_1 metric is invariant to the choice of the dominating measure.

The below given information inequality was independently discovered by many in the late 1960's; among them being Csiszár (1967), Kemperman (1967) and Kullback (1967). We refer to them for its proof which follows essentially from the monotonicity property of the I- divergence.

Lemma 2.1 *Let P and Q be two probability measures. Then with the above notations we have*

$$I_0(P,Q) \geq \frac{(\|P-Q\|)^2}{2}.$$

By symmetry the above is true with I_1 replacing I_0.

Proof. See Csiszár (1967).

□

Now we shall describe what we mean by the I_0 projection of a probability measure, say Q, on a convex set of probability measures, say $\mathcal{T}$. We say that P^* is the I_0 projection of Q

on the set $\mathcal{T}$ if $I_0(P^*, Q)$ is finite and

$$I_0(P^*, Q) = \inf_{P \in \mathcal{T}} I_0(P, Q).$$

If in the above, the other version of the Kullback-Leibler divergence is used then we would have what is called the I_1 projection of Q on the set $\mathcal{T}$. In this paper we shall be dealing with only the former. The existence of P^* in general is altogether another question , but if it exists then it's uniqueness when $\mathcal{T}$ is dominated by Q follows from strict convexity of the I_0 divergence. On this point it is of interest to note that the I_1 projection is unique even if $\mathcal{C}$ is not dominated by Q. Moreover, the I_1 projection exists if the set $\mathcal{C}$ is closed under the total variation metric and there exists at least one measure with finite divergence with Q.

We give some more notations that would be needed. For any space $(\mathcal{Y}, \mathcal{B})$ we shall denote by $\mathcal{P}(\mathcal{Y})$ the set of all probability measures on $(\mathcal{Y}, \mathcal{B})$. As we shall never work with more than one sigma field on a set, this should not cause any confusion. For any subset of $\mathcal{T}$ of $\mathcal{P}(\mathcal{Y})$, we shall denote by $\mathcal{C}(\mathcal{T})$ the convex hull of $\mathcal{T}$. We shall say $P \ll Q$ if P is absolutely continuous with respect to Q. We shall say that a class of probability measures $\mathcal{T}$ is dominated by Q if each P in $\mathcal{T}$ is dominated by Q.

The following is a result which, in particular, implies the existence of I_0 projections on weakly compact sets of probability measures.

Lemma 2.2 *If $(\mathcal{Y}, \mathcal{B})$ is a Polish space or even if $\mathcal{Y}$ is a separable metric space with its associated Borel sigma-field, then I_0 is lower semicontinuous on the space $(\mathcal{P}(\mathcal{Y}) \times \mathcal{P}(\mathcal{Y}))$ with the product of weak topologies.*

PROOF. Follows from a more general theorem of Liese & Vajda (1987).
□

The following result is interesting, in that it is analogous to a well known theorem from the theory of Banach spaces. For any subset of finite measures $\mathcal{T}$ and a finite measure Q we shall denote by $I_0(\mathcal{T}, Q)$ the quantity $\inf_{P \in \mathcal{T}} I_0(P, Q)$.

Lemma 2.3 *Let $\mathcal{T}$ be a non-empty convex subset of probability measures dominated by Q.*

If P^* and $\{P_n\}_{n\geq 1} \in \mathcal{T}$ satisfy

$$\lim_{n\to\infty} I_0(P_n, Q) \;=\; I_0(P^*, Q) \;=\; I_0(\mathcal{T}, Q),$$

then $\|P_n - P^*\| \to\; 0$ as $n \to \infty$.

Proof. Follows from a result for f-divergences given in Liese & Vajda (1987).
□

Remark 2.1 If, instead, $\mathcal{T}$ is a convex subset of finite measures dominated by Q, then we would still have convergence in Q measure of p_n to p^* where the former quantities represent the Radon-Nikodym derivatives of P_n and P^* w.r.t. Q, respectively. To have convergence in the total variation sense we would need stronger conditions which would gives the convergence of the norm of P_n to that of P^* so that the type of argument using Scheffe's theorem goes through.

The following important result is called the monotonicity property of I_0 divergence. See Liese & Vajda (1987) for proof and other implications of this property.

Lemma 2.4 *For every $\pi, \pi^{'} \in \mathcal{P}(\Theta)$ and kernel $K : (\Theta, \mathcal{B}) \Rightarrow (\mathcal{X}, \mathcal{A})$ we have*

$$I_0(\pi, \pi^{'}) \geq I_0(\pi * K, \pi^{'} * K),$$

where equality takes place iff K is sufficient for $(\pi, \pi^{'})$. $\pi * K$ is the measure on $(\mathcal{X}, \mathcal{A})$ given by $\int K(., \theta) d\pi(\theta)$.

We now prove a result about I_0 projection which will play an important role in what follows.

Lemma 2.5 *Let $\mathcal{T}$ be a non-empty convex subset of finite measures dominated by Q such that $I_0(\mathcal{T}, Q) < \infty$. Then P^*, an element of $\mathcal{T}$ with $p^* = \frac{dP^*}{dQ}$, is an I_0 projection of Q onto $\mathcal{T}$ iff*

$$\sup_{P\in\mathcal{T}} \int \frac{1}{p^*} dP \;= 1 \textit{ and } I_0(P^*, Q) < \infty.$$

PROOF. First we note that, since $I_0(P^*, Q) < \infty$, we have

$$I_0((1-\alpha)P^* + \alpha P, Q) < \infty, \forall \alpha \in [0,1).$$

By using the above fact and that $-\log(.)$ is a convex function, one can prove using DCT that the directional derivative of $I_0(., Q)$ at P^* in the direction of P is given by

$$\int \frac{p^* - p}{p^*} dQ = 1 - \int \frac{1}{p^*} dP$$

where $p = \frac{dP}{dQ}$. Now by Whittle's theorem, see Vajda(1989), the lemma follows.

□

Remark 2.2 It follows from the proof that the "if" part holds even if $\mathcal{T}$ is not convex. Also one can conclude as a consequence of the above theorem that $\mathcal{T} \ll P^*$ and that P^* is equivalent to Q.

Remark 2.3 The above result and it's proof is very similar to the result for f-divergences given in Liese & Vajda (1987), see also Rüschendorf (1984). But the important difference is that we do not require that $I_0(P, Q) < \infty$ for all $P \in \mathcal{T}$. Removal of this condition is useful to us in the fourth section. Generalization to finite measures is obvious.

Kullback-Leibler divergence has been important in statistical theory because of the omnipresence of likelihood ratios. This is the main reason for it's appearance in asymptotic statistical theory. It also figures in contingency table theory for an altogether different reason. The iterative proportional fitting method, for example, converges to the I_1 projection of the starting point. We do not want to dwell on the uses of divergence in statistics, but should at least mention, Bahadur (1971), Csiszár (1975), Dykstra (1985), Hoeffding (1965) and Kullback (1959).

3 Convergence Results for T_m

We have split this section into three sub-sections, even though the second sub-section gives results which generalize that of sub-section one, as the first does not make use of any non-trivial results from Information theory. Thus the first sub-section is more accessible. The third subsection deals with the case when we are dealing with finite measures instead of probability measures. Of course, here the interpretations of posterior, etc., breaks down but some of the results do hold in this generality and the goal of this section is to merely point these out.

3.1 Case 1 : $m \in \mathcal{C}\{f(.,\theta) : \theta \in \Theta\}$.

In this sub-section we assume that $m \in \mathcal{C}\{f(.,\theta) : \theta \in \Theta\}$. Let π^* be a probability measure on $(\Theta, \mathcal{B})$ such that $m = m_{\pi^*}$. Henceforth in this sub-section the operator T_m defined before shall be denoted by T_{π^*}.

Lemma 3.1 *Let π be a probability on $(\Theta, \mathcal{B})$ and let ν be a dominating sigma finite measure for π and π^*. Then*

$$\int \pi^*(\theta)\ \log\left(\int \frac{m_{\pi^*}(x)}{m_\pi(x)} P(dx,\theta)\right) d\nu(\theta)$$

is well defined. Moreover when $I_0(\pi,\pi^*)$ is finite the above is finite too.

Proof. We shall show the negative part of the integrand is integrable. To this end note that

$$\begin{aligned}
& \int \pi^*(\theta)\ \left[\log\left(\int \frac{m_{\pi^*}(x)}{m_\pi(x)} P(dx,\theta)\right)\right]^- d\nu(\theta) \\
\leq & \int \pi^*(\theta)\ \left[-\log\left(\int \frac{m_\pi(x)}{m_{\pi^*}(x)} P(dx,\theta)\right)\right]^- d\nu(\theta) \\
= & \int \pi^*(\theta)\ \log\left(\int \frac{m_\pi(x)}{m_{\pi^*}(x)} P(dx,\theta) \vee 1\right) d\nu(\theta) \\
\leq & \int \pi^*(\theta)\ \int \frac{m_\pi(x)}{m_{\pi^*}(x)} P(dx,\theta) d\nu(\theta) \\
= & \ 1.
\end{aligned}$$

This proves the first part. When $I_0(\pi,\pi^*)$ is finite we have

$$I_0(T_{\pi^*}(\pi),\pi^*) = I_0(\pi,\pi^*) - \int \pi^*(\theta)\ \log\left(\int \frac{m_{\pi^*}(x)}{m_\pi(x)} P(dx,\theta)\right) d\nu(\theta).$$

This, in particular, implies that the second integral is finite in view of what we showed above.
□

Lemma 3.2 *Let π be a probability on $(\Theta, \mathcal{B})$ and let ν be a common dominating sigma finite measure for π and π^*. Assume that $I_0(\pi,\pi^*)$ is finite. Then we have,*

$$I_0(T_{\pi^*}(\pi), \pi^*) \leq I_0(\pi, \pi^*) - I_0(m_\pi, m_{\pi^*}).$$

This, in particular, implies that we have equality above iff $m_\pi = m_{\pi^*}$.

Proof. By using Jensen's inequality and in view of the previous lemma we have

$$\begin{aligned}
&I_0(T_{\pi^*}(\pi), \pi^*) \\
&\quad = \int \pi^*(\theta) \log \left(\frac{\pi^*(\theta)}{\pi(\theta) \int \frac{m_{\pi^*}(x)}{m_\pi(x)} P(dx, \theta)} \right) d\nu(\theta) \\
&\quad = \int \pi^* \ \log \left(\frac{\pi^*}{\pi} \right) d\nu - \int \pi^*(\theta) \ \log \left(\int \frac{m_{\pi^*}(x)}{m_\pi(x)} P(dx, \theta) \right) d\nu(\theta) \\
&\quad \leq I_0(\pi, \pi^*) - \int \log \left(\frac{m_{\pi^*}(x)}{m_\pi(x)} \right) dm_{\pi^*}(x) \\
&\quad = I_0(\pi, \pi^*) - I_0(m_\pi, m_{\pi^*})
\end{aligned}$$

The case when equality occurs is clear.

□

Remark 3.1 It is interesting to note that, on each action of the operator T_{π^*}, the Kullback-Leibler distance from π^* is decreased by at least the amount of K-L distance between the marginals. This says that, even if we start far away from π^*, we could find ourselves in the "ball park" of π^* in the first few iterations.

Remark 3.2 The above remark gives the analytical reason as to why most of the change in the Kullback-Leibler distance is observed within the first five iterations, an empirical fact known to users of EM algorithms, see Redner & Walker (1984).

Remark 3.3 Note that the above, in some situations does imply a geometric rate of convergence. This can be easily shown, for example, when the sample space and parameter space are finite. But it is possible that the exponent found in the obvious way can be very close to one; this is what happens when we used naive methods to evaluate the exponent in example 1.

Remark 3.4 The monotonicity can, in some situations, let us relax the usual condition on identifiability if we are just interested in the convergence of the iterations. For instance, when one can prove that there exists a subsequence which converges in the sense of I_0 divergence to some probability measure, say, π' which has the same marginal as π^* then by monotonicity we get the convergence of the whole sequence to this π' in the sense of I_0 divergence. This remark has obvious importance in the case when Θ is a finite set.

Remark 3.5 From the above lemma, one can for the case when the parameter space is finite, observe the following:

$$\begin{aligned} I_0(m_\pi, m_{\pi^*}) &\leq I_0(\pi, \pi^*) \ - \ I_0(T_{\pi^*}(\pi), \pi^*) \\ &= \int \log\left(\frac{T_{\pi^*}(\pi)}{\pi}\right) d\pi^* \\ &\leq \sup_{\theta\in\Theta} \log\left(\frac{T_{\pi^*}(\pi)(\theta)}{\pi(\theta)}\right). \end{aligned}$$

The last expression says that, from successive iterations, one can get a bound for the I_0 distance between the marginals. In the case of the MLE, this gives us a bound on how close we are to the global maximum of the likelihood.

Theorem 3.1 *Let*

(i) π be such that $I_0(\pi, \pi^*)$ is finite.

(ii) $m_{\pi_n} \to m_{\pi^*}$ in total variation metric implies $\pi_n \rightharpoonup \pi^*$.

Then we have $T^n_{\pi^*}(\pi) \rightharpoonup \pi^*$.

PROOF. From the above lemma we have

$$I_0(T^{n+1}_{\pi^*}(\pi), \pi^*) \leq I_0(T^n_{\pi^*}(\pi), \pi^*) - I_0(m_{T^n_{\pi^*}(\pi)}, m_{\pi^*}) \ \forall\ n \geq 0.$$

So either we have $I_0(T^n_{\pi^*}(\pi), \pi^*)$ decreasing to zero or $I_0(m_{T^n_{\pi^*}(\pi)}, m_{\pi^*}) \to 0$. In the former case we have $T^n_{\pi^*}(\pi) \to \pi^*$ in the I_0 divergence sense which, by lemma 2.1, implies convergence in total variation sense and hence, in particular, $T^n_{\pi^*}(\pi) \rightharpoonup \pi^*$. In the latter case we have again, by lemma 2.1, that $m_{T^n_{\pi^*}(\pi)} \longrightarrow m_{\pi^*}$ in the total variation sense which by assumption (ii) implies that $T^n_{\pi^*}(\pi) \rightharpoonup \pi^*$.

□

Remark 3.6 Note that, if Θ is compact and π^* is identifiable, then $m_{\pi_n} \longrightarrow m_{\pi^*}$ in total variation metric would imply that $\pi_n \rightharpoonup \pi^*$, in the case when $f(x,\theta)$ is bounded continuous as a function of θ, a.e. μ.

Example 3.1 Let $P(x,\theta) = Bin(2,\theta)(x)$ and $\Theta = \{0, 0.5, 1\}$ with the usual topology. Note that $\mathcal{X} = \{0,1,2\}$ and is also equipped with the usual topology. Let $\pi^*(0) = 1/6$, $\pi^*(0.5) = 2/3$ and $\pi^*(1) = 1/6$. Further let $\pi(0) = 0.99$, $\pi(0.5) = 0.005$ and $\pi(1) = 0.005$. Since Θ is compact and the other conditions of the theorem also hold, the conclusions of the theorem hold for this example. The table below displays the result of the first ten iterations.

n	$T^n_{\pi^*}(\pi)$	$m_{T^n_{\pi^*}(\pi)}$	$I(T^n_{\pi^*}(\pi), \pi^*)$	$I(m_{T^n_{\pi^*}(\pi)}, m_{\pi^*})$	$J(n)$ *
0	(.990,.005,.005)	(.991,.003,.006)	3.549	2.593	-
1	(.333,.400,.267)	(.433,.200,.366)	0.146	0.0508	3.403
2	(.256,.501,.243)	(.382,.251,.367)	0.0559	0.0172	0.0903
3	(.224,.557,.219)	(.363,.279,.358)	0.0253	0.00730	0.0306
4	(.206,.590,.204)	(.353,.295,.352)	0.0123	0.00342	0.0129
5	(.194,.613,.193)	(.347,.306,.347)	0.00630	0.00170	0.0060
6	(.186,.628,.186)	(.343,.314,.343)	0.00331	0.00088	0.0030
7	(.181,.638,.181)	(.341,.319,.340)	0.00177	0.00046	0.0015
8	(.177,.646,.177)	(.339,.323,.338)	0.00096	0.00025	0.0008
9	(.174,.651,.175)	(.337,.326,.337)	0.00053	0.00014	0.0004
10	(.172,.655,.173)	(.336,.328,.336)	0.00029	0.000074	0.0002

* $J(n) := I(n-1) - I(n); I(n) := I(T^n_{\pi^*}(\pi), \pi^*)$

Table 1: Result of ten iterations

□

Lemma 3.3 *Let Θ be a locally compact metric space and we shall denote by ∞ the adjoint single point of the one point compactification. We shall assume that we can extend $f(.,\theta)$ such that it is bounded upper semi-continuous at ∞ for μ a.e. and $\int f(x,\infty)d\mu(x) < 1$. Then $m_{\eta_n} \to m_\eta$ in the total variation sense would imply the tightness of the sequence $\{\eta_n : n \geq 1\}$. Moreover if η is identifiable we would have weak convergence of η_n to η.*

Proof. If you consider $\{\eta_n : n \geq 1\}$ as a sequence of probability measures in the extended space, we would have a subsequence $\{\eta_m\}$ such that $\eta_m \rightharpoonup \eta^*$. We will write η^* as $\eta^* = a\eta_0^* + (1-a)\delta_\infty$, where η_0^* is a probability measure on Θ. Note that this representation is unique. Below we shall show that a is equal to one, which shall imply the tightness of the whole sequence. That identifiability implies weak convergence can be shown by a standard subsequence argument. Assume that a is less than one. By upper semi-continuity of the function $f(x,\theta)$ for almost every x, we have

$$\limsup m_{\eta_m}(x) \leq a m_{\eta_0^*}(x) + (1-a)f(x,\infty)$$

But since $m_{\eta_n} \to m_\eta$ in the total variation sense, we have

$$\begin{aligned} & \int m_\eta(x)d\mu(x) \\ \leq \; & a\int m_{\eta_0^*}(x)d\mu(x) + (1-a)\int f(x,\infty)d\mu(x) \\ < \; & 1. \end{aligned}$$

This is a contradiction; hence the proof.

□

Remark 3.7 It is important to note that even though Θ may have a natural topology , it should be given the topology which makes the density functions $f(x,.)$ continuous. This is the natural topology for Θ in this problem. This is similar to what one does in the consistency proofs for the MLE, see Wald(1949). In this regard, Landers & Rogge(1972) is interesting reading.

Example 3.2 Let $P(.,\theta) = N(\theta,\sigma_0^2)$, $\pi^* = N(\mu_0,\eta_0^2)$ and $\pi = N(\mu,\eta)$. It is easy to see that

$$T_{\pi^*}(\pi) = N\left(\frac{\mu\sigma^2+\mu_0\eta^2}{\sigma^2+\eta^2}, \eta^2+\frac{\eta^4(\eta_0^2-\eta^2)}{(\eta^2+\sigma^2)^2}\right).$$

The conditions of the previous lemma are satisfied with $f(.,\infty)$ being defined to be the constant zero function. Hence condition (iii) of the theorem is satisfied. Conditions (i) and (ii) are easily seen to be satisfied too. Hence we have convergence of the iterations from the theorem. In this case, though, it is also easy to see from elementary considerations that the iterations do converge.

□

3.2 <u>Case 2</u> : Arbitrary probability measure m.

In this section we shall revert back to the original notation, T_m, for the operator under study. In this case the first question to answer is what are the iterations likely to converge to if they converge? The answer is that they are likely to converge to the π^* which satisfies

$$I_0(m_{\pi^*},m) = \inf_{\pi} I_0(m_\pi,m).$$

In other words π^* is such that m_{π^*} is the I_0 projection of m on $\mathcal{C}\{f(.,\theta) : \theta \in \Theta\}$. So let us first go ahead and make the assumption that such a π^* exists. We shall also assume that m dominates $\mathcal{C}\{f(.,\theta) : \theta \in \Theta\}$. Note that the part of the measures defined by $f(.,\theta)$ and not dominated by m does not play any role, but if we relax this assumption then we enter the case where we do not assume that the measures are probability measures but just finite measures, see the next sub-section. This assumption implies that π^* is unique when we have identifiability. But we do not assume identifiability.

The following lemma shows immediately that our "guess" is at least a fixed point of the operator T_m and hence a sensible "guess".

Lemma 3.4 *The above defined π^* is a fixed point of the operator T_m.*

PROOF. We have, by definition of π^* and lemma 2.5, that

$$\int \frac{m(x)}{m_{\pi^*}(x)} f(x,\theta) d\mu(x) \ \leq 1 \ \forall \theta.$$

Since

$$\int\int \frac{m(x)}{m_{\pi^*}(x)} f(x,\theta) d\mu(x) d\pi^*(\theta) \ = 1$$

we have

$$\int \frac{m(x)}{m_{\pi^*}(x)} f(x,\theta) d\mu(x) \ = 1 \ \pi^* a.e.$$

Hence, we have $T_m(\pi^*) = \pi^*$.

□

Let us define

$$\psi(\pi) = \int \log\left(\int \frac{m(x)}{m_\pi(x)} f(x,\theta) d\mu(x)\right) d\pi^*(\theta).$$

Lemma 3.5 *The function ψ is a well defined function on $\mathcal{P}(\Theta)$. Moreover, when π is such that $I_0(\pi, \pi^*) < \infty$, then $\psi(\pi)$ is finite.*

PROOF.

$$\begin{aligned} & \int \left[\log\left(\int \frac{m(x)f(x,\theta)}{m_\pi(x)} \cdot d\mu(x)\right)\right]^- d\pi^*(\theta) \\ = & \int \left[\log\left(\int \frac{m_{\pi^*}(x)}{m_\pi(x)} \cdot \frac{m(x)f(x,\theta)}{m_{\pi^*}(x)} d\mu(x)\right)\right]^- d\pi^*(\theta). \end{aligned}$$

Now, since

$$\int \frac{m(x)}{m_{\pi^*}(x)} f(x,\theta) d\mu(x) = 1 \ \pi^* a.e.$$

by Cauchy-Schwartz, we have

$$\begin{aligned} & \int \left[\log\left(\int \frac{m_{\pi^*}(x)}{m_\pi(x)} \cdot \frac{m(x)f(x,\theta)}{m_{\pi^*}(x)} d\mu(x)\right)\right]^- d\pi^*(\theta) \\ \leq & \int \left[-\log\left(\int \frac{m_\pi(x)}{m_{\pi^*}(x)} \cdot \frac{m(x)f(x,\theta)}{m_{\pi^*}(x)} d\mu(x)\right)\right]^- d\pi^*(\theta) \\ = & \int \log\left(1 \vee \int \frac{m_\pi(x)}{m_{\pi^*}(x)} \cdot \frac{m(x)f(x,\theta)}{m_{\pi^*}(x)} d\mu(x)\right) d\pi^*(\theta) \\ \leq & \int\int \frac{m_\pi(x)}{m_{\pi^*}(x)} \cdot \frac{m(x)f(x,\theta)}{m_{\pi^*}(x)} d\mu(x) d\pi^*(\theta) \\ = & \int \frac{m(x)m_\pi(x)}{m_{\pi^*}(x)} d\mu(x) \\ \leq & \ 1. \end{aligned}$$

The above implies that $\psi(\pi)$ is well defined.Let us assume that π is such that $I_0(\pi, \pi^*) < \infty$. Then we have, after some elementary algebra,

$$I_0(T_m(\pi), \pi^*) = I_0(\pi, \pi^*) - \psi(\pi).$$

Since $I_0(T_m(\pi), \pi) > -\infty$ we have $\psi(\pi) < \infty$.

□

Lemma 3.6 *The function ψ is non-negative on $\mathcal{P}(\Theta) \cap \{\pi : I_0(\pi, \pi^*) < \infty\}$ and, moreover,*

$$\psi(\pi) \geq I_0(m_\pi, m) - I_0(m_{\pi^*}, m).$$

The above, in particular, implies that $\psi(\pi)$ is zero iff $m_\pi = m_{\pi^*}$.

Proof. We have, from observations made above and Jensen's inequality, that

$$\begin{aligned}
&\psi(\pi) \\
&\quad= \int \log\left(\int \frac{m_{\pi^*}(x)}{m_\pi(x)} \cdot \frac{m(x)f(x,\theta)}{m_{\pi^*}(x)} d\mu(x)\right) d\pi^*(\theta) \\
&\quad\geq \int\int \log\left(\frac{m_{\pi^*}(x)}{m_\pi(x)}\right) \cdot \frac{m(x)f(x,\theta)}{m_{\pi^*}(x)} d\mu(x) d\pi^*(\theta) \\
&\quad= \int \log\left(\frac{m_{\pi^*}(x)}{m_\pi(x)}\right) dm(x) \\
&\quad= I_0(m_\pi, m) - I_0(m_{\pi^*}, m).
\end{aligned}$$

The last quantity is non-negative by definition of π^*.

□

Below we state an inequality which, in fact, is a generalization of the inequality given in lemma 3.2.

Theorem 3.2 *Let $\pi \in \mathcal{P}(\Theta) \cap \{\pi : I_0(\pi, \pi^*) < \infty\}$. Then*

$$I_0(T_m(\pi), \pi^*) \leq I_0(\pi, \pi^*) - (I_0(m_\pi, m) - I_0(m_{\pi^*}, m)).$$

PROOF. As we know when $\pi \in \mathcal{P}(\Theta) \cap \{\pi : I_0(\pi, \pi^*) < \infty\}$,

$$I_0(T_m(\pi), \pi^*) = I_0(\pi, \pi^*) - \psi(\pi).$$

The theorem now follows from the previous lemma.

□

Remark 3.8 To get lemma 3.2 from the previous theorem just equate m to m_{π^*}. Most of the remarks made immediately after lemma 3.2 apply here too.

Remark 3.9 It is important to note that in all the above results of this section we have never used the fact that $f(\cdot, \theta)$ define probability measures. We just needed that

$$\sup_{\theta \in \Theta} C(\theta) < \infty, \text{ where } C(\theta) = \int f(x, \theta) d\mu(x).$$

The next result is an exception.

Theorem 3.3 *Assume that the following hold, in addition to from the assumptions made in the start of the sub-section.*
(i) π *is such that* $I_0(\pi, \pi^*) \leq \infty$.
(ii) $m_{\pi_n} \rightarrow m_{\pi^*}$ *in total variation metric implies* $\pi_n \rightharpoonup \pi^*$.
Then $T_m^n(\pi) \rightharpoonup \pi^*$.

PROOF. From the above lemma we have

$$I_0(T_m^{n+1}(\pi), \pi^*) \leq I_0(T_m^n(\pi), \pi^*) - (I_0(m_{T_m^n(\pi)}, m) - I_0(m_{\pi^*}, m)) \ \forall \ n \geq 0.$$

So either $I_0(T_m^n(\pi), \pi^*)$ is decreasing to zero or $I_0(m_{T_m^n(\pi)}, m) \rightarrow I_0(m_{\pi^*}, m)$. In the former case, $T_m^n(\pi) \rightarrow \pi^*$ in the I_0 divergence sense, which by lemma 2.1 implies convergence in the total variation sense and hence, in particular, $T_m^n(\pi) \rightharpoonup \pi^*$. In the latter case we have by lemma 2.3 that $m_{T_m^n(\pi)} \rightarrow m_{\pi^*}$ in the total variation sense. Now by assumption *(ii)*, $T_m^n(\pi) \rightharpoonup \pi^*$.

□

Remark 3.10 Note that the assumption *(ii)* in the above theorem is satisfied if the conditions of Lemma 3.3 are satisfied.

Remark 3.11 As a corollary to the above theorem, when $f(x,\theta)$ is bounded as a function of θ for all x, we get $I_0(m_{T_m^n(\pi)}, m) \to I_0(m_{\pi^*}, m)$. This is what gives the assurance of convergence to the global maximum when applied to the EM algorithm.

We give a simple lemma which establishes the existence of π^* in the case when Θ is a compact metric space.

Lemma 3.7 *Assume that $f(x,\theta)$ is bounded continuous as a function of θ for almost every x w.r.t μ. Let $\phi(\pi) = I_0(m_\pi, m)$ for π a probability measure on $(\Theta, \mathcal{B})$. Then ϕ is lower semicontinuous on $\mathcal{P}(\Theta)$ w.r.t. the weak topology. Moreover, when Θ is a compact metric space, there exists a π^* such that*

$$I_0(m_{\pi^*}, m) = \inf_{\pi \in \mathcal{P}(\Theta)} I_0(m_\pi, m).$$

PROOF. First note that from Scheffe's theorem it follows that if $\pi_n \rightharpoonup \pi_0$ then $m_{\pi_n} \to m_{\pi_0}$ in $L^1(\mu)$. The lower semicontinuity of ϕ now follows from the lower semicontinuity of the I_0 divergence w.r.t. the total variation metric, see Lemma 2.2. The rest follows from the compactness of $\mathcal{P}(\Theta)$ w.r.t. the weak topology.

□

3.3 Case 3 : $\{f(.,\theta) : \theta \in \Theta\}$; Class of Finite Measures.

The main motivation of this sub section is to present results applicable to maximum likelihood estimation of mixtures. The other example that can be addressed is the maximization of log investment return. These will be discussed in the next section. We shall specialize directly to the case when Θ is a finite set. Let us denote the elements of Θ by $\{\theta_i : 1 \leq i \leq K\}$ for some K.

Theorem 3.4 *Let π be a probability measure with support Θ and m be a probability measure which dominates $\{f(.,\theta) : \theta \in \Theta\}$. Let $\mathcal{T} = \mathcal{C}(\{f(.,\theta) : \theta \in \Theta\})$. Assume that $I_0(\mathcal{T}, m) < \infty$. Then $T_m^n(\pi)$ converges and the limit point, say π^*, is such that $I_0(\mathcal{T}, m) = I_0(m_{\pi^*}, m)$.*

PROOF. Since Θ is a finite set the condition of remark 3.9 is satisfied. Also since $\mathcal{P}(\Theta)$ is compact, so is $\mathcal{T}$ and, we have the existence of a unique m^* such that $I_0(m^*, m) = \inf_{\eta \in \mathcal{P}(\Theta)} I_0(m_\eta, m)$. Let $\pi^* \in \mathcal{P}(\Theta)$ be such that $m^* = m_{\pi^*}$. Now since $\mathcal{P}(\Theta)$ is compact there exists a subsequence $\{l\}$ of $\{n\}$ such that $T^l_m(\pi) \to \pi'$, for some $\pi' \in \mathcal{P}(\Theta)$. From theorem 3.2 we get either $T^n_m(\pi) \to \pi^*$ or $I_0(m_{T^l_m(\pi)}, m) \to I_0(m^*, m)$. In the former case we are done. Hence let us suppose the latter occurs. From remark 2.1 we get that

$$\frac{dm_{T^l_m(\pi)}}{dm} \to^m \frac{dm^*}{dm}.$$

But since $T^l_m(\pi) \to \pi'$ we have $m_{\pi'} = m^*$. Now apply theorem 3.2 with π' instead of π^*. Then we get that $I_0(T^n_m(\pi), \pi')$ is non-increasing and since $I_0(T^l_m(\pi), \pi') \downarrow 0$ (since Θ is finite) we have $I_0(T^n_m(\pi), \pi') \downarrow 0$. Hence the proof.

□

Remark 3.12 The assumption on the support of π is an obvious necessity. Since we are working with Θ as a finite set, the convergence of $T^n_m(\pi)$ is in all possible senses. Note also that we have made no assumption of identifiability above.

3.4 Application

We shall discuss the problem of estimating the finite mixture using the method of maximum likelihood estimation when the distribution of individual types is completely known.

Let us suppose that $X_1, X_2, ..X_n$ is an iid sample from m_π, where

$$m_\pi = \int_\Theta f(\cdot, \theta) d\pi(\theta)$$

and π is unknown. We wish to estimate π from the data using the principle of maximum likelihood estimation. We assume that Θ is a finite set and that the μ densities $f(\cdot, \theta)$ are completely known. So the problem is to find a π^* such that

$$\prod_{i=1}^{i=n} m_{\pi^*}(x_i) = \sup_{\pi \in \mathcal{P}(\Theta)} \prod_{i=1}^{i=n} m_\pi(x_i).$$

This is the same as finding π^* such that

$$\sum_{i=1}^{i=n} \log(m_{\pi^*}(x_i)) \;=\; \sum_{i=1}^{i=n} \log(m_{\pi}(x_i)).$$

The above π^* can now be immediately recognized as one such that m_{π^*} is the I_0 projection F_n, the empirical distribution function, on the space of finite measures on the set $\{x_i : 1 \leq i \leq n\}$ given by the convex hull of the set of measures $\{(f(x_i, \theta))_{1 \leq i \leq n} : \theta \in \Theta\}$. This is so because

$$I_0(m_\pi, F_n) \;=\; \frac{1}{n} \sum_{i=1}^{i=n} -\log(m_\pi(x_i)) - \log(n).$$

It is important to note that in the above expression we are not interpreting m_π as μ absolutely continuous measures but measures on the finite set $\{x_i : 1 \leq i \leq n\}$. Now it is clear that theorem 3.4 applies and it implies that, by starting with any π which gives mass to every θ, one can by the iterations of the operator T_{F_n} on π converge to one such π^*. Uniqueness of the limit point holds under the assumption of identifiability. That m_{π^*} is an MLE follows from an argument similar to that in remark 3.11.

4 Data Augmentation Method

4.1 Conditions and Preliminary Results

Denote, by π_0, the initial starting point of the iteration. Assume that there exists a fixed point of the operator W, whose density w.r.t. ν shall be denoted by π^*. The conditions that we need are as follows:

(a) $M = \sup_\theta \frac{\pi_0(\theta)}{\pi^*(\theta)} < \infty$.

(b) $f(.,.)$ is a strictly positive and bounded function.

(c) f is continuous in each argument.

(d) $g(\theta, x)$ is a bounded function in x for each θ.

(e) π_0 satisfies the condition $\int \log \frac{\pi^*(\theta)}{\pi_0(\theta)} \pi^*(\theta)\ d\nu(\theta) \quad < +\infty$.

Remark 4.1 *Condition (e) can be weakened to condition (é), which is given below.*

(é) There exists an n_0 such that $\int \pi^*(\theta) log(\frac{\pi^*(\theta)}{W^{n_0}(\pi_0)(\theta)})\ d\nu(\theta) \quad < +\infty$

We shall list needed preliminary facts that can be deduced from these conditions. Below we shall denote by, m_h, the μ density

$$m_h(x) = \int f(x,\theta)\ h(d\theta) \qquad \forall x \in \mathcal{X},$$

where h is any measure on $(\Theta, \mathcal{B})$. We shall in this paper, find occasions when it is convenient to talk in terms of either the measure or the density, and shall do so without using different notations. By $\eta_{\pi_n}(.,.)$, we shall denote the posterior of the measure given by the $\mu \times \nu$ density $f(.,.) \cdot \pi_n(.)$.

Lemma 4.1 *The following are true.*

(i) For each $n \geq 0$,

$$\sup_x \frac{m_{W^n(\pi_0)}(x)}{m_{\pi^*}(x)} \leq M \quad and \quad \sup_\theta \frac{W^n(\pi_0)(\theta)}{\pi^*(\theta)} \leq M.$$

(ii) The sequence of probability measures $\{W^n(\pi_0)\}_{n\geq 0}$ is tight.

(iii) For any h, a ν density, m_h is a strictly positive continuous function and $W(h)$ has a strictly positive ν density.

Proof.

(i) The argument uses some elementary manipulations and induction.

(ii) This follows from (i) and the fact that a single probability measure on a Polish space is tight, see Parthasarathy (1967).

(iii) Follows from conditions (b) and (c) above.

□

In what follows, we shall denote $W^n(\pi_0)$ by π_n for all $n \geq 0$.

Lemma 4.2 *For every subsequence of* $\{\pi_n\}_{n\geq 0}$, *we can get a further subsequence which converges in the total variation metric. Moreover, if this subsequence is denoted by* $\{\pi_m\}_{m\in\mathcal{I}}$ *and the limiting* ν *density is denoted by* h, *then* h *is strictly positive and* m_{π_m} *converges in* L_1 *distance, and uniformly, to* m_h.

PROOF. Let $\{\pi_m\}_{m\in\mathcal{I}_0}$ be any given subsequence. Let us consider the subsequence $\{\pi_m\}_{m\in\mathcal{I}_0-1}$ (ignoring the index -1, in the case when $\mathcal{I}$ contains 0). Now this subsequence is tight by the lemma above; hence we have a subsequence which converges in the weak topology. Let the limiting measure be denoted by β and the subsequence by $\{\pi_m\}_{m\in\mathcal{J}}$. Note that, by virtue of conditions (b) and (c) above,

$$m_{\pi_m}(x) = \int f(x,\theta)\pi_m(\theta)\, d\nu(\theta) \;\longrightarrow\; \int f(x,\theta)\; \beta(d\theta) \;\text{ as } m \to \infty, m \in \mathcal{J}.$$

Hence we have, using condition (d), that, for each θ,

$$W(\pi_m)(\theta) = \int g(\theta,x)m_{\pi_m}(x)\; d\mu(x) \quad\longrightarrow\quad \int g(\theta,x)m_\beta(x)\; d\mu(x) \;\text{ as } m \to \infty, m \in \mathcal{J}.$$

From Scheffe's theorem we have the L_1 convergence of $W(\pi_m)$ to $W(\beta)$, as $m \to \infty, m \in \mathcal{J}$. Denote the ν density of $W(\beta)$ by h. That h is strictly positive follows from Lemma 4.1 (iii). But note that $\{W(\pi_m)\}_{m\in\mathcal{J}}$ is a subsequence of $\{W(\pi_m)\}_{m\in\mathcal{I}_0}$. Hence, denoting $\mathcal{J}+1$ by $\mathcal{I}$, we have that $\{\pi_m\}_{m\in\mathcal{I}}$ is a required such sequence. The second assertion in the lemma follows by using condition (b), the first assertion and another application of Scheffe's theorem. □

4.2 The New Approach

Our approach is based on the fact that the operator W can be described as an operator which is derived from composing two 'projections', in quotes because the divergence is not even a metric. Let $\mathcal{C}_f$ denote the set of all probability measures on the product measure space $(\mathcal{X} \times \Theta, \mathcal{A} \otimes \mathcal{B})$, defined by

$$\mathcal{C}_f = \{P : P \text{ has conditional } \mu \text{ density } f(.,\theta) \text{ given } \{\phi, \mathcal{X}\} \otimes \mathcal{B}\}.$$

Similarly, we shall denote by $\mathcal{C}_g$, the set of all probability measures on $(\mathcal{X} \times \Theta, \mathcal{A} \otimes \mathcal{B})$, defined by

$$\mathcal{C}_g = \{P : P \text{ has conditional } \nu \text{ density } g(.,x) \text{ given } \mathcal{A} \otimes \{\phi, \Theta\}\}.$$

The theorem below says that, when we operate W on π_n for some $n \geq 0$, we are essentially projecting the measure with a $\mu \times \nu$ density $f(.,.) \cdot \pi_n$ on $\mathcal{C}_g$ and then projecting this resulting measure on $\mathcal{C}_f$.

Theorem 4.1 *For any $n \geq 0$,*

(i) The I_0 projection of the probability with $\mu \times \nu$ density $f(.,.) \cdot \pi_n$ on $\mathcal{C}_g$ is the measure with the x-marginal μ density m_{π_n}.

(ii) The I_0 projection of the probability with $\mu \times \nu$ density $g(.,.) \cdot m_{\pi_n}$ on $\mathcal{C}_f$ is the measure with the θ-marginal ν density π_{n+1}.

PROOF. It is important to note that $\mathcal{C}_f \bigcap \mathcal{C}_g$ contains the probability measure with $\mu \times \nu$ density $f(.,.) \cdot \pi^*$ (or $g(.,.) \cdot m_{\pi^*}$).

(i) The divergence between $f(.,.) \cdot \pi_n$ and $f(.,.) \cdot \pi^*$ is finite, as it is equal to $I_0(\pi^*, \pi_n)$, which in turn is finite as a consequence of Lemma 4.1 (i). Note that the above divergence is also equal to

$$\begin{aligned}
& I_0(g(.,.) \cdot m_{\pi^*}, f(.,.) \cdot \pi_n) \\
= \;& I_0(g(.,.) \cdot m_{\pi^*}, \eta_{\pi_n}(.,.) \cdot m_{\pi_n}) \\
= \;& \int \log\left(\tfrac{\eta_{\pi_n}(\theta,x) \cdot m_{\pi_n}(x)}{g(\theta,x) \cdot m_{\pi^*}(x)}\right) \eta_{\pi_n}(\theta, x) \cdot m_{\pi_n}(x) \; d\mu(x) d\nu(\theta) \\
= \;& \int \log\left(\tfrac{\eta_{\pi_n}(\theta,x)}{g(\theta,x)}\right) \; \eta_{\pi_n}(\theta, x) \cdot m_{\pi_n}(x) \; d\mu(x) d\nu(\theta) + I_0(m_{\pi^*}, m_{\pi_n}) \\
= \;& \int I_0(g(.,x), \eta_{\pi_n}(.,x)) \; dm_{\pi_n}(x) + I_0(m_{\pi^*}, m_{\pi_n}).
\end{aligned}$$

Note that the second term is finite and the left hand side was observed above to be finite, hence we have the finiteness of the first term. Now let us look at the projection problem. Note that it is sufficient to minimize the Kullback-Leibler divergence in the sub-class of measures which have a m_{π_n} absolutely continuous x-marginal, as otherwise the divergence would be $+\infty$. Let $g(.,.)\cdot m(.)$ be any arbitrary element of this $\mathcal{C}_g$. Then the divergence of this measure from $f(.,.)\cdot\pi_n$ would be

$$\begin{aligned} & I_0(g(.,.)\cdot m(.), f(.,.)\cdot\pi_n) \\ = \ & \textstyle\int I_0(g(.,x), \eta_{\pi_n}(.,x))\ dm_{\pi_n}(x) + I_0(m, m_{\pi_n}) \end{aligned}$$

by similar manipulations as above. Hence the search for the projection reduces to minimizing the second term w.r.t. m, which is obviously minimized at m_{π_n}.

(ii) The second assertion is proved in the same way as above, the symmetry being clear.

□

4.3 Main Results

The following theorem gives the crucial inequality on which this section is based. The intuition for the inequality from the previous subsection.

Theorem 4.2 *For all $n \geq 0$,*

$$I_0(\pi_{n+1}, \pi^*) \leq I_0(\pi_n, \pi^*) - I_0(\int \eta_{\pi_n}(.,x) m_{\pi^*}(x)\ d\mu(x), \pi^*)$$

PROOF. We shall prove this only for for $n = 0$. From the proof and Lemma 4.1 (i) it is clear that the result holds for all $n \geq 0$. Note that

$$I_0(\pi_1, \pi^*) = I_0(\pi_0, \pi^*) + \int \log(\frac{\pi_0(\theta)}{\pi_1(\theta)})\pi^*(\theta)\ d\nu(\theta).$$

We now deal with the second term of the above equation.

$$\begin{aligned} & \textstyle\int \log(\frac{\pi_0(\theta)}{\pi_1(\theta)})\pi^*(\theta)\ d\nu(\theta) \\ = \ & \textstyle\int \log(\frac{\pi_0(\theta)}{\pi^*(\theta)\cdot\int f(x,\theta)\cdot\frac{m_{\pi_0}(x)}{m_{\pi^*}(x)}\ d\mu(x)})\pi^*(\theta)\ d\nu(\theta) \\ \leq \ & \textstyle\int \log(\frac{\pi_0(\theta)\cdot\int f(x,\theta)\cdot\frac{m_{\pi^*}(x)}{m_{\pi_0}(x)}\ d\mu(x)}{\pi^*(\theta)})\pi^*(\theta)\ d\nu(\theta) \end{aligned}$$

$$\begin{aligned} &= \quad \textstyle\int \log\left(\frac{\int \eta_{\pi_0}(\theta,x) m_{\pi^*}(x)\ d\mu(x)}{\pi^*(\theta)}\right)\pi^*(\theta)\ d\nu(\theta) \\ &= \quad -I_0(\textstyle\int \eta_{\pi_n}(.,x) m_{\pi^*}(x)\ d\mu(x), \pi^*). \end{aligned}$$

The inequality above follows by the use of Jensen's inequality. Combining we get

$$I_0(\pi_1, \pi^*) \leq I_0(\pi_0, \pi^*) - I_0(\int \eta_{\pi_0}(.,x) m_{\pi^*}(x)\ d\mu(x), \pi^*).$$

□

Theorem 4.3 *If*

$$\lim_{n \to \infty} I_0(\int \eta_{\pi_n}(.,x) m_{\pi^*}(x)\ d\mu(x), \pi^*) = 0,$$

then we have L_1 convergence of the iterates to π^*, i.e.

$$\lim_{n \to \infty} \|W^n \pi_0 - \pi^*\| = 0.$$

PROOF. Let $\{\pi_m\}_{m \in \mathcal{J}}$ be any arbitrary sequence. We shall prove that the above has a further subsequence converging in L_1 topology to π^*. Let $\{\pi_m\}_{m \in \mathcal{J}'}$ be the subsequence with index set $\mathcal{J}' = \mathcal{J} - 1$, with deletion of negative indices if necessary. From Lemma 4.2 and Lemma 4.1 (iii), we have for the subsequence $\{\pi_m\}_{m \in \mathcal{J}'}$, a further subsequence $\{\pi_m\}_{m \in \mathcal{I}}$ and a ν density h satisfying

(i) $\lim_{m \to \infty, m \in \mathcal{I}} \|\pi_m - h\| = 0$.

(ii) $\lim_{m \to \infty, m \in \mathcal{I}} \|m_{\pi_m} - m_h\| = 0$ and $m_{\pi_m} \to m_h$ uniformly.

(iii) m_h and m_{π_m} are continuous and strictly positive.

(iv) h and π_m are strictly positive.

In the sequel it will be assumed that the index always belongs to the set $\mathcal{I}$. Consider the ν densities $\eta_{\pi_m}(\theta, x)$ and $\eta_h(\theta, x)$. We have pointwise convergence for each x and θ respectively, from (i) & (ii) above. Since these are densities, Scheffe's theorem yields

$$\lim_{m \to \infty, m \in \mathcal{I}} \|\eta_{\pi_m}(.,x) - \eta_h(.,x)\| = 0, \quad \forall x$$

Moreover, the above limit is uniform on compact subsets of $\mathcal{X}$. To prove this, let $K \subseteq \mathcal{X}$ be a compact set. By using (ii) & (iii), we can choose a $c > 0$ and $N > 0$ such that

$$m_{\pi_m}(x) > c \;\; \forall\, x \in K, \;\; m \in \mathcal{I} \; m \geq N.$$

By using the uniform boundedness of $f(.,.)$, the above fact and some elementary manipulations one can get

$$\lim_{m\to\infty, m\in\mathcal{I}} \|\eta_{\pi_m}(.,x) - \eta_h(.,x)\| \;= 0, \quad \text{uniformly on compacts of } \mathcal{X}.$$

For convenience, denote by Q_m the measure

$$\int \eta_{\pi_m}(.,x) m_{\pi^*}(x) \; d\mu(x)$$

and by Q the measure

$$\int \eta_h(.,x) m_{\pi^*}(x) \; d\mu(x).$$

Next, we shall prove that

$$\lim_{m\to\infty} \|Q_m - Q\| = 0.$$

Fix an $\epsilon > 0$ and let $\epsilon' = \epsilon/4$. Let K be a compact set such that $m_{\pi^*}(K) < \epsilon'$. Choose N large such that

$$\|\eta_{\pi_m}(.,x) - \eta_{\pi^*}(.,x)\| < \epsilon' \;\; \forall x \in K, \forall m \geq N.$$

Then for an arbitrary set $A \subseteq \Theta$ and $m \geq N$,

$$\begin{aligned}
& |Q_m(A) - Q(A)| \\
= \;& |\textstyle\int \eta_{\pi_m}(A,x) m_{\pi^*}(x) \; d\mu(x) - \int \eta_h(A,x) m_{\pi^*}(x) \; d\mu(x)| \\
\leq \;& |\textstyle\int_K \eta_{\pi_m}(A,x) m_{\pi^*}(x) \; d\mu(x) - \int_K \eta_h(A,x) m_{\pi^*}(x) \; d\mu(x)| \\
& + |\textstyle\int_{K^c} \eta_{\pi_m}(A,x) m_{\pi^*}(x) \; d\mu(x) - \int_{K^c} \eta_h(A,x) m_{\pi^*}(x) \; d\mu(x)| \\
\leq \;& \textstyle\int_K |\eta_{\pi_m}(A,x) - \eta_h(A,x)| m_{\pi^*}(x) \; d\mu(x) + 2 \cdot \epsilon' \\
\leq \;& \textstyle\int_K \|\eta_{\pi_m}(.,x) - \eta_h(.,,x)\| m_{\pi^*}(x) \; d\mu(x) + 2 \cdot \epsilon' \\
\leq \;& 3 \cdot \epsilon' < \epsilon.
\end{aligned}$$

Hence we have L_1 convergence of Q_m to Q. But this and the hypothesis of the theorem, along with the lower semicontinuity of the divergence w.r.t. the total variation distance, Vajda (1989), implies that

$$I_0(Q,\pi^*) = 0 \Rightarrow \int \eta_h(.,x) m_{\pi^*}(x)\ d\mu(x) = \pi^*\ [\nu].$$

But, this implies that

$$h(\theta) \int f(x,\theta) \frac{m_{\pi^*}(x)}{m_h(x)}\ d\mu(x) = \pi^*(\theta)\ [\nu].$$

Hence,

$$\frac{\pi^*(\theta)}{h(\theta)} = \int f(x,\theta)\frac{m_{\pi^*}(x)}{m_h(x)}\ d\mu(x) \geq \left(\int f(x,\theta)\frac{m_h(x)}{m_{\pi^*}(x)}\ d\mu(x)\right)^{-1}.$$

The last inequality follows from Jensen's inequality. The above inequality implies

$$\int f(x,\theta)\frac{m_h(x)}{m_{\pi^*}(x)}\ d\mu(x) \geq \frac{h(\theta)}{\pi^*(\theta)}.$$

Integrating the above w.r.t. π^* measure yields one on both sides. Hence this implies that all along the inequality was an equality, for almost all ν. This implies $m_{\pi^*} = m_h\ [\mu]$. Hence, $W(h) = \pi^*$. But, since $\{W(\pi_m)\}_{m\in\mathcal{I}}$ is a subsequence of $\{\pi_m\}_{m\in\mathcal{J}}$, by L_1 continuity of W there exists a subsequence with L_1 limit π^*. Hence,

$$\lim_{n\to\infty} \|W^n\pi_0 - \pi^*\| = 0.$$

□

Theorem 4.4 *Under the conditions (a)-(e),*

$$\lim_{n\to\infty} \|W^n\pi_0 - \pi^*\| = 0.$$

PROOF. From condition (e) and Theorem 4.2, $I_0(\pi_n,\pi^*)$ is non-increasing in n. This implies the existence of the limit of $I_0(\pi_n,\pi^*)$ as $n \to \infty$. Let this limit be c. Now c is either strictly positive or zero. If it is zero, then we are done, as convergence in the Kullback-Leibler distance implies L_1 convergence. If it is positive then we get from Theorem 4.2 that

$$\lim_{n\to\infty} I_0\left(\int \eta_{\pi_n}(.,x) m_{\pi^*}(x)\ d\mu(x), \pi^*\right) = 0.$$

This, by Theorem 4.3, implies L_1 convergence of the iterates. Hence the proof. □

Acknowledgement

This paper arose from a problem suggested by Professors James O. Berger and William E. Strawdermann. The author is indebted to them for their encouragement and support without which this article would not have seen it's present form.

References

Bahadur, R. R. (1971). *Some Limit Theorems in Statistics.* **SIAM**, Philadelphia.

Boyles, R.A. (1983). On the convergence of the EM algorithm. *J. Statist. Plann. Inference* **45**, No. 1, 47-50.

Cover, T.M. (1984). An algorithm for maximizing expected log investment. *IEEE transactions in Information theory.* **IT-30**, No. 2, 369-373.

Csiszár, I. (1964). Über topologische und metrische Eigenschaften der relativen Information der Ordnung α. *Trans. Third Prague Conference of Information Theory etc.* Publishing House of the Czehoslovak Academy of Sciences, Prague, 63-73.

Csiszár, I. (1967). Information type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica.* **2**, 299-318.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3**, 1, 146-158.

Csiszár, I. & Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statist. Decisions* **Suppl. 1**, 205-237.

Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). Maximum Likelihood from Incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Dykstra, R. L. (1985). An iterative procedure for obtaining I-projections onto the intersection of convex sets. *Ann. Probab.* **13**, No. 3, 975-984.

Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions (with discussion). *Ann. Math. Statist.* **36**, 369-408.

Jeffreys, H. (1948). *Theory of Probability.* **2**nd ed. Clarendon Press, Oxford.

Kemperman, J.H. (1967). On the optimum rate of transmitting information. In *Prob. and Inf. Th.*. 126-129. Lecture Notes in Mathematics. Springer-Verlag: Berlin.

Kullback, S. (1959). *Information Theory and Statistics.* Wiley, New York.

Kullback, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Trans. Information Theory.* **13**, 126-127.

Landers, D. & Rogge, L. (1972). Characterization of the topologies used in the theory of maximum likelihood estimation. *Z. Wahrsch. Verw. Gebiete.* **21**, 197-200.

Liese, F. (1975). On the existence of *f*-projections. Colloq. Math. Soc. J. Bolyai. **16**, Budapest, 431-446.

Liese, F. & Vajda, I. (1987). *Convex Statistical Distances.* Teubner, Leipzig.

Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces.* Academic Press, New York.

Redner, R.A. & Walker, H.F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review.* **26**, No. 2, 195-239.

Rüschendorf, L. (1984). On the minimum information discrimination theorem. *Statist. Decisions* **Suppl. 1.**

Tanner, M. A. & Wong, W. H. (1987). The calculation of the posterior distribution data augmentation. *J. Amer. Statist. Assoc.* **82**, No. 398, 528-540.

Vajda, Igor. (1989). *Theory of Statistical Inference and Information.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimators. *Ann. Math. Statist.* **20**, 595-601.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, No. 1, 95-103.

Zangwill, W.I. (1969). *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, New Jersey.

Department of Statistics
Purdue University
West Lafayette, Indiana 47907