Two Iterative Algorithms for Finding Minimax Solutions

CHEIN-I CHANG, MEMBER, IEEE, AND LEE D. DAVISSON, FELLOW, IEEE

Abstract—Two iterative minimax algorithms are presented with associated convergence theorems. Both algorithms consists of iterative procedures based on a sequence of finite parameter sets; in general these finite parameter sets are subsets of an infinite parameter space. To show their applicabilities, several commonly used examples are presented. It is also shown that minimax problems with or without finite parameter sets can be solved by these two algorithms numerically to any assigned degree of accuracy.

I. INTRODUCTION

MINIMAX problems in statistical decision theory have received considerable attention over the years [1]–[5]. The importance of such studies resides in the fact that the prior knowledge of observations obtained from experiments usually are incomplete and insufficient; the minimax risk simply provides the best performance under worst case conditions. However, most of the work in today's minimax theory is analytic; little has been done numerically. To the authors' best knowledge, Nelson's paper [6] is the only work devoted to solving general minimax problems by iteration methods. Unfortunately, Nelson proved the existence of such methods but failed to propose constructive schemes for finding solutions. Since present-day digital computers can perform sophisticated and complicated calculations, developing computational algorithms for finding numerical solutions is increasingly important.

We present two iterative algorithms for solving minimax problems. The problem to be considered is formulated as a general statistical decision game so that the two algorithms can be applied to diverse areas including communications and control. The algorithms (to be called Algorithms I and II henceforth) consist of iterative procedures based on a sequence of finite parameter sets; these finite parameter sets are subsets of a parameter space that may or may not be finite. The associated convergence theorems show that the algorithms eventually converge to the same minimax value. The difference between these two algorithms is that Algorithm I iterates on a sequence of parameter sets with fixed size; while Algorithm II iterates with varying sizes. More precisely, at each iteration, both algorithms find all possible locally maximizing points of the risk function over the original parameter space with respect to the minimax rule obtained at that iteration. Then to generate a new parameter set for the next iteration, Algorithm I replaces those parameters in the current parameter set with low probabilities by those locally maximizing points just found at that iteration. Instead of replacing points as does Algorithm I, Algorithm II simply collects and includes these locally maximizing parameters in the present parameter set to generate the next parameter set. Both algorithms produce a monotonically increasing sequence of estimated minimax values which eventually converges to the desired minimax risk.

In general, the performance of Algorithm II is superior to Algorithm I on the basis of the advantage that the size of the initial parameter set can be chosen arbitrarily. Computationally, however, Algorithm II is inferior to Algorithm I because, while Algorithm I uses a buffer of fixed size to store the data, Algorithm II needs more storage to accommodate the increasing data after an iteration cycle and thus requires more computations. Moreover, it will be shown that, if the risk function has a finite number of locally maximizing points for all nonequalizer decision rules N_0, then by choosing an initial parameter set with size sufficiently enough both algorithms converge and yield essentially the same performance.

Since the finiteness of N_0 is practically true and generally satisfied for many common probability distribution, this condition will be assumed throughout the paper. As shown in a step of the proof of convergence for Algorithm I under this assumption, a new property (to be called the Bayesian transistivity property or BTP) is introduced and also studied by some examples for illustration. The Bayesian transitivity property is necessary and very important because it determines the size of an initial parameter for Algorithm I. To extend the BTP to cover more general cases, a general theory (stated in the Appendix) is proven which says that, for a given error tolerance $\epsilon > 0$, if a condition probability distribution $p(x|\theta)$, $\theta \in \Theta$ is continuous and Θ is compact, then there exists a probability distribution $\tilde{p}(x|\theta)$, a polynomial approximation of p, such that the minimax risk using p and $\tilde{p}$ differ by no more than ϵ. With any polynomial conditional distribution and jointly continuous loss function, the associated risk function, $r(\cdot|\cdot)$ will have at most N_0 local maxima and the BTP can be proven satisfied for continuous probability

Manuscript received July 17, 1987; revised February 23, 1988.

L. D. Davisson is with the Electrical Engineering Department, University of Maryland, College Park, MD 20742.

C.-I. Chang is with the Electrical Engineering Department, University of Maryland, Catonsville, MD 21228.

IEEE Log Number 8933118.

distributions. As a result, this theorem can be applied to a variety of Bayes problems to reduce computational complexity. It is believed that in most practical problems, Algorithm I will be applicable and the storage we really need is generally much less than that for Algorithm II although the Bayesian transistivity property must be justified beforehand for Algorithm I.

In essence, Algorithm I originates in an algorithm which was used for source matching problems [7] and is a generalization suited for general statistical decision problems. Algorithm II is a slightly different version of Algorithm I. The idea was implicitly used in one of Nelson's examples [6]. He showed that a sequence of prior distributions for a fixed parameter space can be generated, and further proved that this sequence converges to a least favorable distribution. Unfortunately, he did not explicitly construct this sequence. Therefore, in this paper, a modified Nelson's algorithm for a fixed parameter space is described in which a sequence of successive prior distribution can be generated until a least favorable distribution is found. Then including this modified version of Nelson's algorithm as a subalgorithm, Algorithms I and II can actually produce a desired least favorable distribution throughout iterations.

Finally, two numerical examples are presented to exemplify the relative performance of Algorithms I and II. According to the numerical results, two algorithms have different advantages and disadvantages, and so it cannot be concluded that one is better than the other. The preference really depends upon applications.

This paper is organized as follows. In Section II, Algorithm I is stated, and its convergence theorem is proven. In Section III, Algorithm II is given, and its convergence theorem is also proven with a slightly different approach. In Section IV, the Bayesian transitivity property is defined and studied. Two approaches for proving this property are proposed. In Section V, two numerical examples are given for cases of two different performance criteria—are relative entropy loss and squared error loss. These examples provide evidence that Algorithms I and II indeed have respective advantages in different applications. In the Appendix, a general theorem is proven for an extension of the Bayesian transistivity property to a broader class including continuous probability distributions.

II. An Algorithm for Solving Minimax Solutions Based on a Sequence of Fixed-Size Parameter Sets

We shall consider a general statistical decision problem (Θ, D, R) of fixed sample size and also establish the following definitions and assumptions described in [6].

1) Let $\boldsymbol{X}$ be a random variable with observations x in a sample space X, and let $B(X)$ be the Borel field of X. There exists a σ-finite measure μ on $B(X)$ such that, for $\theta \in \Theta$, a probability distribution $\boldsymbol{P}_\theta$, has a density $p(x|\theta)$ with respect to μ and $p(\cdot|\cdot)$ is measurable on the σ-field $B(X) \times B(\Theta)$ where $B(\Theta)$ is the Borel field of Θ. Furthermore, for each $x \in X$, $p(x|\cdot)$ is continuous on Θ.

2) Let $\boldsymbol{C}$ be an essentially complete compact class of decision rules for the game $(\Theta, \boldsymbol{D}, R)$.

3) The parameter space Θ and the action space A are compact.

4) The loss function $L(\cdot,\cdot)$ is real-valued and jointed continuous on $\Theta \times A$. Consequently, L is bounded and uniformly continuous on $\Theta \times A$.

5) Let τ be any prior distribution over Θ. For any $x \in X$, except possibly on a set of μ-measure 0, there is at most one decision $d(x) \in A$ that minimizes $\int_\Theta L(\theta, d(x)) p(x|\theta) \tau(\theta) \mathrm{d}\theta$.

6) For any *a priori* distribution τ that does not satisfy 5), $r(\tau, \delta_\tau) = 0$, where $r(\tau, \delta_\tau)$ is the Bayes risk with respect to the prior distribution τ, whereas, $\sup_\alpha \{r(\alpha, \delta_\alpha)\} > 0$.

7) Let J be the size of the given initial parameter set for Algorithm I, $N_b \equiv \arg[\max_{\theta \in \Theta} R(\theta, \delta)]$ be the set of parameters which maximize $R(\theta, \delta)$ locally, and $\boldsymbol{N}_0 = \sup_{\delta \in \boldsymbol{D}'} |\boldsymbol{N}_\delta| < \infty$ where $\boldsymbol{D}'$ is the set of all nonequalizer rules in $\boldsymbol{D}$. Note that if there is an equalizer rule δ_0 in $\boldsymbol{D}$, then $R(\theta, \delta_0)$ is flat over Θ and thus, if $\boldsymbol{N}_0$ were defined on $\boldsymbol{D}$ instead of $\boldsymbol{D}'$, we would have $\boldsymbol{N}_0 = \infty$. In this case and if we know δ_0, we can check whether or not it is a minimax rule. Otherwise, we will assume that $\infty > J \geq \boldsymbol{N}_0$.

Here are some comments on the assumptions.

a) Assumption 6 prevents the algorithms from dealing with pathological cases. This was discussed in [6].

b) Since both algorithms operate on finite parameter sets, it is desirable to assume $\boldsymbol{N}_0 < \infty$. This is not a restrictive assumption because in most practical cases this assumption will be satisfied by invoking the Stone–Weierstrass approximation theorem which states that any continuous function on a compact set can be approximated uniformly by a sequence of finite degree polynomials to any degree of accuracy.

c) As an example, let $\Theta = A = [0,1]$, $X = \{0,1\}$, $p(x|\theta) = \theta^x (1-\theta)^{1-x}$, and $L(\theta, a) = (\theta - a)^2$. Then for each θ and each nonrandomized decision rule d, the risk function is given by

$$\begin{aligned} R(\theta, d) &= E_\theta[L[\theta, d(x)]] \\ &= (d(0)-\theta)^2(1-\theta) + (d(1)-\theta)^2\theta \\ &= (1 + 2d(0) - 2d(1))\theta^2 \\ &\quad + (d^2(1) - d^2(0) - 2d(0))\theta + d^2(0). \end{aligned} \tag{2.1}$$

Clearly, $\boldsymbol{N}_0 \geq 2$ because we can find a nonrandomized decision rule $d(x)$ defined by $d(0) = d(1) = 1/2$ so that the maximizing points of $R(\theta, d)$ over $[0,1]$ are $\{0,1\}$; on the other hand, from (2.1), $\boldsymbol{N}_0 \leq 2$ since for each nonrandomized rule d the risk function R is quadratic in θ. Therefore, $N_0 = 2$. (Notice that from (2.1) it is easy to see that there exists only one equalizer rule d_0: $\{d_0(0) = 1/4, d_0(1) = 3/4\}$, which happens to be the minimax rule such that if N_0 were defined over $\boldsymbol{D}$ instead of $\boldsymbol{D}'$, $\boldsymbol{N}_0 = \infty$.) If we choose $J = 2$ in this example, it is shown in [8] that, with a choice of an initial parameter set of size 2, Algorithm I does not work. Nevertheless, it does work for

$J \geq 3$. This simply means that the size of an initial parameter set, $J = 2$, is too small and is not equal to N_0. In other words, we have to pick enough parameters for initialization before we execute Algorithm I. As also shown in [8], Algorithm II does not have this defect. More details on this example were discussed in [8].

A. Description of Algorithm I

The following algorithm is used in Algorithm I (and Algorithm II) as a subalgorithm to generate a least favorable distribution on a fixed finite parameter space and to find the corresponding minimax risk.

Modified Nelson's Algorithm (Based on a Finite Parameter Set Θ^n):

1) Initialization: Given a parameter set $\Theta^n = \{\theta_1^n, \cdots, \theta_J^n\}$ which is the input from the main algorithm (either Algorithm I or II), choose an arbitrary initial prior distribution α^1 on Θ^n and find the corresponding Bayes rule $\delta_{\alpha^1}^n$. Set $m = 0$.

2) Set $m = m + 1$. Determine if α^m is least favorable on Θ^n by checking to see if

$$\max_{\theta \in \Theta^n} R(\theta, \delta_{\alpha^m}^n) = r(\alpha^m, \delta_{\alpha^m}^n). \tag{2.2}$$

3) If (2.2) holds, then output $\alpha^m, \delta_{\alpha^m}^n$, and return to the main algorithm. Let $r^n = \alpha^m$.

4) Otherwise, construct a new prior distribution α^{m+1} as follows.

a) Find $\Theta^*, n, m = \arg[\max_{\theta \in \Theta^n} R(\theta, \delta_{\alpha^m}^n)]$, all maximizing points of $R(\theta, \delta_{\alpha^m}^n)$ over Θ^n.

b) Define a distribution $\beta^m = \{\beta_1^m, \cdots, \beta_{|\Theta^*, n, m|}^m\}$ on Θ^n by

$$\beta_j^m \equiv \begin{cases} \beta^m(\theta_j^n) = \dfrac{1}{|\Theta^{*,n,m}|}, & \text{for } \theta_j^n \in \Theta^{*,n,m} \\ 0, & \text{otherwise.} \end{cases}$$

c) Construct a family of distributions $\{\alpha^{m,\lambda}\}$ indexed by $\lambda \in (0,1]$, where

$$\alpha^{m,\lambda} = \lambda\beta^m + (1-\lambda)\alpha^m, \qquad \text{for } 0 < \lambda \leq 1.$$

d) Find the corresponding Bayes rule $\delta_{\alpha^{m,\lambda}}^n$ and the risk $r(\alpha^{m,\lambda}, \delta_{\alpha^{m,\lambda}}^n)$. Let

$$\Lambda^m = \arg\left[\max_{\lambda \in (0,1]} r(\alpha^{m,\lambda}, \delta_{\alpha^{m,\lambda}}^n)\right] \neq \phi.$$

e) Define $\alpha^{m+1} = \lambda^m\beta^m + (1-\lambda^m)\alpha^m$ where $\lambda^m \in \Lambda^m$. Go to step 2).

Note: For computational purposes, (2.2) in step 2) can be replaced by an error range

$$\max_{\theta \in \Theta^n} R(\theta, \delta_{\alpha^m}^n) - r(\alpha^m, \delta_{\alpha^m}^n) < \epsilon_0$$

for a prescribed tolerable error threshold ϵ_0. The convergence proof is in [6, corollary to theorem 6, p. 1650].

The following recursive algorithm is used based on a sequence of finite fixed size parameter subsets in the finite set Θ.

Algorithm I:

1) Initialization: Given an error threshold ϵ and a positive integer $J \geq N_0$, choose an arbitrary initial parameter set $\Theta^1 = \{\theta_1^1, \cdots, \theta_J^1\}$. Set $n = 0$.

2) Set $n = n + 1$. Apply the modified Nelson algorithm to Θ^n to find a least favorable distribution τ^n on Θ^n and the corresponding Bayes rule $\delta_{\tau^n}^n$.

3) Compute $\max_{\theta \in \Theta} R(\theta, \delta_{\tau^n}^n)$ and check the error

$$\epsilon_n = \max_{\theta \in \Theta} R(\theta, \delta_{\tau^n}^n) - r(\tau^n, \delta_{\tau^n}^n).$$

4) If $\epsilon_n < \epsilon$, then halt and output the τ^n, $\delta_{\tau^n}^n$ and $r(\tau^n, \delta_{\tau^n}^n)$. Otherwise, let $\{\theta_j^n\}$ be arranged in a nonincreasing order according to the probability distribution τ^n such that

$$\tau^n(\theta_i^n) > \tau^n(\theta_j^n) \text{ iff } i < j.$$

Let $\Theta^{*,n} = \arg[\max_{\theta \in \Theta} R(\theta, \delta_{\tau^n}^n)]$, all locally maximizing points in Θ with risks $R(\theta, \delta_{\tau^n}^n) \geq r(\tau^n, \delta_{\tau^n}^n)$, and define Θ^{n+1} as follows. If $|\Theta^{*,n}| = J$, then let $\Theta^{n+1} = \Theta^{*,n}$; otherwise, let

$$\begin{aligned} \theta_j^{n+1} &= \theta_j^n, \qquad 1 \leq j \leq J - |\Theta^{*,n}| \\ &= \theta_j^{*,n}, \qquad J - |\Theta^{*,n}| + 1 \leq j \leq J \end{aligned}$$

where $\theta_j^n \in \Theta^n$ and $\theta_j^{*,n} \in \Theta^{*,n}$, and then, let $\Theta^{n+1} = \{\theta_j^{n+1}\}$. Go to step 2).

The following two remarks are relevant here. First, it is worth noting that in step 4) $\Theta^{*,n}$ is chosen to be the set of all locally maximizing parameters $\theta^{*,n}$ in Θ whose risks $R(\theta^{*,n}, \delta_{\tau^n}^n) \geq r(\tau^n, \delta_{\tau^n}^n)$. The condition that $R(\theta^{*,n}, \delta_{\tau^n}^n) \geq r(\tau^n, \delta_{\tau^n}^n)$ will guarantee inequality (2.3) valid in the proof of Lemma 1. However, another possibility to make inequality (2.3) valid is to choose $\Theta^{*,n}$ to be the set of all possible globally maximizing parameters rather than locally maximizing parameters. Henceforth, we shall call the replacement procedure done by locally maximizing parameters the *local maxima replacement*, while the replacement done by globally maximizing parameters the *global maxima replacement*. The relative performance will be studied through numerical examples in Section V.

Second, with a slight modification the replacements made in step 4) according to probabilities on parameters can be also done by risks that are generated by parameters yield such that inequality (2.3) is still satisfied. A numerical result based on this modified algorithm applied to channel capacity problem can be found in [4].

B. Convergence Theorems for Algorithm I

Before stating the main theorem, we define a property that will be needed to prove the convergence of Algorithm I. Note that in step 4) of Algorithm I, the $(n+1)$st parameter set Θ^{n+1} is obtained by replacing those parameters in Θ^n with the lowest probabilities by the parameters in $\Theta^{*,n}$. It is not immediately obvious that the procedure will converge. For instance, if $|\Theta^{*,n}| = J$, then $\Theta^{n+1} = \Theta^{*,n}$ because all elements in Θ^n are replaced by all elements $\theta_j^{*,n}$ in $\Theta^{*,n}$. Thus $\Theta^n \cap \Theta^{n+1} = \phi$. This situation reveals a

lack of connection between Θ^{n+1} and Θ^n that could force the process to go back to the initial status. That is, without taking into account the information obtained previously we restart an initial parameter set Θ^{n+1} and repeat the whole procedure again. Accordingly, the following property is given for this purpose to bridge the gap between two consecutive iterations and conveys the least information of the present iteration to the next stage of iteration such that the iterative processes will eventually converge.

Definition (the Bayesian Transitivity Property): Given two statistical decision games (Θ^1, D, R) and (Θ^2, D, R), a decision rule δ is said to satisfy the (Θ^1, Θ^2) Bayesian transitivity property if the δ that is Bayes with respect to a prior on Θ^1 is also Bayes with respect to some prior on Θ^2.

Note: Θ^1 and Θ^2 can be defined on different spaces and need not have the same cardinalities.

The BTP can best be illustrated by the channel capacity problem considered in [8] or [10]. In these papers, for a given prior $\alpha = (\alpha_1, \cdots, \alpha_J)$ on the nth parameter set $\Theta^n = \{\theta_1^n, \cdots, \theta_J^n\}$, the nth Bayes rule on the output k is given by $\hat{q}_k^n = \sum_{j=1}^J \alpha_j p_k^{\theta_j^n}$ where $k \in$ the output space $= \{1, \cdots, L\}$ and $J \geq L$. If we solve the equation

$$\sum_{j=1}^{J} \alpha_j p_k^{\theta_j^n} = \sum_{j=1}^{J} \beta_j p_k^{\theta_j^{n+1}}$$

for a probability vector $\beta = (\beta_1, \cdots, \beta_J)$ on $\Theta^{n+1} = \{\theta_1^{n+1}, \cdots, \theta_J^{n+1}\}$, the resulting $\hat{q}_k^n$ is also a Bayes rule with respect to the prior β on Θ^{n+1}. Hence it satisfies the (Θ^n, Θ^{n+1}) BTP. More details will be given in Section IV and the Appendix. The main result of this paper is to provide implementable algorithms for the following minimax theorem and to show that the minimax risk computed by these algorithms is indeed convergent to the desired minimax risk.

Theorem 1: If there exists positive N_0 such that $R(\theta, \delta)$ has at most N_0 local maxima for every nonequalizer rule δ, then there exists a least favorable distribution on Θ and corresponding Bayes rule δ_τ such that

$$V = r(\tau, \delta_\tau) = \max_{\theta \in \Theta} R(\theta, \delta_\tau) = \min_{\delta \in D} \max_{\theta \in \Theta} R(\theta, \delta)$$

where V is the minimax risk.

In what follows, we will show that a sequence of priors $\{\tau^n\}$ (and the corresponding Bayes rule $\delta_{\tau^n}^n$) constructed in Algorithm I will converge to the desired τ (and to δ_τ a.s.), and thus $\lim_{n \to \infty} r(\tau^n, \delta_{\tau^n}^n) = V$.

Lemma 1: Suppose that for each n $\delta_{\tau^n}^n$ satisfies the (Θ^n, Θ^{n+1}) Bayesian transitivity property. Then $\{r(\tau^n, \delta_{\tau^n}^n)\}$ is a nondecreasing sequence in n, i.e.,

$$r(\tau^n, \delta_{\tau^n}^n) \leq r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1})$$

with equality iff $r(\tau^n, \delta_{\tau^n}^n) = \max_{\theta \in \Theta} R(\theta, \delta_{\tau^n}^n)$.

Proof: By the definition of $r(\tau^n, \delta_{\tau^n}^n)$,

$$r(\tau^n, \delta_{\tau^n}^n) = \sum_j \tau_j^n R(\theta_j^n, \delta_{\tau^n}^n) \leq \sum_j w_j R(\theta_j^{n+1}, \delta_{\tau^n}^n) \quad (2.3)$$

for any probability vector w on Θ^{n+1}.

By the assumption $\delta_{\tau^n}^n$ satisfies the Bayesian transitivity property. Thus there is a prior distribution β^{n+1} on Θ^{n+1} such that the corresponding Bayes rule $\delta_{\beta^{n+1}}^{n+1} = \delta_{\tau^n}^n$ and $\delta_{\beta^{n+1}}^{n+1}$ is also a rule for the game $(\Theta^{n+1}, \boldsymbol{D}, R)$. Since ω is arbitrary in (2.3), we can choose $w = \beta^{n+1}$, and thus we have

$$\begin{aligned} r(\tau^n, \delta_{\tau^n}^n) &= \sum_j \tau_j^n R(\theta_j^n, \delta_{\tau^n}^n) \leq \sum_j \beta_j^{n+1} R(\theta_j^{n+1}, \delta_{\beta^{n+1}}^{n+1}) \\ &\leq \sum_j \tau_j^{n+1} R(\theta_j^n, \delta_{\tau^{n+1}}^{n+1}) = r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1}). \end{aligned} \quad (2.4)$$

The last inequality holds because τ^{n+1} is least favorable on Θ^{n+1}. This shows that $r(\tau^n, \delta_{\tau^n}^n) \leq r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1})$ and that $r(\tau^n, \delta_{\tau^n}^n)$ is nondecreasing in n.

Proof of Equality:

1) Necessity: If $r(\tau^n, \delta_{\tau^n}^n) = r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1})$, then from (2.4) we have

$$\begin{aligned} r(\tau^n, \delta_{\tau^n}^n) &= \sum_j \tau_j^n R(\theta_j^n, \delta_{\tau^n}^n) = \sum_j \beta_j^{n+1} R(\theta_j^{n+1}, \delta_{\beta^{n+1}}^{n+1}) \\ &= \sum_j \tau_j^{n+1} R(\theta_j^{n+1}, \delta_{\tau^{n+1}}^{n+1}) = r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1}). \end{aligned}$$

This implies that β^{n+1} is also least favorable on Θ^{n+1}, so

$$\begin{aligned} r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1}) &= \max_{\theta \in \Theta^{n+1}} R(\theta, \delta_{\tau^{n+1}}^{n+1}) = \max_{\theta \in \Theta^{n+1}} R(\theta, \delta_{\beta^{n+1}}^{n+1}) \\ &\geq \max_{\theta \in \Theta} R(\theta, \delta_{\tau^n}^n) \geq r(\tau^n, \delta_{\tau^n}^n) \\ &= r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1}). \end{aligned}$$

The second equality holds because $\delta_{\beta^{n+1}}^{n+1}$ is also a minimax rule for the game $(\Theta^{n+1}, \boldsymbol{D}, R)$, and the first inequality is true because of the choice of θ_j^{n+1} and $\delta_{\beta^{n+1}}^{n+1} = \delta_{\tau^n}^n$. Consequently, $r(\tau^n, \delta_{\tau^n}^n) = \max_{\theta \in \Theta} R(\theta, \delta_{\tau^n}^n)$.

2) Sufficiency: If $r(\tau^n, \delta_{\tau^n}^n) = \max_{\theta \in \Theta} R(\theta, \delta_{\tau^n}^n)$, we want to prove that $r(\tau^n, \delta_{\tau^n}^n) = r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1})$. Obviously, $\delta_{\tau^n}^n$ is also a decision rule for the game (Θ^{n+1}, D, R), and $\delta_{\tau^{n+1}}^{n+1}$ is a minimax rule for $(\Theta^{n+1}, \boldsymbol{D}, R)$. This implies that

$$\begin{aligned} r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1}) &= \max_{\theta \in \Theta^{n+1}} R(\theta, \delta_{\tau^{n+1}}^{n+1}) \\ &\leq \max_{\theta \in \Theta^{n+1}} R(\theta, \delta_{\tau^n}^n) \\ &\leq \max_{\theta \in \Theta} R(\theta, \delta_{\tau^n}^n) \\ &= r(\tau^n, \delta_{\tau^n}^n). \end{aligned}$$

The first equality follows because $\delta_{\tau^{n+1}}^{n+1}$ is a minimax rule, and the first inequality holds because $\delta_{\tau^n}^n$ is a decision rule in $\boldsymbol{D}$. The second inequality is true since Θ^{n+1} is a subset of Θ, and the last equality holds by the assumption of sufficiency. However, from the first part of the theorem $\{r(\tau^n, \delta_{\tau^n}^n)\}$ is nondecreasing, and so we have $r(\tau^n, \delta_{\tau^n}^n) = r(\tau^{n+1}, \delta_{\tau^{n+1}}^{n+1})$. □

An immediate consequence of Lemma 1 is the following corollary, which can be proven by an appropriate modification of the proof of Lemma 1.

Corollary 1: Under the assumption of Lemma 1 that $\delta_{\tau^n}^n$ satisfies the (Θ^n, Θ^{n+1}) BTP for each n, Algorithm I

converges and terminates at step N if and only if $\epsilon_N = \max_{\theta\in\Theta} R(\theta, \delta_{\tau^N}^N) - r(\tau^N, \delta_{\tau^N}^N) < \epsilon$ for a given error threshold ϵ.

Lemma 2: Let $\{\delta^n\}$ be a sequence of decision rules in D^* and $\delta^* \in D^*$. If $\lim_{n\to\infty}\delta^n = \delta^*$ a.e. μ on X. Then

$$\lim_{n\to\infty}\max_{\theta\in\Theta} R(\theta, \delta^n) = \max_{\theta\in\Theta} R(\theta, \delta^*).$$

Proof: The theoretical foundation of this lemma was done in [6]. We refer all straightforward technical details to [8].

It has been shown in Lemma 1 that the sequence $\{r(\tau^n, \delta_{\tau^n}^n)\}$ is monotonically decreasing and thus converges to some number $\hat{r}$. In the following lemma we prove that the limit of $\max_{\theta\in\Theta} R(\theta, \delta_{\tau^n}^n)$ also converges to $\hat{r}$. As a result, we can further conclude by a theorem (Theorem 2 stated as follows) that the number $\hat{r}$ is actually equal to the minimax V in Theorem 1 whereby Theorem 1 is justified as well. □

Lemma 3:

$$\lim_{n\to\infty}\max_{\theta\in\Theta} R(\theta, \delta_{\tau^n}^n) = \hat{r}.$$

Proof: Based on the work in [6], it can be easily shown that τ^n converges weakly to τ and thus $\delta_{\tau^n}^n \to \delta_\tau$ a.e. Furthermore, by Lemma 2 we can derive

$$\lim_{n\to\infty}\max_{\theta\in\Theta} R(\theta, \delta_{\tau^n}^n) = \max_{\theta\in\Theta} R(\theta, \delta_\tau).$$

However, by straightforward justification τ is also a least favorable distribution on Θ if we define $\tau(\theta) = 0$ for all $\theta \in \Theta - \hat{\Theta}$ where $\hat{\Theta}$ is defined to be the limit of nth parameter set Θ^n on which Algorithm I iterates at nth iteration. This results in

$$\max_{\theta\in\Theta} R(\theta, \delta_\tau) = \max_{\theta\in\Theta} R(\theta, \delta_\tau) = r(\tau, \delta_\tau).$$

Moreover, as shown in Lemma 1, the sequence $r(\tau^n, \delta_{\tau^n}^n) \to r(\tau, \delta_\tau) = \hat{r}$. This shows that $\lim_{n\to\infty}\max_{\theta\in\Theta} R(\theta, \delta_{\tau^n}^n) = \hat{r}$. □

Finally, we prove by the following theorem that the limit value $\hat{r}$ to which the sequence $\{r(\tau^n, \delta_{\tau^n}^n)\}$ converges is indeed the desired minimax value

$$V = \min_{\delta\in D}\max_{\theta\in\Theta} R(\theta, \delta)$$

in Theorem 1.

Theorem 2: The sequence $\{r(\tau^n, \delta_{\tau^n}^n)\}$ converges to the limit $V = \min_{\delta\in D^*}\max_{\theta\in\Theta} R(\theta, \delta)$.

Proof: Since V is the minimax value,

$$V \le \max_{\theta\in\Theta} R(\theta, \delta_\tau). \tag{2.5}$$

On the other hand, by Lemma 1 we have

$$\lim_{n\to\infty}\max_{\theta\in\Theta} R(\theta, \delta_{\tau^n}^n) = \max_{\theta\in\Theta} R(\theta, \delta_\tau) = r(\tau, \delta_\tau) = \lim_{n\to\infty} r(\tau^n, \delta_{\tau^n}^n) = \hat{r} \le V. \tag{2.6}$$

The last inequality holds because for each n, $r(\tau^n, \delta_{\tau^n}^n)$ is bounded by V from Lemma 1. Combining (2.5) and (2.6) yields that $V = \hat{r}$. So, the sequence $\{r(\tau^n, \delta_{\tau^n}^n)\}$ converges to the desired minimax risk. □

III. A Second Algorithm for Finding Minimax Solutions Based on a Sequence of Finite Varying Parameter Sets

A. Description of Algorithm II

In the first algorithm the number of parameters is fixed at $J < \infty$. We now present a second algorithm that allows the number of parameters to vary.

Algorithm II:

1) Initialization: Given an error threshold ϵ and an arbitrary positive integer J^1, choose an arbitrary initial parameter set $\Theta^1 = \{\theta_1^1, \cdots, \theta_{J^1}^1\}$. Set $n = 0$.

2) Set $n = n + 1$. Apply the modified Nelson algorithm to Θ^n to find a least favorable distribution σ^n on Θ^n and the corresponding Bayes rule $\delta_{\sigma^n}^n$.

3) Compute $\max_{\theta\in\Theta} R(\theta, \delta_{\sigma^n}^n)$ and check the error

$$\epsilon_n = \max_{\theta\in\Theta} R(\theta, \delta_{\sigma^n}^n) - r(\sigma^n, \delta_{\sigma^n}^n).$$

4) If $\epsilon_n < \epsilon$, then halt and output the least favorable distribution σ^n and the corresponding Bayes rule $\delta_{\sigma^n}^n$. Otherwise, delete the parameters in Θ^n with zero probability, relabel the remaining parameters and denote the resulting parameter set by $\tilde{\Theta}^n$. Let $\Theta^{*,n} = \arg[\max_{\theta\in\Theta} R(\theta, \delta_{\sigma^n}^n)]$, all locally maximizing points of $R(\theta, \delta_{\sigma^n}^n)$ over Θ and define

$$\theta_j^{n+1} = \theta_j^n;\ 1 \le j \le |\tilde{\Theta}^n|, \quad \text{for } \theta_j^n \in \tilde{\Theta}^n,$$
$$= \theta^*{}_j^n;\ 1 + |\tilde{\Theta}^n| \le j \le |\tilde{\Theta}^n| + |\Theta^{*,n}|$$

where

$$\theta^*{}_j^n \in \Theta^{*,n}.$$

Let $\Theta^{n+1} = \{\theta_j^{n+1}\}$. Go to step 2).

In analogy with Algorithm I, two similar comments can be made. In particular, policy analogous to step 4) can be adopted by adding the set of all globally maximizing parameters instead of locally maximizing parameters. We will call the *local* and *global maxima additions* respectively afterwards.

Note: Under the assumption $N_0 < \infty$, the nth parameter set Θ^n always has its size $\le 3N_0$. This fact will be verified in the next section (Corollary 3).

B. Convergence Theorem for Algorithm II

The convergence theorems for Algorithm II are nearly the same except Lemma 1. In the following we will prove a lemma analogous to Lemma 1 for Algorithm II. However, unlike Lemma 1, the assumption of the BTP is not necessary for this lemma. It will be seen in the proof of the following lemma that the BTP is automatically satisfied.

Lemma 1A (Convergence Theorem for Algorithm II): $\{r(\sigma^n, \delta_{\sigma^n}^n)\}$ is a nondecreasing sequence in n, i.e., $r(\sigma^n, \delta_{\sigma^n}^n) \le r(\sigma^{n+1}, \delta_{\sigma^{n+1}}^{n+1})$ with equality iff $r(\sigma^n, \delta_{\sigma^n}^n) = \max_{\theta\in\Theta} R(\theta, \delta_{\sigma^n}^n)$.

Proof: Since σ^{n+1} is a least favorable distribution on Θ^{n+1},

$$r\left(\sigma^{n+1}, \delta_{\sigma^{n+1}}^{n+1}\right) \geq r\left(\alpha^{n+1}, \delta_{\alpha^{n+1}}^{n+1}\right)$$

for any probability vector α^{n+1} on Θ^{n+1}. In particular, we can choose α^{n+1} as follows:

$$\alpha_j^{n+1} = \begin{cases} \sigma_j^n, & \theta_j^{n+1} \in \Theta^n \\ 0, & \theta_j^{n+1} \in \Theta^{*,n}. \end{cases}$$

Thus we obtain

$$\begin{aligned} r\left(\sigma^{n+1}, \delta_{\sigma^{n+1}}^{n+1}\right) &\geq \sum_{\theta_j^{n+1} \in \tilde{\Theta}^n} \sigma_j^n R\left(\theta_j^n, \delta_{\alpha^{n+1}}^{n+1}\right) \\ &\quad + \sum_{\theta_j^{n+1} \in \Theta^{*,n}} 0 \times R\left(\theta_j^{n+1}, \delta_{\alpha^{n+1}}^{n+1}\right) \\ &= \sum_{\theta_j^n \in \Theta^n} \sigma_j^n R\left(\theta_j^n, \delta_{\sigma^n}^n\right) \\ &= \sum_{\theta_j^n \in \Theta^n} \sigma_j^n R\left(\theta_j^n, \delta_{\sigma^n}^n\right) \\ &= r\left(\sigma^n, \delta_{\sigma^n}^n\right), \end{aligned} \tag{3.1}$$

so $\{r(\sigma^n, \delta_{\sigma^n}^n)\}$ is nondecreasing in n.

Proof of Equality:

1) Necessity: If $r(\sigma^n, \delta_{\sigma^n}^n) = r(\sigma^{n+1}, \delta_{\sigma^{n+1}}^{n+1})$, then from (3.1), we have

$$r\left(\sigma^{n+1}, \delta_{\sigma^{n+1}}^{n+1}\right) = r\left(\alpha^{n+1}, \delta_{\alpha^{n+1}}^{n+1}\right) = r\left(\sigma^n, \delta_{\sigma^n}^n\right)$$

where

$$\alpha_j^{n+1} = \begin{cases} \sigma_j^n, & \theta_j^{n+1} \in \Theta^n \\ 0, & \theta_j^{n+1} \in \Theta^{*,n}. \end{cases}$$

This implies that α^{n+1} is not only least favorable on Θ^{n+1} but also least favorable on Θ^n. Hence the corresponding Bayes rule $\delta_{\alpha^{n+1}}^{n+1}$ is a minimax rule for the game $(\Theta^{n+1}, \boldsymbol{D}, R)$ and is also a mininax rule for the game $(\Theta^n, \boldsymbol{D}, R)$. However, from [6], the minimax rule is essentially unique, and thus the rule $\delta_{\alpha^{n+1}}^{n+1}$ is equivalent to the rule $\delta_{\sigma^n}^n$. Thus

$$\begin{aligned} \max_{\theta \in \Theta} R\left(\theta, \delta_{\sigma^n}^n\right) &\leq \max_{\theta \in \Theta^{n+1}} R\left(\theta, \delta_{\sigma^n}^n\right) \\ &= \max_{\theta \in \Theta^{n+1}} R\left(\theta, \delta_{\alpha^{n+1}}^{n+1}\right) \\ &= r\left(\alpha^{n+1}, \delta_{\alpha^{n+1}}^{n+1}\right) \\ &= r\left(\sigma^n, \delta_{\sigma^n}^n\right). \end{aligned}$$

The first inequality follows from the choice of θ^{n+1}, and the first equality holds because $\delta_{\alpha^{n+1}}^{n+1}$ is equivalent to $\delta_{\sigma^n}^n$. Clearly, however, $\max_{\theta \in \Theta} R(\theta, \delta_{\sigma^n}^n) \geq r(\sigma^n, \delta_{\sigma^n}^n)$. Therefore, $r(\sigma^n, \delta_{\sigma^n}^n) = \max_{\theta \in \Theta} R(\theta, \delta_{\sigma^n}^n)$.

2) Sufficiency: If $r(\sigma^n, \delta_{\sigma^n}^n) = \max_{\theta \in \Theta} R(\theta, \delta_{\sigma^n}^n)$, we want to show that

$$r\left(\sigma^n, \delta_{\sigma^n}^n\right) = r\left(\sigma^{n+1}, \delta_{\sigma^{n+1}}^{n+1}\right).$$

Obviously, $\delta_{\sigma^n}^n$ is also a decision rule for the game $(\Theta^{n+1}, \boldsymbol{D}, R)$, since for each $\theta_j^{n+1} \in \Theta^{n+1}$,

$$\begin{aligned} R\left(\theta_j^{n+1}, \delta_{\sigma^n}^n\right) &\leq \max_{\theta \in \Theta} R\left(\theta, \delta_{\sigma^n}^n\right) \\ &= r\left(\sigma^n, \delta_{\sigma^n}^n\right) < +\infty, \\ r\left(\sigma^{n+1}, \delta_{\sigma^{n+1}}^{n+1}\right) &= \max_{\theta \in \Theta^{n+1}} R\left(\theta, \delta_{\sigma^{n+1}}^{n+1}\right) \\ &\leq \max_{\theta \in \Theta^{n+1}} R\left(\theta, \delta_{\sigma^n}^n\right) \\ &\leq \max_{\theta \in \Theta} R\left(\theta, \delta_{\sigma^n}^n\right) \\ &= r\left(\sigma^n, \delta_{\sigma^n}^n\right). \end{aligned}$$

The first equality holds because $\delta_{\sigma^{n+1}}^{n+1}$ is a minimax rule. The first inequality follows from the fact that $\delta_{\sigma^{n+1}}^{n+1}$ is minimax and the rule $\delta_{\sigma^n}^n$ is in $\boldsymbol{D}$. The second inequality holds since Θ^{n+1} is a subset of Θ. The last equality is true because of the assumption. However, it has been shown that $\{r(\sigma^n, \delta_{\sigma^n}^n)\}$ is nondecreasing. This implies that

$$r\left(\sigma^{n+1}, \delta_{\sigma^{n+1}}^{n+1}\right) = r\left(\sigma^n, \delta_{\sigma^n}^n\right). \qquad \square$$

An immediate consequence of Lemma 1A is the following corollary.

Corollary 2: Algorithm II converges and terminates at step N if and only if

$$\epsilon_N = \max_{\theta \in \Theta} R\left(\theta, \delta_{\sigma^N}^N\right) - r\left(\sigma^N, \delta_{\sigma^N}^N\right) < \epsilon$$

for a given error threshold ϵ.

Remark: A significant implication of Algorithms I and II is that these algorithms present a general approach to solving minimax solutions using the modified Nelson algorithm. However, in some specific problems the modified Nelson algorithm can be replaced by more efficient algorithms, e.g., in the source matching problems considered in [7] or in the channel capacity problems in [9], [10], the modified Nelson algorithm is replaced by the Arimoto-Blahut algorithm.

Thus far we have described two algorithms for finding minimax rules. As we have seen, Algorithm I iterates on fixed size parameter sets; whereas Algorithm II iterates with varying sizes. Seemingly, they look like different schemes, but they have common characteristics. Recall that N_0 is defined to be the maximum number of locally maximizing points of the risk function on Θ over all nonequalizer rules. If we start with an initial parameter set of size $3N_0$, then Algorithms I and II essentially perform the same iterative processes; this will be shown in the following corollary. Of course, if Algorithm II begins with an initial parameter set with an arbitrary size less than $3N_0$, in general, both algorithms will not produce the same iteration at each step.

Corollary 3: Under the same assumptions made in Section II, if Algorithms I and II are initiated by any arbitrary parameter set with a size of $3N_0$, algorithms I and II are identical in the sense that, for every n at the nth iteration, Algorithm II arranges the parameters in Θ^n as in Algorithm I, deletes the last $|\Theta^{*,n}|$ parameters, and then adds parameters in $\Theta^{*,n}$ to the resulting parameter set.

Proof: The minimax rule $\delta_{\tau^n}^n$ places nonzero probability on those values in Θ^n for which $R(\theta, \delta_{\tau^n}^n)$ equals the minimax risk, because, for each nonequalizer rule δ the risk function $R(\theta, \delta)$ has at most N_0 locally maximizing points over all θ, there can be at most $2N_0$ of these points in Θ^n. Since there are at least N_0 parameters with zero probabilities, i.e., $\delta_{\tau^n}^n(\theta_j^n) = 0$ for $2N_0 + 1 \le j \le 3N_0$. However, to find Θ^{n+1} Algorithm I replaces by $\{\theta_j^{*,n}\}$ at most N_0 parameters $\{\theta_j^n\}_{j=3N_0-|\theta^{*,n}|+1}^{3N_0}$ (which in fact have zero probabilities). Thus this process results in the same parameter set Θ^{n+1} produced by Algorithm II. □

Some remarks on Corollary 3 are given as follows.

1) In Corollary 3, it has been shown that under a mild condition Algorithms I and II could be regarded as the same algorithm as long as the size of the initial parameter set was chosen to be $3N_0$. Therefore, at any time the iterated parameter sets in both algorithms are always the same and have a constant size $3N_0$.

2) Step 4) (i.e., the last step) of Algorithm II requires deleting all zero probability parameters to prevent the size of the parameter sets from growing; the algorithm does not have to arrange parameters as does Algorithm I. So, under the assumption $N_0 < +\infty$, all parameter sets are always $\le 3N_0$. On the other hand, in Corollary 3 the elements in Θ^n of Algorithm II are well-arranged and only the last $|\Theta^{*,n}|$ parameters are deleted to keep the parameter sets with a constant size $3N_0$ all the time. The reason that we adopted this technique is to ensure that Algorithm II deletes the same parameters in Θ^n which are being replaced by Algorithm I.

3) As indicated in the beginning, if Algorithm II starts with an initial parameter set of size $\le 3N_0$, then the iterative processes may be terminated before the size of the parameter set in the last iteration reaches $3N_0$, but it may require more iterations than when the algorithm starts with an initial parameter set of size $3N_0$. So, the choices depend on the trade-off between costs.

4) Note that we have referred only to nonequalizer rules. If at any step an equalizer rule, $\delta_{\tau^n}^n$, is found, the algorithm automatically terminates.

IV. A Further Study of the Bayesian Transitivity Property

In Section II-A we defined the BTP, a property required for Algorithm I. In what follows we study this property further by looking at two commonly used examples; we will see later that the buffer size J needed for Algorithm I is determined by the BTP, not by N_0. A more detailed study on the Bayesian transitivity property can be found in [8].

Example 1 (Estimation Problems with Relative Entropy Loss)

Most problems of this kind arise in communication theory and have already been investigated extensively. In this example we consider the probability mass function of a random variable X defined on the samples space X which is specified by a binomial distribution with $N+1$ observations. More precisely, we let

Θ $[0,1]$, parameter space,
A $[0,1]$, action space,
X $\{0,1,\cdots,N\}$, sample space,
$p(x|\theta)$ $= C_{N,x}\,\theta^x(1-\theta)^{N-x}$ where $C_{N,x} = N!/x!(N-x)!$, conditional probability defined on X given $\theta \in \Theta$,
$\boldsymbol{D}$ $= \{d = (d(0), d(1),\cdots, d(N)) | d(x) \in A$ and $\sum_{x=0}^N d(x) = 1\}$, decision space (i.e., any nonrandomized decision rule d can be specified by an $N+1$-dimensional probability vector $(d(0),\cdots, d(N)) \in \boldsymbol{D}$.)
$L(\theta, d)$ $= \log[p(x|\theta)/d(x)]$, loss function,
$R(\theta, d)$ $\sum_{x=0}^N p(x|\theta)[p(x|\theta)/d(x)]$ = relative entropy, risk function.

For any subset Θ^1 contained in Θ and any given prior α on Θ^1, a Bayes rule is given by

$$d_\alpha(x) = \int_{\Theta^1} p(x|\theta)\alpha(\theta)\,d\theta = \int_{\Theta^1} C_{N,x}\theta^x(1-\theta)^{N-x}\alpha(\theta)\,d\theta, \quad \text{for } x = 0,\cdots,N \tag{4.1}$$

and so,

$$d_\alpha(x) = C_{N,x}\int_{\Theta_1} \theta^x(1-\theta)^{N-x}\alpha(\theta)\,d\theta, \quad x = 0,1,\cdots,N \tag{4.2}$$

Let Θ^2 be another subset in Θ with the same size as Θ^1. For the decision rule $d_\alpha = (d(0), d(1),\cdots, d(N))$ to satisfy the (Θ^1, Θ^2) BTP, we must be able to find a prior β on Θ^2 such that for each $x = 0,1,\cdots,N$,

$$d_\alpha(x) = d_\beta(x) = C_{N,x}\int_{\Theta^2} \theta^x(1-\theta)^{N-x}\beta(\theta)\,d\theta. \tag{4.3}$$

In the following we propose two approaches to proving the (Θ^1, Θ^2) BTP. Since we are only interested in finite parameter sets, the sets Θ^1, Θ^2 are assumed to be finite and of the same size.

1) Algebraic Approach: Let $\Theta^1 = \Theta^n = \{\theta_1^n,\cdots,\theta_J^n\}$ and $\Theta^2 = \Theta^{n+1} = \{\theta_1^{n+1},\cdots,\theta_J^{n+1}\}$, then (4.1) becomes

$$d_\alpha(x) = \sum_{j=1}^J \alpha_j p(x|\theta_j^n)$$

where

$$\alpha_j = \alpha(\theta_j^n), \qquad p(x|\theta_j^n) = C_{N,x}(\theta_j^n)^x(1-\theta_j^n)^{N-x}.$$

To find a probability vector β defined on Θ^{n+1} such that d_α satisfies (Θ^n, Θ^{n+1}) Bayesian transitivity property, we only have to solve for a β the following set of $N+1$ linear algebraic equations that is equivalent to (4.3):

$$d_\alpha(x) = \sum_{j=1}^J \alpha_j p(x|\theta_j^n) = \sum_{j=1}^J \beta_j p(x|\theta_j^{n+1}) = d_\beta(x), \quad \text{for all } x = 0,1,\cdots,N. \tag{4.4}$$

Since α, β, $p(x|\theta_j^n)$ and $p(x|\theta_j^{n+1})$ are all probability vectors, (4.4) can be solved if $J \geq N+1$. This shows that the BTP is valid whenever $p(x|\theta)$ is a probability mass function of a discrete random variable $\boldsymbol{X}$, in particular this is true for binomial distributions.

2) Moment Approach: From (4.2) we derived that d_α is uniquely determined by the first N moments of α, and so is d_β from (4.3). This observation reveals an important fact —that the Bayesian transitivity property essentially hinges on the moments induced by the prior α.

If we are given a probability mass function of a polynomial form in θ, and if a Bayes rule d_α defined on X with respect to a prior α on a parameter subset Θ^1 is uniquely determined by the first N moments of the prior α, then for any other parameter subset Θ^2 with the same size as Θ^1, we are able to find a prior β on Θ^2 such that, for each x in X, $d_\alpha(x) = d_\beta(x)$ where d_β is a Bayes rule defined on X with respect to β. According to (4.2), for a binomial distribution $p(x|\theta)$, d_β is uniquely determined by the first N moments of the prior α. Therefore, the trick is that instead of directly solving for a β on Θ^2 for the given probability mass function, we use a binomial distribution as a base to consider a binomial distribution with N observations with respect to the same prior α so that based on (4.4), a β can be found. Since the binomial distribution with $N+1$ observations is a polynomial in θ with the degree N, (4.2) also uniquely determines the first N moments of the α and so does the β from (4.3). It is important that each x in X determines a moment of the prior α and vice versa, e.g., if $x = k$, then it determines the kth moment of α. As a result, this β is exactly what we need. However, note that, in this case, β need not be unique, because any prior $\hat{\beta}$ defined on Θ^2 satisfying the first N moments determined by (4.2) is also a candidate for $d_\alpha(x)$ satisfying (Θ^1, Θ^2)-BTP.

Example 2 (Estimation Problems with Squared Error Loss)

In this example we consider estimation problems with squared error loss where all assumptions made in Example 1 are the same except that the loss function is chosen to be squared error. It is well-known that a Bayes rule for an estimation problem with respect to square error loss is obtained by calculating a posterior conditional mean. More precisely, for any prior α on a parameter space Θ, a Bayes rule d_α with respect to α is given by $E_\alpha[\Theta|\boldsymbol{X} = x]$ for every x in the sample space X, that is

$$d_\alpha(x) = \frac{\int_{\Theta^1} \theta^{x+1}(1-\theta)^{N-x}\alpha(\theta)\,d\theta}{\int_{\Theta^1} \theta^{x}(1-\theta)^{N-x}\alpha(\theta)\,d\theta}, \quad \text{for } x = 0,1,2,\cdots,N \quad (4.5)$$

where the term $C_{N,x}$ appears in both numerator and denominator and has been canceled out.

Obviously, the algebraic method suggested in Example 1 cannot be directly applied to proving the BTP. However, if we compare (4.5) with (4.1) we will find that for each x the Bayes rule $d_\alpha(x)$ in (4.1), is exactly the denominator of the $d_\alpha(x)$ in (4.5) scaled by the constant $C_{N,x}$. On the other hand, the numerator in (4.5) is simply specified by the moments of the prior α on Θ^1. Therefore $d_\alpha(x)$ in (4.5) is uniquely determined by $N+1$ moments of α on Θ^1. As shown in the expression of (4.5), one more moment is required than that in (4.1) (i.e., $N+1$ st moment). This extra term is due to the numerator in (4.5) when $x = N$. Consequently, the moment approach is readily applied here. A simple example to illustrate how the moment method is applied to finding the desired β which will yield $d_\beta(x) = d_\alpha(x)$ for $x = 0,1,\cdots,N$ is given in [8].

Although we only considered binomial distributions for $p(x|\theta)$, the argument can be carried out to deal with Poisson and Gaussian distributions by using the Stone–Weierstrass approximation theorem. The details can be also found in [8].

We close this section with some comments on the relationship between BTP and N_0. Recall that, in the previous examples, the sample distributions with which we dealt were polynomials in θ. It follows that N_0 is finite and the BTP is valid by applying either the algebraic approach or the moment approach. As a result, Algorithms I and II are applicable. This idea reduces the problems of proving $N_0 < +\infty$ and BTP to that whether or not we can approximate a sample distribution by a polynomial uniformly on Θ. Fortunately, under some regularities (e.g., Θ is compact and the risk function is continuous on Θ) this can be done within any assigned degree of accuracy by a well-known theorem, the Stone-Weierstrass approximation theorem (see the Appendix). The significant implication of this theorem offers a connection that, in the some sense, requiring $N_0 < +\infty$ is equivalent to justifying the BTP, and thus Algorithms I and II have the same extent in applications to which we have freedom to choose either one for implementations. However, note that before applying Algorithm I we ought to find J which is equivalent to proving the Bayesian transitivity property. Hence, whenever there is a difficulty with determining J, Algorithm II is always preferred. Nevertheless Algorithm I has an advantage that it only needs J buffers, and thus it is more efficient than Algorithm II when J or N_0 is large.

V. Numerical Results

In the last section we studied the theoretical bases for Algorithms I and II on estimation problems with relative entropy loss and squared error loss. Now we study two numerical examples corresponding to Example 1 and 2, respectively, and analyze the relative performance of Algorithms I and II.

Example 3 (Channel Capacity Problems)

Basically, this example was studied in [5], and numerical results obtained based on Algorithms I and II were also given there. However, to see how Algorithms I and II apply to channel capacity problems, we briefly discuss this application and include some numerical results regarding

the global maxima replacement for Algorithm I and the global maxima addition for Algorithm II not available in [10] but in [9] and also refer all the details to [9], [10].

Let us consider a generalized binarylike memoryless channel that is specified by the input space $X=[0,1]$, the output space $Y_{L+1}=\{0,1,2,\cdots,L\}$, and the channel transition probabilities $\{P(k|x)\}_{x\in,k\in Y_{L+1}}$ given by

$$P(k|x)=C_{L,k}x^k(1-x)^{L-k}$$

where $C_{L,k}=L!/k!(L-k)!$. Then the channel capacity is defined by

$$C_{L+1}=\max_{p(x)}\left\{\sum_{k=0}^{L}\int_X p(x)P(k|x)\log\frac{P(k|x)}{q_p^*(k)}\,dx\right\}.$$

where $q_p^*(k)=\int_X p(x)P(k|x)\,dx$.

It is easy to see that the parameter space is characterized by the input space, the sample space by the output space, the action space by [0,1], a decision rule by an $L+1$-dimensional probability vector and the risk function is

$$\sum_{k=0}^{L}P(k|x)\log\frac{P(k|x)}{q(k)}.$$

For any input probability vector p the optimum output probability vector q_p^* is the Bayes rule with respect to p. Furthermore, by the algebraic method the BTP is satisfied by choosing $J=L+1$, the number of channel outputs.

The numerical results are obtained based on two different policies for forming new parameter sets after completing an iteration cycle, i.e., the global maxima replacement (addition) and the local maxima replacement (addition) for Algorithm I (II). $L+1$ ranges from 2 to 30 and the error threshold is set to be 6×10^{-4}. Fig. 1 shows that for each L Algorithms I and II converge to nearly the same value for both cases (global maxima and local maxima). (Note that in Figs. 1, 2, and 3, G and L are abbreviations of global maxima and local maxima.) Table I also shows that the global maxima replacement (addition) generally requires more iterations than does the local maxima replacement (addition). Moreover, from Table I we also learn that both algorithms are indeed very efficient. In most cases only two or three iterations are needed to terminate execution (no more than 6 iterations overall). A surprising observation from Table I shows that the global maxima addition for Algorithm II outperforms the local maxima addition for Algorithm II and even Algorithm I. This is because Algorithm II deletes all zero probability parameters (in this example, we delete all parameters with probabilities less than 10^{-8}) before adding new parameters. For instance, when $L+1\geq 20$ the global maxima addition for Algorithm II requires less parameters than $L+1$ which is required for Algorithm I and also less than does the local maxima addition for Algorithm II. On the other hand, for $L+1=14$, the local maxima addition for Algorithm I performs better than both Algorithm I and the global maxima addition for Algorithm II. By and large, this example shows that the global maxima addition for Algorithm II has better performance than Algorithm I and the local maxima addition for Algorithm II at the expense of requiring more iterations. However, as long as the size for the initial parameter set can be preset by $L+1$ in advance without considering buffer problems through the entire execution, Algorithm I is generally preferred to Algorithm II.

Example 4 (Estimation Problems with Squared Error Loss)

In this example we continue to investigate Example 2 by implementing Algorithms I and II on computers for a general binomial distribution where the sample space consisting of observations $N+1$ ranging from 2 to 30 and the loss function is $L(\theta,a)=(\theta-a)^2$ where θ and $a\in[0,1]$.

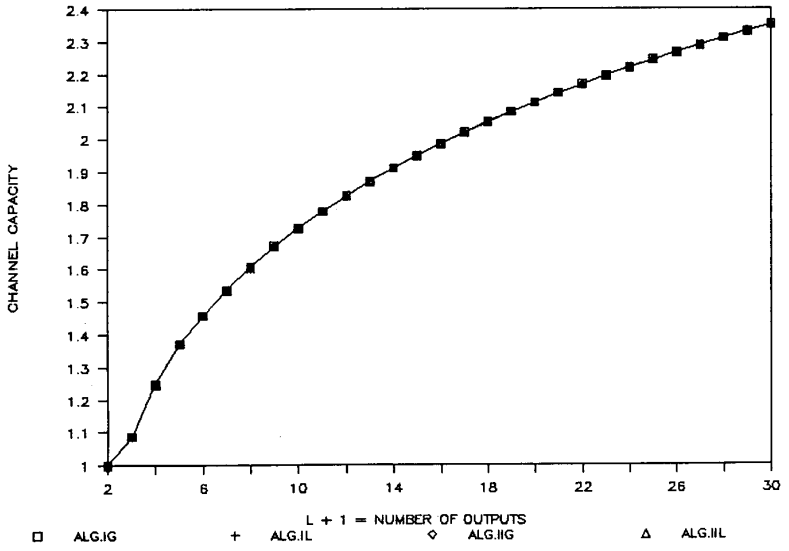

Fig. 1. Channel capacity versus $L+1$ = number of outputs.

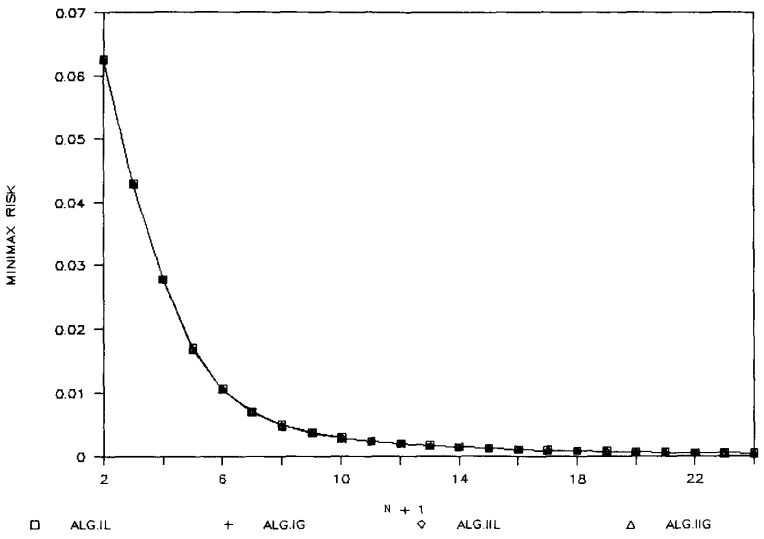

Fig. 2. Minimax risk versus $N+1$ = number of observations.

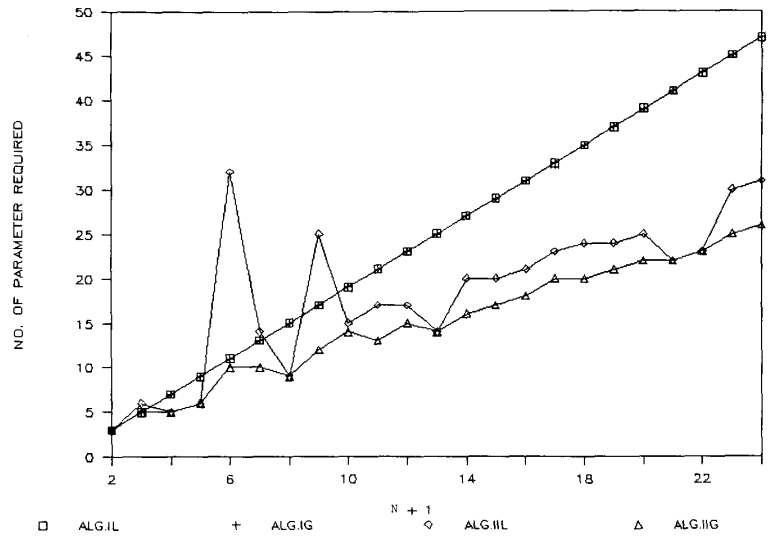

Fig. 3. Comparison of number of parameters required for Algorithms IG, IL, IIG and IIL.

The error threshold is set to be 10^{-5}. Since the conditional binomial distribution is symmetric with respect to $1/2$, we can confine ourselves to the range $[0,1/2)$. Therefore, whenever a $\theta \in [0,1/2)$ is selected, its symmetric point $1-\theta \in (1/2,1]$ is also chosen. On the other hand, according to the moment approach, the size of iterating parameter sets, J for Algorithm I is $N+2$, for we need determine $N+1$ moments and plus one extra $N+2$nd moment resulted from the numerator in (4.5). Because $1/2$ is the midpoint of $[0,1]$, $1/2$ is always included in the initial parameter set Θ^1. So, we set $J=2N+1$ by incorporating the symmetric points with respect to those parameters chosen from $[0,1/2)$. For instance, for $N=1$, we have two observations, and thus we need three parameters (i.e., $0,1/2,1$). This fact has been seen in the third comment following the assumptions made in the very beginning of Section II. At this moment, we would like to point out that in this example we chose $2N+1$ for J, but it does not mean that J must be at least $2N+1$. This can be seen from Table II where, in general, the number of parameters for Algorithm II is actually less than the J chosen for Algorithm I for most cases, particularly, when N is increasing. The reason for choosing $2N+1$ for Algorithm I is to ensure that there are always enough parameters. The numerical results show that, whatever parameters we start with for Algorithms I and II, the parameters $0,1/2,1$ are always chosen and the results in Fig. 2 yielded by these two algorithms are very close. Fig. 3 also shows that, although Algorithm II iterates finite parameter sets with varying sizes, it generally requires much fewer parameters

TABLE I
A COMPARISON OF PERFORMANCE BETWEEN ALGORITHM I AND ALGORITHM II IN EXAMPLE 3 (TAKEN FROM [4])[a]

	Algorithm I				Algorithm II				
	Global Maximum		Local Maximum		Global Maximum			Local Maximum	
L	Channel Capacity	Number of Iterations	Channel Capacity	Number of Iterations	Number of Input	Channel Capacity	Number of Iterations	Number of Input	Channel Capacity	Number of Iterations
2	1.00000000	1	1.00000000	1	2	1.00000000	1	2	1.00000000	1
3	1.08746283	1	1.08746283	1	3	1.08746283	1	3	1.08746283	1
4	1.24790640	2	1.24790640	2	5	1.24790661	2	5	1.24790661	2
5	1.37227190	1	1.37227190	1	5	1.37277190	1	5	1.37227190	1
6	1.45797721	1	1.45797721	1	6	1.45797721	1	6	1.45797721	1
7	1.53585166	4	1.53587065	3	7	1.53585162	4	11	1.53586957	3
8	1.60792294	4	1.60795459	3	7	1.60791719	4	12	1.60795116	3
9	1.67148683	5	1.67149199	3	8	1.67148664	5	13	1.67149204	3
10	1.72680318	5	1.72682231	3	12	1.72680062	5	15	1.72682026	3
11	1.77801272	5	1.77802079	3	12	1.77801270	5	16	1.77802110	3
12	1.82583107	6	1.82584637	3	12	1.82583121	6	15	1.82584434	3
13	1.87026751	5	1.87028438	3	15	1.87026667	5	21	1.87028312	3
14	1.91131471	3	1.91139536	3	14	1.91131471	3	13	1.91138040	3
15	1.94962998	3	1.94964308	2	14	1.94963024	3	18	1.94963458	2
16	1.98584264	5	1.98588584	3	17	1.98584285	5	27	1.98587671	3
17	2.02019351	5	2.02021339	3	19	2.02019326	5	22	2.02022505	3
18	2.05274739	5	2.05280604	3	18	2.05274735	5	21	2.05280645	3
19	2.08367546	3	2.08372756	2	18	2.08367550	3	23	2.08372758	2
20	2.11304347	3	2.11305547	2	20	2.11304347	3	24	2.11307271	2
21	2.14113732	3	2.14116808	2	18	2.14113735	3	23	2.14116790	2
22	2.16805408	3	2.16810322	2	21	2.16805421	3	27	2.16810326	2
23	2.19392388	3	2.19397183	2	20	2.19392395	3	26	2.19397166	2
24	2.21887692	3	2.21881660	2	22	2.21876916	3	28	2.21881651	2
25	2.24267607	3	2.24269341	2	23	2.24267615	3	28	2.24269314	2
26	2.26567689	3	2.26568557	2	24	2.26567699	3	31	2.26568556	2
27	2.28799788	3	2.28701704	2	23	2.28789823	2	32	2.28791477	2
28	2.30938179	3	2.30940827	2	25	2.30938190	3	34	2.30940754	2
29	2.33017246	3	2.33020603	2	27	2.33017246	3	35	2.33020748	2
30	2.35033665	3	2.35035823	2	28	2.35033668	3	35	2.35035611	2

[a] $\epsilon = 6 \times 10^{-4}$, $L = 1 - 29$.

TABLE II
A COMPARISON OF PERFORMANCE BETWEEN ALGORITHM I AND ALGORITHM II IN EXAMPLE 4[a]

	Algorithm I						Algorithm II				
	Global Maximum			Local Maximum			Global Maximum			Local Maxim	
$N+1$	Number of θ	Minimax Risk	Number of Iteration	Number of θ	Minimax Risk	Number of Iterations	Number of θ	Minimax Risk	Number of Iterations	Number of θ	Minimax Risk
2	3	0.06250000	1	3	0.06250000	1	3	0.06250000	1	3	0.0625000
3	5	0.04289029	1	5	0.04289029	1	5	0.04289029	2	6	0.0428902
4	7	0.02777444	1	7	0.02777444	1	5	0.02777445	1	5	0.0277744
5	9	0.01705320	1	9	0.01705320	1	6	0.01678058	1	6	0.0167805
6	11	0.01052182	2	11	0.01053708	3	10	0.01052687	4	32	0.0105357
7	13	0.00701030	1	13	0.00701030	1	10	0.00694753	3	14	0.0069963
8	15	0.00500124	1	15	0.00500124	1	9	0.00487785	1	9	0.0048778
9	17	0.00377652	1	17	0.00377652	1	12	0.00374143	3	25	0.0037725
10	19	0.00299463	2	19	0.00299463	2	14	0.00299448	4	15	0.0029830
11	21	0.00244896	1	21	0.00244896	1	13	0.00243175	2	17	0.0024486
12	23	0.00205113	1	23	0.00205113	1	15	0.00205554	3	17	0.0020569
13	25	0.00174578	1	25	0.00174578	1	14	0.00170329	1	14	0.0017032
14	27	0.00150441	1	27	0.00150441	1	16	0.00148033	2	20	0.0015173
15	29	0.00131612	2	29	0.00132544	2	17	0.00131607	2	20	0.0013246
16	31	0.00116728	2	31	0.00116737	2	18	0.00116724	2	21	0.0011664
17	33	0.00103580	2	33	0.00103580	2	20	0.00103497	3	23	0.0010349
18	35	0.00092505	2	35	0.00092505	2	20	0.00092446	2	24	0.0009244
19	37	0.00081958	1	37	0.00081958	1	21	0.00083067	2	24	0.0008306
20	39	0.00074127	1	39	0.00074127	1	22	0.00075043	2	25	0.0007504
21	41	0.00067393	1	41	0.00067393	1	22	0.00065163	1	22	0.0006516
22	43	0.00061521	1	43	0.00061521	1	23	0.00059350	1	23	0.0005935
23	45	0.00056378	1	45	0.00056378	1	25	0.00055123	2	30	0.0005685
24	47	0.00051843	1	47	0.00051843	1	26	0.00051355	2	31	0.0005224

[a] $\epsilon = 1 \times 10^{-5}$, $N = 1 - 23$.

than does Algorithm I. It is particularly true when N becomes large because J grows linearly in N with slope 2. This example shows that Algorithm II is generally superior to Algorithm I in computer implementations for squared error loss. In fact, Algorithm II needs much less computing time than Algorithm I. A potential application of this example to quantization has been discussed in some detail in [8].

VI. Conclusion

Two iterative algorithms (Algorithm I and Algorithm II) for solving minimax rules for general decision problems were presented. Both algorithms are designed based on iterative processes that successively select a finite set of parameters from the original parameter space that is generally uncountable. By means of a sequence of improved finite approximations, the algorithms eventually generate the desired minimax rule. It has been shown in Example 3 of Section V that Algorithm I is preferred to Algorithm II in the sense that at Algorithm I iterates on a finite fixed-size parameter set. On the other hand, Example 4 shows that Algorithm II is better than Algorithm I in the sense that Algorithm I needs more parameters for iterations than does Algorithm II, albeit algorithm II utilizes parameter sets with different sizes. The main difference between these two algorithms is that the Bayesian transitivity property is automatically satisfied for Algorithm II, whereas the BTP must be justified before Algorithm I is used. Consequently, whenever it is not clear that the BTP is valid, Algorithm II is always desirable.

It is worth noting that in a recent study [11] we have shown that there is a resemblance between the algorithms proposed in this paper and Remez's algorithms arising in Chebyshev approximation theory. Based on implementational techniques, Algorithm I is analogous to Remez's second algorithm (or Remez's exchange algorithm) and Algorithm II corresponds to Remez's first algorithm. In particular, the Haar condition imposed in Remez's algorithms has a property similar to the Bayesian transitivity property. This surprising discovery suggests that Algorithms I and II may find applications in digital filter design.

Appendix
An Approximation Theorem of a Continuous Conditional Distribution by a Polynomial Conditional Distribution

In Section IV we showed that, with respect to some common loss functions, the BTP is satisfied for either a finite sample space or a conditional polynomial probability density function. Here we will prove that this property can be even carried through a continuous conditional probability density function by any desired degree of accuracy. Moreover, to prove this assertion, we further establish a general theorem that has its own interest and can also be applied to various Bayes problems to make arguments tractable.

Recall that in the earlier assumptions Θ and A are compact and $L(\theta, a)$ is jointly continuous on $\Theta \times A$. Therefore, L is uniformly bounded by a positive number M. In addition, by the Stone–Weierstrass approximation theorem, for any x in X a continuous conditional probability density function $p(x|\theta)$ on Θ can be approximated by a polynomial $P(x|\theta)$ uniformly on Θ within any assigned degree of accuracy. Let $\{P_n(x|\theta)\}$ be such a sequence of polynomials which uniformly approximate the $p(x|\theta)$. Then for any arbitrarily small $0 < \epsilon < 1$ there is a positive integer $N(x)$ depending on x such that for $n \geq N(x)$ we have

$$|p(x|\theta) - P_n(x|\theta)| < \epsilon/2 \text{ uniformly on } \Theta,$$

i.e.,

$$0 \leq p(x|\theta) - (\epsilon/2) < P_n(x|\theta) < p(x|\theta) + (\epsilon/2). \quad \text{(A.1)}$$

In inequality (A.1) we note that the integer $N(x)$ varies when x ranges over the sample space X. To find an N independent of x we further look at the given probability density function $p(x|\theta)$ which is continuous on $X \times \Theta$. It is apparent that if Θ is compact $p(x|\theta)$ converges to 0 uniformly on Θ as $\|x\| \to +\infty$. Accordingly, for this given $\epsilon > 0$, a positive number C exists such that $p(x|\theta) < \epsilon/2$ uniformly on Θ whenever $\|x\| > C$, and in this case we simply let $P(x|\theta) = 0$. On the other hand, for $\|x\| \leq C$, x lies in a compact set K bounded by C. Without loss of generality, we may assume that $K = \{x \in X | \|x\| \leq C\}$.

Let $K_x = \{y \in K | |p(y|\theta) - p(x|\theta)| < \epsilon/2 \text{ for all } \theta \in \Theta\}$. Then $K = \bigcup_{x \in K} K_x$ and K_x is open because $p(x|\theta)$ is continuous on X. Since K is compact then there is a finite set $F = \{x_i \in K\}$ such that $K = \bigcup_{x_i \in F} K_{x_i}$, where

$$K_{x_i} \equiv \left\{ y \in K | |p(y|\theta) - p(x_i|\theta)| < \frac{\epsilon}{2} \text{ for all } \theta \text{ in } \Theta \right.$$
$$\left. \text{and some } x_i \text{ in } K \right\}.$$

Now let $N = \max_{x_i \in F} N(x_i)$. Then for $n \geq N$ and any x in K there exists an x_i such that $x \in K_{x_i}$ and from (A.1) we have

$$|p(x|\theta) - P_n(x_i|\theta)| \leq |p(x|\theta) - p(x_i|\theta)| + |p(x_i|\theta) - P_n(x_i|\theta)| < (\epsilon/2) + (\epsilon/2) = \epsilon,$$

i.e.,

$$0 < p(x|\theta) - \epsilon < P_n(x_i|\theta) < p(x|\theta) + \epsilon < 1 + \epsilon \quad \text{(A.2)}$$

It is important to note that the (A.2) holds for all $x \in K$, i.e., all $\|x\| \leq C$. In particular, (A.2) holds for probability density functions belonging to exponential families. Of course, if X is finite, (A.2) follows immediately by simply letting $N = \max_{x_i \in X} N(x_i)$.

As we defined earlier, if $p(x_i|\theta) < \epsilon/2$, let $P_n(x_i|\theta) = 0$. Hence, from inequality (A.1) $P_n(x_i|\theta)$ is nonnegative and $\sum_{x_i \in F} P_n(x_i|\theta)$ is bounded and greater than zero. We can define a new conditional probability mass function given θ, $f_{P_n}(x_i|\theta)$, defined on F associated with $P_n(x_i|\theta)$. Let $h(\theta) = \sum_{x_i \in F} P_n(x_i|\theta)\mu(K_{x_i})$ where $\mu(K_{x_i}) = \int_{K_{x_i}} dx$ and μ is the Lebesgue measure. Then $f_{P_n}(x_i|\theta)$ is defined by the following:

$$f_{P_n}(x_i|\theta) = \begin{cases} \dfrac{P_n(x_i|\theta)}{h(\theta)}, & \text{if } p(x_i|\theta) \geq \epsilon/2 \\ 0, & \text{if } p(x_i|\theta) < \epsilon/2 \end{cases} \quad \text{(A.3)}$$

and

$$F_{P_n}(x_i|\theta) = f_{P_n}(x_i|\theta)\mu(K_{x_i}).$$

Obviously, $\Sigma_{x_i \in F} F_{P_n}(x_i|\theta) = 1$ and $F_{P_n}(x_i|\theta)$ is a conditional probability mass function on F given θ. Notice that, although due to a factor $h(\theta)$ appearing in the denominator of $F_{P_n}(x_i|\theta)$, $F_{P_n}(x_i|\theta)$ is not a polynomial in θ we will show, in the sequel, that the $h(\theta)$ can be absorbed into the given prior considered in the problem by defining a new prior. Namely, if a prior α on Θ is considered in a Bayes problem, then we can introduce a modified version of $\alpha(\theta)$ induced by $\alpha(\theta), \hat{\alpha}(\theta)$, as follows:

$$\hat{\alpha}(\theta) \equiv \frac{\alpha(\theta)h(\theta)}{\int_{\Theta} \alpha(\theta)h(\theta)\,d\theta}. \tag{A.4}$$

Therefore, if we let

$$C_{P_n,\alpha} \equiv \int_{\Theta} \alpha(\theta)h(\theta)\,d\theta, \tag{A.5}$$

then $f_{P_n}(x_i|\theta)\hat{\alpha}(\theta)C_{P_n,\alpha} = P_n(x_i|\theta)\alpha(\theta)$. The sequence of $\{P_n(x_i|\theta)\}$ constructed by the Stone-Weierstrauss theorem are all polynomials in θ unless $P_n(x_i|\theta) = 0$ and the constant $C_{P_n,\alpha}$ depends only on the given prior α and $P_n(x_i|\theta)$ which are fixed throughout the problem. In addition, notice that $C_{P_n,\alpha}$ is an expectation of $h(\theta)$ with respect to the prior α, of which $h(\theta)$ is a finite sum of $P_n(x_i|\theta)$ over F, and thus it is again a polynomial in θ. This implies that $C_{P_n,\alpha}$ is indeed determined by all moments of the given prior $\alpha(\theta)$ generated by the polynomials $P_n(x_i|\theta)$ for all $x_i \in F$.

According to the moment method described in Section IV, the Bayesian transitivity property works for any arbitrary polynomial probability density function. Hence, instead of using $p(x|\theta)$ we would rather deal with the sequence of $\{P_n(x_i|\theta)\}$.

For any prior $\alpha \in \Xi$ and any decision function $\delta \in \boldsymbol{D}^*$, we define a Bayes risk $\tilde{r}(\alpha, \delta)$ by

$$\tilde{r}_n(\alpha,\delta) = \int_{\Theta}\left[\sum_{x_i \in F}\int_{K \cap K_{x_i}} L(\theta,\delta(x)) \cdot f_{P_n}(x_i|\theta)\hat{\alpha}(\theta)C_{P_n,\alpha}\,dx\right] d\theta \tag{A.6}$$

and define $r_C(\alpha,\delta) \equiv \int_{\Theta}\int_{\|x\| \le C} L(\theta,\delta(x))p(x|\theta)\alpha(\theta)\,dx\,d\theta$, then it can be shown that for an arbitrarily small ϵ

$$|r_C(\alpha,\delta) - \tilde{r}_n(\alpha,\delta)| < \epsilon.$$

This verifies the following result.

Lemma A1: Given continuous conditional probability $p(x|\theta)$ and a jointly continuous loss function, there exist a finite subset $\boldsymbol{F}$ in X and a sequence of polynomials $\{P_n(x_i|\theta)\}$ for some $x_i \in \boldsymbol{F}$ such that for any prior α on Θ and any decision rule δ in $\boldsymbol{D}^*$, the Bayes risks defined by (A.6) converge to the original Bayes risk for $p(x|\theta)$.

As a matter of fact, a more compact form for Lemma A1 can be proven by straightforward justification and stated as the following theorem:

Theorem A1: Given continuous conditional probability density function $p(x|\theta)$, $\theta \in \Theta$, Θ compact, and an $\epsilon > 0$, then there exists a polynomial approximation $\tilde{p}(x|\theta)$ such that the minimax risk using p and $\tilde{p}$ differ by no more than ϵ. Furthermore, under any polynomial conditional distribution and a jointly continuous loss function, $r(\cdot,\cdot)$ will have at most N_0 local maxima, and thus the BTP is satisfied.

Although the above theorem was proven under the assumption that the sample space X is compact, it can be extended to the case of X not compact, particularly, countably infinite. By means of a truncation technique this can be easily justified by truncating the tails of X and replacing it with a single probability for the truncated tails, such that an increasing nested sequence of such truncations will converge to $r(\alpha,\delta)$.

The significant implication of this theorem is that whenever a Bayes problem is considered it suffices for us to restrict a continuous conditional probability density function to a class of specific probability mass functions on a compact space induced by polynomials constructed from the Stone–Weierstrass approximation theorem so that the resulting Bayes risks will only differ from the original Bayes risk by a negligible amount. As a result, the BTP can be carried through continuous conditional probability by this technique. Furthermore, if we let r_{x_i} be the degree of $P_n(x_i|\theta)$ and $r = \max_{x_i \in F} r_{x_i}$, then the BTP determines J, the size of an initial parameter set, which is chosen for Algorithm I beforehand. In other words, J depends on r and is a function of r. We demonstrate below how this technique is applied to proving the BTP.

If the error between $r(\alpha,\delta)$ and $\tilde{r}_n(\alpha,\delta)$ is negligible, then in any Bayes problem it suffices to consider $\tilde{r}_n(\alpha,\delta)$ rather than $r(\alpha,\delta)$. Let $\delta_{p,\alpha} \in \arg[\min_{\delta} \tilde{r}_n(\alpha,\delta)]$ where $\delta_{p,\alpha}$ is a Bayes rule with respect to the prior α and the probability density function p. Since the following loss functions are convex, we can restrict a decision policy to a nonrandomized rule and denote it by $d_{p,\alpha}$.

1) If the loss function is square error, i.e., $L(\theta, d(x)) = (\theta - d(x))^2$, then the Bayes rule with respect to a prior $\hat{\alpha}$ is given by a posterior conditional mean as follows:

$$\begin{aligned} d_{f_{P_n},\hat{\alpha}}(x) &= E[\Theta|X = x_i] \\ &= \frac{\int_{\Theta} \theta\hat{\alpha}(\theta) f_{P_n}(x_i|\theta)\,d\theta}{\int_{\Theta} \hat{\alpha}(\theta) f_{P_n}(x_i|\theta)\,d\theta} \\ &= \frac{\frac{1}{C_{P_n,\alpha}}\int_{\Theta} \theta\alpha(\theta)h(\theta) f_{P_n}(x_i|\theta)\,d\theta}{\frac{1}{C_{P_n,\alpha}}\int_{\Theta} \alpha(\theta)h(\theta) f_{P_n}(x_i|\theta)\,d\theta} \\ &= \frac{\int_{\Theta} \theta\alpha(\theta) P_n(x_i|\theta)\,d\theta}{\int_{\Theta} \alpha(\theta) P_n(x_i|\theta)\,d\theta}. \end{aligned} \tag{A.7}$$

So, this implies that the decision $d_{f_{P_n},\hat{\alpha}}(x)$ is determined by all moments of α generated by the polynomials $P_n(x_i|\theta)$ for all $x_i \in F$ and θ.

To prove the BTP, we assume that Θ^1 and Θ^2 are two parameter sets in Θ with the same cardinality and the Bayes rule $d_{f_{P_n},\hat{\alpha}}$ is determined by (A.7) with respect to the parameter set Θ^1. To show that d_{α} satisfies the (Θ^1,Θ^2) BTP, we must find a prior β on Θ^2 such that $d_{f_{P_n},\hat{\alpha}}(x) = d_{f_{P_n},\hat{\beta}}(x)$ for every x in X. As we have seen, for any x in X there exists an $x_i \in F$, $x \in K_{x_i}$ and from (A.7) $d_{f_{P_n},\hat{\alpha}}(x)$ is determined by all moments of α on Θ^1 generated by the polynomial $P_n(x_i|\theta)$ and θ in Θ^1. Now applying the moment approach described in Section V to $d_{f_{P_n},\hat{\alpha}}$,

there is a β defined on Θ^2 such that, for all x_i in F,

$$\int_{\Theta^1}\theta\alpha(\theta)\,P_n(x_i|\theta)\,d\theta=\int_{\Theta^2}\theta\beta(\theta)\,P_n(x_i|\theta)\,d\theta \qquad \text{(A.8)}$$

and

$$\int_{\Theta^1}\alpha(\theta)\,P_n(x_i|\theta)\,d\theta=\int_{\Theta^2}\beta(\theta)\,P_n(x_i|\theta)\,d\theta. \qquad \text{(A.9)}$$

These two equations enable us to find a $\hat{\beta}$ defined on Θ^2 which corresponds to $\hat{\alpha}$ in $\tilde{r}_n(\alpha,\delta)$. Let $C_{P_n,\beta}\equiv\int_{\Theta^2}\beta(\theta)h(\theta)\,d\theta$. Then, from (A.8) and (A.9),

$$\frac{\int_{\Theta^1}\theta\alpha(\theta)\,P_n(x_i|\theta)\,d\theta}{\int_{\Theta^1}\alpha(\theta)\,P_n(x_i|\theta)\,d\theta}=\frac{C_{P_n,\beta}\int_{\Theta^2}\theta\left[\dfrac{\beta(\theta)h(\theta)}{C_{P_n,\beta}}\right]\left[\dfrac{P_n(x_i|\theta)}{h(\theta)}\right]d\theta}{C_{P_n,\beta}\int_{\Theta^2}\left[\dfrac{\beta(\theta)h(\theta)}{C_{P_n,\beta}}\right]\left[\dfrac{P_n(x_i|\theta)}{h(\theta)}\right]d\theta}. \qquad \text{(A.10)}$$

Let $\hat{\beta}(\theta)=\beta(\theta)h(\theta)/C_{P_n,\beta}$; then $\hat{\beta}$ is a prior on Θ^2 and (A.10) can be rewritten as follows:

$$\frac{\int_{\Theta^2}\hat{\theta}\beta(\theta)f_{P_n}(x_i|\theta)\,d\theta}{\int_{\Theta^2}\hat{\beta}(\theta)f_{P_n}(x_i|\theta)\,d\theta}$$

which is exactly a Bayes decision on x, $d_{f_{P_n},\hat{\beta}}(x)$ with respect to the prior $\hat{\beta}$ on Θ^2. Moreover, from (A.10)

$$d_{f_{P_n},\hat{\beta}}(x)=d_{f_{P_n},\hat{\alpha}}(x), \qquad \text{for all } x\in X.$$

This proves that $d_{f_{P_n},\hat{\alpha}}$ satisfies (Θ^1,Θ^2) Bayesian transitivity property. In the meantime, to apply Algorithm I we have to know how large J is. From (A.7) we easily derive that $J\geq r+1$ where $r=\max_{x_i\in F}r_{x_i}$ and r_{x_i} is the degree of $P_n(x_i|\theta)$.

2) If the loss function is relative entropy, i.e., $L(\theta,d(x))=\log[p(x|\theta)/d(x)]$, then for any x in X there exists an x_i such that $x\in K_{x_i}$ and the Bayes decision $d_{f_{P_n},\hat{\alpha}}(x)$ with respect to a prior $\hat{\alpha}$ is given by

$$d_{f_{P_n},\hat{\alpha}}(x)=\int_{\Theta}f_{P_n}(x_i|\theta)\,\hat{\alpha}(\theta)\,d\theta=\frac{1}{C_{P_n,\alpha}}\int_{\Theta}P_n(x_i|\theta)\,\alpha(\theta)\,d\theta. \qquad \text{(A.11)}$$

So, the decision $d_{f_{P_n},\hat{\alpha}}(x)$ is also determined by all moments of α on Θ generated by the polynomials $P_n(x_i|\theta)$ for all x_i in F and $C_{P_n,\alpha}$. However, recall that

$$C_{P_n,\alpha}=\sum_{x_i\in F}\int_{\Theta}\alpha(\theta)\,P_n(x_i|\theta)\,d\theta \qquad \text{(A.12)}$$

which implies that the constant $C_{P_n,\alpha}$ is also determined by all moments of α generated by the same polynomials $P_n(x_i|\theta)$ for all x_i in F.

Now let Θ^1 and Θ^2 be two parameter subsets in Θ with the same cardinality. By the moment approach, there exists a β on Θ^2 such that, for any x_i in F,

$$\int_{\Theta^1}P_n(x_i|\theta)\,\alpha(\theta)\,d\theta=\int_{\Theta^2}P_n(x_i|\theta)\,\beta(\theta)\,d\theta \qquad \text{(A.13)}$$

and thus (A.11) can be expressed as

$$\begin{aligned}d_{f_{P_n},\hat{\alpha}}(x)&=\frac{1}{C_{P_n,\alpha}}\int_{\Theta^2}P_n(x_i|\theta)\beta(\theta)\,d\theta\\&=\frac{1}{C_{P_n,\alpha}}\int_{\Theta^2}f_{P_n}(x_i|\theta)\beta(\theta)h(\theta)\,d\theta\\&=\frac{C_{P_n,\beta}}{C_{P_n,\alpha}}\int_{\Theta^2}f_{P_n}(x_i|\theta)\hat{\beta}(\theta)\,d\theta\\&=\frac{C_{P_n,\beta}}{C_{P_n,\alpha}}d_{f_{P_n},\hat{\beta}}(x)\end{aligned}$$

where

$$C_{P_n,\beta}=\int_{\Theta^2}\beta(\theta)h(\theta)\,d\theta \quad\text{and}\quad \hat{\beta}(\theta)=\frac{\beta(\theta)h(\theta)}{C_{P_n,\beta}}.$$

Once again, we note that $C_{P_n,\beta}$ is also determined by all moments of β on Θ^2 generated by the polynomials $P_n(x_i|\theta)$ for all $x_i\in F$.

To prove the (Θ^1,Θ^2) Bayesian transitivity property, we have to show that, for each x in X,

$$d_{f_{P_n},\hat{\alpha}}(x)=d_{f_{P_n},\hat{\beta}}(x),$$

i.e.,

$$\int_{\Theta^1}\alpha(\theta)h(\theta)\,d\theta=\int_{\Theta^2}\beta(\theta)h(\theta)\,d\theta. \qquad \text{(A.14)}$$

Obviously, (A.14) is not generally true. However, as we have seen previously, both constants $C_{P_n,\alpha}$ and $C_{P_n,\beta}$ are, respectively, determined by all moments of priors α and β generated by the same polynomials $P_n(x_i|\theta)$ for all $x_i\in F$. By the moment approach and (A.14), $C_{P_n,\alpha}=C_{P_n,\beta}$. Therefore, the (Θ^1,Θ^2) BTP is satisfied for relative entropy loss. In this case from (A.11) the size of an initial parameter set chosen for Algorithm I, J, is no less than r where r is defined as the same as case **1)**.

Finally, when we make comparisons between the two different loss functions considered in cases **1)** and **2)**, it turns out that

1) For case **1)**, (i.e., square error loss) it follows from (A.8) that the Bayes rule $d_{f_{P_n},\hat{\alpha}}$ with respect to the prior $\hat{\alpha}$ does not depend on the constant $C_{P_n,\alpha}$ since $C_{P_n,\alpha}$ is cancelled out during computations. However, for case **2)** (i.e., relative entropy loss), it can be seen from (A.11) that the Bayes rule $d_{f_{P_n},\hat{\alpha}}$ with respect to the prior $\hat{\alpha}$ does depend on the constant $C_{P_n,\alpha}$ which appears in the denominator of $d_{f_{P_n},\hat{\alpha}}$.

2) As we have noticed, the size of an initial parameter set chosen for Algorithm I relies on the BTP, e.g., $J\geq r+1$ for case **1)**, and $J\geq r$ for case **2)**. This is because in case **1)** the Bayes rule $d_{f_{P_n},\hat{\alpha}}$ in (A.8) is determined by all moments of α generated by all polynomials $P_n(x_i|\theta)$ for all $x_i\in F$ and θ; by contrast, in case **2)** there is no θ generating an extra moment of α.

3) In spite of these differences, the BTP is satisfied for both cases. What is more, since in both cases the constants $C_{P_n,\alpha}$ and $C_{P_n,\beta}$ are, respectively, determined only through by the moments of α and β generated by the same polynomials $P_n(x_i|\theta)$ for all x_i in F, it yields that $C_{P_n,\alpha}=C_{P_n,\beta}$, and thus these two Bayes rules are indeed the same, i.e., $d_{f_{P_n},\hat{\alpha}}=d_{f_{P_n},\hat{\beta}}$.

REFERENCES

[1] D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions*. New York: Wiley, 1954.

[2] E. L. Lehmann, *Testing Statistical Hypotheses*. New York: Wiley, 1959.

[3] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*. Orlando, FL: Academic, 1967.
[4] J. O. Berger, *Statistical Decision Theory*. New York: Springer-Verlag, 1980.
[5] E. L. Lehmann, *Theory of Point Estimation*. New York: Wiley, 1983.
[6] W. Nelson, "Minimax solution of statistical decision problems by iteration," *Ann. Math. Statist.*, vol. 37, pp. 1643–1657, 1966.
[7] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, no. 2, pp. 166–174, Mar. 1980.
[8] C.-I. Chang, "A generalized minimax approach to statistical decision problems with applications to information theory," Ph.D. dissertation, Elec. Eng. Dep., Univ. of Maryland, College Park, May 1987.
[9] C.-I. Chang, S. C. Fan, and L. D. Davisson, "On numerical methods of calculating the capacity of continuous-input discrete-output channels," *Inform. Comput.*, to be published.
[10] C.-I. Chang and L. D. Davisson, "On calculating the capacity of an infinite-input finite (infinite)-output channel," *IEEE Trans. Inform. Theory*, vol. IT-34, no. 5, pp. 1180–1184, Sept. 1988.
[11] ____, "A counterpart of Remez's algorithms in statistical decision theory: Chang–Davisson's algorithms," presented at the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Glasgow, Scotland, May 23–26, 1989, pp. 1345–1348.