SEQUENTIAL CLASSIFICATION ON PARTIALLY ORDERED SETS

Curtis Tatsuoka[1] and Thomas Ferguson[2]

[1] *Department of Statistics, The George Washington University, Washington, DC 20052, USA*
[2] *Department of Statistics, UCLA, Los Angeles, CA 90095, USA*

Abstract

A general theorem on asymptotically optimal sequential selection of experiments is presented and applied to a Bayesian classification problem when the parameter space is a finite partially ordered set. The main results include establishing conditions under which the posterior probability of the true state converges to one almost surely, and determining optimal rates of convergence. Properties of various classes of experiment selection rules are explored.

KEY WORDS: partially ordered set, cognitive diagnosis, group testing, sequential selection of experiment, optimal rates of convergence, Kullback-Leibler information.

1. Introduction: Background and Motivation.

Partially ordered sets are natural models for many statistical applications. As an example of how partially ordered sets (posets) can be used in cognitively diagnostic analysis and computerized intelligent tutoring systems, consider the following example (e.g. see K. Tatsuoka (1995)). Suppose a student is to be tested on a certain subject domain for which there is a known finite set of knowledge states, denoted by S. It is of interest to determine the student's knowledge state in S. Responses from sequentially selected test items (i.e. experiments) will be used to classify the person into one of the states. A natural model for S is to assume that certain states are at higher levels than others. Two states i and j in S may be related to each other in the following manner. If a student in state i has the knowledge to answer correctly all the test items that a student can who is in state j, we denote this by $j \leq i$. It is thus natural to assume that S is a partially ordered set.

Another example of the use of partially ordered sets in statistics with a rich body of literature, is group testing, originated by Dorfman (1943) (see also Ungar (1960), Sobel and Groll (1959) and (1966), Yao and Hwang (1990), and Gastwirth and Johnson (1994)). This is the problem of identifying all defectives in a set of finite objects by experiments that find for a given subset if there is at least one defective in the subset. Note that the classification states, consisting of all subsets of the objects, can also be viewed as partially ordered.

In general, an experiment consists in observing a random variable or vector, X, whose distribution depends on the true unknown state, call it $s \in S$. We assume that for each experiment e and state $s \in S$, the corresponding class conditional response distribution of X has some density $f(x|e, s)$. We assume that the prior distribution of the true state is known, and consider the Bayesian approach. The basic problem is to choose a sequence of experiments sequentially and a stopping rule to determine the true state as quickly as possible.

An experiment is said to *partition* S in the sense that states in the same partition share the same response distribution. We give special consideration to the case when S is a finite poset and experiments are associated with a state $e \in S$ in such a way that the distribution of X has density $f(x)$ if $e \leq s$, and density $g(x)$ otherwise. Such experiments partition S into two subsets. In the context of cognitive diagnosis, this can be interpreted as an experiment, $e \in S$, eliciting one distribution of response, $f(x)$, if the true state of the subject has at least the knowledge contained in state e, and another distribution of response, $g(x)$, otherwise. As an example, f and g could be the densities for Bernoulli responses with corresponding probabilities of success p_u and p_l, where $p_u > p_l$. The value $1 - p_u$ could represent the probability of making a careless mistake on a test item, while p_l could represent the probability of making a lucky guess. In group testing, experiments also have two partitions in S. One of these partitions consists of the union of the states which contain at least one of the objects being tested.

The most basic partially ordered set of interest is the two-element lattice, $S = \{\hat{0}, \hat{1}\}$ with $\hat{0} < \hat{1}$. For this particular model, there is no choice of experiment to complicate the analysis, since experiments associated with state $\hat{0}$ give no information. Indeed, it is a test of two simple hypotheses, and the literature concerning the sequential probability ratio test (SPRT) applies directly here (as in Ferguson (1967) or Chernoff (1972)).

This research is an extension of the work done in cognitive diagnosis by Tatsuoka and Tatsuoka (1987), Tatsuoka (1990), and Tatsuoka (1995). In those papers, student misconceptions are diagnosed using responses to test items. The classification states are collections of discrete cognitive attributes related to the subject domain, and represented by ideal response patterns. Other literature of interest in the area of cognitive modeling includes Falmagne and Doignon (1988). Their work discusses models in which states are partially ordered and closed under union.

In the next section, we present some general theorems on sequential selection of experiments where the state space, S, and the set of experiments, $\mathcal{E}$, are arbitrary finite sets. The concept of separation is introduced and it is noted that an infinite sequence of experiments identifies the true state with probability one if and only if the true state is separated from all the other states infinitely often. We then consider the rate of convergence and find the optimal rate of convergence of the posterior probability of the true state to one. In Section 3, these results are applied to a special case when the experimental response distributions reflect the order structure of an underlying poset model.

In Section 4, we suggest a class of ad hoc rules that have nice asymptotic properties. The rules are simple and shown to be asymptotically consistent. When the state space is a lattice, these rules achieve the optimal rate of convergence as well. However, an example is given to show that when S is not a lattice, the optimal rate may not be achieved in general by a member of this class. All proofs are contained in the Appendix.

A framework for model-fitting and analysis of experiments has been developed in C. Tatsuoka (2002). Using data of actual student responses to test items described in K. Tatsuoka (1990), the proposed sequential methods have been applied to partially ordered cognitive models. Markov Chain Monte Carlo parameter estimation techniques incorporating order constraints have been employed, as described in Gelfand et al. (1992). Poset models and corresponding techniques can be employed in other contexts. One such area

is in medical diagnosis. It may also be useful to use poset models and sequential methods in determining neuropsychological performance (e.g. see The Diabetes Control and Complications Trial Research Group (1996) and Jaeger et al. (1992)).

2. Optimal Rates for Experiment Selection.

We present some general results on sequential selection of experiments. Let $\mathcal{E} = \{1, 2, \ldots, m\}$ be a finite class of experiments, let S be a finite parameter space, and for $e \in \mathcal{E}$, let $f(x|e, j)$ for $j \in S$ be a family of probability densities with respect to a σ-finite measure μ_e on a measurable space $(\mathcal{X}_e, \mathcal{B}_e)$.

Let $\pi_0(j)$ denote the prior distribution of $j \in S$. We assume $\pi_0(j) > 0$ for all $j \in S$. At the first stage, an experiment $e_1 \in \mathcal{E}$ is chosen and a random variable X_1 having density $f(x|e_1, s)$ is observed. The posterior distribution on the parameter space, denoted by $\pi_1(j)$, is proportional to the prior times the likelihood, that is, $\pi_1(j) \propto \pi_0(j) f(x_1|e_1, j)$, where x_1 represents the observed value of X_1. Inductively, at stage n for $n > 1$, conditionally on having chosen experiments e_1, e_2,..., e_{n-1}, and having observed $X_1 = x_1, X_2 = x_2, \ldots, X_{n-1} = x_{n-1}$, an experiment $e_n \in \mathcal{E}$ is chosen and X_n with density $f(x|e_n, s)$ is observed. The posterior distribution then becomes

$$\pi_n(j) \propto \pi_0(j) \prod_{i=1}^{n} f(x_i|e_i, j). \tag{1}$$

The posterior probability structure on S at stage n will be denoted by $\boldsymbol{\pi}_n$.

Let $s \in S$ denote the true parameter value. We seek a sequential selection of experiments, $e_1, e_2, \ldots$, for which the posterior probability of the true state, $\pi_n(s)$, converges a.s. to 1 at the fastest rate. Typically, $1 - \pi_n(s)$ converges to zero at the order $e^{-\alpha n}$ for some α, and α is called the rate of convergence. The mathematical definition of the rate of convergence is taken to be

$$\alpha = \liminf_{n \to \infty} -\frac{1}{n} \log(1 - \pi_n(s)). \tag{2}$$

The basic notion for the following theorems is separation.

Definition. *An experiment, $e \in \mathcal{E}$, is said to separate two states, $i \in S$ and $j \in S$, if the probability distributions $f(x|e, i)$ and $f(x|e, j)$ are distinct.*

In order to obtain convergence, it is sufficient for the sequence of administered experiments to separate the true state, s, from all the others infinitely often (i.o.). If we rule out the case in which there is an experiment, e, and a state, j, such that one observation using e distinguishes between s and j with probability one, the result to follow establishes that doing so is necessary as well as sufficient.

Theorem 2.1. *Let $s \in S$ denote the true state, and let $\boldsymbol{e} = \{e_m\}_{m=1}^{\infty}$ denote a fixed sequence of experiments. Then $\pi_n(s) \to 1$ a.s. as $n \to \infty$ if $\boldsymbol{e}$ separates s from j i.o. for all $j \in S \backslash \{s\}$. Conversely, if for all $e \in \mathcal{E}$ and $j \in S$ the distributions $f(x|e, s)$ and $f(x|e, j)$*

are not mutually singular, then $\pi_n(s) \to 1$ *a.s. as* $n \to \infty$ *only if* $\boldsymbol{e}$ *separates* s *from* j *i.o. for all* $j \in S \backslash \{s\}$.

The proof follows from standard martingale arguments (e.g. Neveu (1974)), and is omitted. An experiment selection procedure which separates i.o. the true state from all others with probability one, no matter what be the true state, is called *convergent*.

Let $K(f, g)$ represent the Kullback-Leibler information for distinguishing f and g when f is true, defined as

$$K(f, g) = \int \log(f(x)/g(x)) f(x) \mu(dx), \tag{3}$$

where μ is a σ-finite measure. The basic property we use of Kullback-Leibler Information is that $K(f, g) \geq 0$ (possibly $+\infty$) for all f and g, and that $K(f, g) = 0$ if and only if f and g are identical as probability measures. We will be dealing with distributions f and g chosen from $f(x|e, j)$, and we use the notation, $K_{e,j}(s)$ to denote the Kullback-Leibler information for distinguishing states s and j when s is true and experiment e is used,

$$K_{e,j}(s) = K(f(\cdot|e, s), f(\cdot|e, j)). \tag{4}$$

For $e \in \mathcal{E}$, let n_e denote the number of administrations of experiment e in the first n stages.

Definition. *An experiment is said to be chosen in limiting proportion* $\beta \geq 0$ *if* $\beta = \liminf_{n\to\infty} n_e/n$. *If* $\beta > 0$, *the experiment is said to be administered in positive limiting proportion.*

The optimal rate of convergence of $\pi_n(s)$ to 1 is related to the following linear program:
Find $p_1, \ldots, p_m$ *and* v *to*

$$\textit{maximize } v \tag{5}$$

subject to

$$p_1 \geq 0, p_2 \geq 0, \ldots, p_m \geq 0 \quad \textit{and} \quad \sum_{e=1}^{m} p_e = 1 \tag{6}$$

and

$$v \leq \sum_{e=1}^{m} p_e K_{e,j}(s) \qquad \textit{for } j \in S,\ j \neq s. \tag{7}$$

The following theorem gives the optimal rate when the true state $s \in S$ is known. In Theorem 2.3, we note that the optimal rate can be achieved even when the true state is unknown.

Theorem 2.2. *Assume all* $K_{e,j}(s)$ *are finite. Let* $v^*(s), p_1^*(s), \ldots, p_m^*(s)$ *be any solution to the linear program (5)-(7). Then if* $v^*(s) > 0$, *the optimal rate of convergence of* $\pi_n(s)$ *to 1 is* $v^*(s)$ *a.s., attained by any sequential experimental selection procedure for which the limiting proportion of experiment* e *is* $p_e^*(s)$ *for all* $e \in \mathcal{E}$.

Remarks.

1. When some of the Kullback-Leibler numbers are allowed to be infinite, the conclusion of this theorem still holds, but under a slight modification. If $K_{e,k}(s)$, say, is infinite, then $\pi_n(k)$ can be made to converge to zero faster than any linear rate, using rules that use experiment e infinitely often but with limiting proportion zero. Such a state, k, can be taken care of using e occasionally, without sacrificing any positive limiting proportion for any of the other states. In other words, in finding the correct limiting proportion for the other states, variable k may be removed from S in (7). When all such variables are removed from S, the linear program gives the correct value for $v^*(s)$ as the optimal rate. If all of S is removed by this process, then $v^*(s) = \infty$ and one can obtain a superlinear rate of convergence.

2. Note that experiment e separates states s_1 and s_2 if and only if $K_{e,s_1}(s_2) > 0$, or equivalently, if $K_{e,s_2}(s_1) > 0$. We say that *a state $s_1 \in S$ is identifiable through $\mathcal{E}$* if for every $s_2 \neq s_1$, there exists an experiment $e \in \mathcal{E}$ that separates s_1 and s_2. We say that S *is identifiable through $\mathcal{E}$* if every $s \in S$ is identifiable through $\mathcal{E}$.

The hypothesis $v^*(s) > 0$ in the above theorem is equivalent to the assumption that s be identifiable through $\mathcal{E}$. This is because if s is identifiable through $\mathcal{E}$, then the simple rule that cycles through the experiments in $\mathcal{E}$ indefinitely yields a positive rate of convergence.

3. We would like to find an experiment selection rule that achieves the optimal rate without knowing in advance the true state. If S is identifiable through $\mathcal{E}$, then a rule that presumes the most probable state is the true one and acts in some appropriate way will achieve the optimal rate not knowing the true state. For example, consider the following simple randomized sequential rule for selecting experiments.

R: At stage $n+1$, find any state $\hat{k}_n \in S$ with the largest posterior probability, $\pi_n(\hat{k}_n) = \max\{\pi_n(k) : k \in S\}$. Then choose as e_{n+1} an experiment $e \in \mathcal{E}$ according to the probability distribution $p_e^*(\hat{k}_n)$.

Theorem 2.3. *If S is identifiable through $\mathcal{E}$, and if $K_{e,j}(s)$ is finite for all e, j and s, then R attains the optimal rate of convergence no matter which state $s \in S$ is the true state.*

A related procedure is described in Chernoff (1972, p. 72) for the problem of testing multiple hypotheses. Though R achieves the optimal rate not knowing the true state, it is unlikely to be efficient for small samples. We would like to choose an experiment that separates the most likely state from those that are next most likely, but R gives no consideration to those other states. In the next section, we specialize to a class of problems structured by a partial order. For these problems, we suggest in Section 4 some efficient algorithms for finding the true state quickly.

3. A Poset of Experiments and States.

For a complete discussion of partially ordered set theory, see Davey and Priestley (1990) or Stanley (1986). Briefly, let S be a partially ordered set. For an element $i \in S$, the set, $\uparrow i = \{j \in S : i \leq j\}$, is known as the up-set of i. The set $\downarrow i = \{j \in S : j \leq i\}$, is known as the down-set of i. If there exists a greatest element $\hat{1}$ in S such that $i \leq \hat{1}$ for all $i \in S$, $\hat{1}$ will be referred to as the top element. Similarly, if there exists a least element $\hat{0}$ in S such that $\hat{0} \leq i$ for all $i \in S$, $\hat{0}$ will be referred to as the bottom element. A *lattice*

is a poset such that any two elements have both a unique least upper bound (join) and a unique greatest lower bound (meet). Note that a finite lattice has both a top and a bottom element.

We assume that the set of classification states, S, is a finite poset containing at least two elements. Experiments are identified with states in S as follows. If X represents the response random variable, then the density of X for a given experiment $e \in S$ and true state $s \in S$ is given by

$$f_X(x|e,s) = \begin{cases} f(x) & \text{if } e \leq s \\ g(x) & \text{otherwise.} \end{cases} \tag{9}$$

If S has a bottom element $\hat{0}$, then the experiment $e = \hat{0}$ gives no information since all states in S will have the same response distribution f. Thus we may take $\mathcal{E} = S \backslash \{\hat{0}\}$ as the set of experiments. The Kullback-Leibler numbers simplify to

$$K_{e,j}(s) = \begin{cases} K(f,g) & \text{if } e \leq s \text{ and } e \not\leq j \\ K(g,f) & \text{if } e \not\leq s \text{ and } e \leq j \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

We assume f and g are not identical distributions and not mutually singular. We note that e separates j and s if the up-set of e contains exactly one of j and s.

An element j is said to cover i if $i < j$ and there does not exist another $k \in S$ such that $i < k < j$. We denote the set, possibly empty, of all covers of an element $i \in S$ by C_i. The *cover separators of* i are

$$K_i = \{k \in S : k \leq c \text{ for some } c \in C_i \text{ and } k \not\leq i\}. \tag{11}$$

As an illustration, consider Figure 3.1 (a lattice). $C_A = \{D, E\}$, and $K_A = \{B, C, D, E\}$.

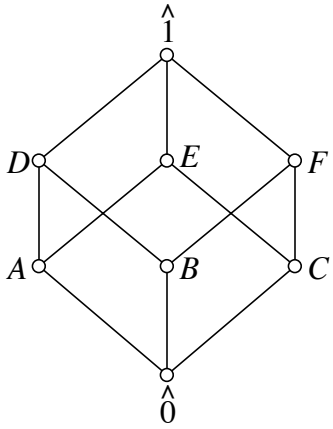

Figure 3.1

If S is a lattice, each element in K_s separates s from exactly one cover in C_s. To see this, suppose there exists an experiment type $j \in K_s$ that separates s from two or more covers. Since s is the greatest lower bound of the two covers, and j is a lower bound of the two covers, we must have $j \leq s$ by the lattice property. This contradicts the assumption that $j \in K_s$. Because of this property, a solution to the linear program of Theorem 2.2 can be found explicitly for such models.

Theorem 3.1. *Suppose that S is a lattice and that both Kullback-Leibler information numbers, $K(f,g)$ and $K(g,f)$, are finite. Let $s \in S$ denote the true state.*

(i) *If $s = \hat{0}$, then the optimal rate of convergence is $K(g,f)/|C_s|$, attained a.s. by experiment selection rules that use the elements in C_s in limiting proportion $1/|C_s|$ each.*

(ii) *Otherwise, the optimal rate is $K(f,g) \cdot K(g,f)/(|C_s| \cdot K(f,g) + K(g,f))$. A rule attains the optimal rate if and only if it uses s itself in limiting proportion $K(g,f)/(|C_s| \cdot K(f,g)+K(g,f))$, and separates each cover of s from s in limiting proportion $K(f,g)/(|C_s| \cdot K(f,g) + K(g,f))$. (Note that if $s = \hat{1}$, then $|C_s|$ is zero, in which case using s in limiting proportion one attains the optimal rate, $K(f,g)$.)*

If $K(f,g) = \infty$ and $s = \hat{1}$ in case (ii), then we can get superlinear convergence of $\pi_n(s)$ to 1 a.s.; in fact, if $\log(f(x)/g(x)) = \infty$ with positive probability under f, then we get the best rate possible: $\pi_n(s)$ will be equal to 1 from some n on a.s. Similarly for case (i) if $K(g,f) = \infty$.

If $s \neq \hat{1}$ in case (ii), the convergence will be linear unless both $K(f,g) = \infty$ and $K(g,f) = \infty$. If $K(f,g) < \infty$ and $K(g,f) = \infty$, then the optimal rate is $K(f,g)$ attained using experiment s in limiting proportion one, provided the covers of s are taken infinitely often. If $K(f,g) = \infty$ and $K(g,f) < \infty$, the optimal rate is $K(g,f)/|C_s|$, attained for instance by using the covers of s in limiting proportions $1/|C_s|$ each, provided s is taken infinitely often.

Although in part (ii) of Theorem 3.1, the optimal rate is attained using any of the cover separators in the correct proportions, second order considerations imply that a cover itself should not be used if there is another separator for that cover. Consider Figure 3.1 with A as the true state. The covers are D and E. But use of B and C instead, has the advantage of making $\pi_n(F)$ go down faster. For small samples, such considerations may lead to substantial improvement. In general, let G_s be the set of all minimal elements of $K_s, s \in S$. Given s is the true state, note that the rules suggested in Section 4 satisfy these second order considerations, as only experiments of type s or elements in G_s are used eventually (cf. Proposition 2 in the Appendix).

Consider again the problem of cognitive diagnosis, and suppose that a discrete set of cognitive attributes has been identified for a given subject domain. Knowledge states can then be associated with subsets of the attributes that are mastered. If all possible subsets of attributes are represented, such a model would be a lattice. For instance, if three attributes are identified, the lattice would be isomorphic to Figure 3.1, and a student in the top element would have mastery of all of the attributes. As for the exponential rates of convergence for these models, this has significant practical ramifications for reducing the number of test items that need to be administered through the use of sequential methods (cf. Tatsuoka (2002)).

A Bayesian formulation of the group testing problem can be treated using (9) by considering the true state, s, as the set of non-defective objects. An experiment, e is just a subset of the objects. The outcome has one distribution if $e \subseteq s$ (no defectives in e) and another if $e \not\subseteq s$ (at least one defective in e). Taking the partial order to be that of inclusion, $e \subseteq s$ is identified with $e \leq s$, giving (9).

4. Experiment Selection Procedures.

Below, several heuristic approaches to the problem of sequential selection of experiments in the poset model of Section 3 are considered. These methods depend on the quantity

$$m_n(e) = \sum_{j \geq e} \pi_n(j),$$

the mass on $\uparrow e, e \in S$, at stage n.

The first heuristic to be proposed is very intuitive. It will be referred to as the *halving algorithm* $H : \boldsymbol{\pi}_n \to \mathcal{E}$, and it chooses the experiment that minimizes

$$h(\boldsymbol{\pi}_n, e) = |m_n(e) - 0.5|, \tag{12}$$

for $e \in \mathcal{E}$ and where $m_n(e)$ is viewed as a function on $\mathcal{E}$. The heuristic H thus selects the experiment that partitions the poset into two parts closest to one-half in terms of mass. Computationally this algorithm is very simple. It does not depend on f or g except through the posterior distributions $\boldsymbol{\pi}_n$.

The basis for the second experiment selection rule to be studied comes from information theory. Denote Shannon entropy by $En(\boldsymbol{\pi}_n) = \sum_{j \in S}(-\log(\pi_n(j)) \cdot \pi_n(j))$. Define the *Shannon entropy procedure* $Sh : \boldsymbol{\pi}_n \to \mathcal{E}$ to select the experiment $e \in \mathcal{E}$ that minimizes

$$sh(\boldsymbol{\pi}_n, e) = \int En(\boldsymbol{\pi}_{n+1}(e, x))(f(x)m_n(e) + g(x)(1 - m_n(e)))\mu(dx) \tag{13}$$

where $\boldsymbol{\pi}_{n+1}(e, x)$ denotes the posterior probability distribution updated to stage $n + 1$ given $e_{n+1} = e$ and $X_{n+1} = x$. Note that the function sh is just the expected entropy taken with respect to the marginal distribution of X_{n+1} having the mixed density

$$f_{X_{n+1}}(x|\boldsymbol{\pi}_n, e_{n+1} = e) = f(x) \cdot m_n(e) + g(x) \cdot (1 - m_n(e)). \tag{14}$$

A related version of sh is

$$sh^*(\boldsymbol{\pi}_n, e) = sh(\boldsymbol{\pi}_n, e) - En(\boldsymbol{\pi}_n).$$

Since $En(\boldsymbol{\pi}_n)$ is not a function of $e \in \mathcal{E}$, one may equivalently minimize sh^* over $\mathcal{E}$ in experiment selection. This heuristic does depend on f and g and thus is more sensitive than H. The advantage of sh^* over sh is that $sh^*(\boldsymbol{\pi}_n, e)$ depends on $\boldsymbol{\pi}_n$ only through $m_n(e)$ as in the following lemma (given without proof).

Lemma 4.1. *The function $sh^*(\boldsymbol{\pi}_n, e)$ depends on $\boldsymbol{\pi}_n$ only through $m_n(e)$ and is convex in $m_n(e)$ for $m_n(e) \in [0, 1]$. Moreover, $sh^*(\boldsymbol{\pi}_n, e) = 0$ if $m_n(e) = 0$ or 1.*

We can thus define a class of experiment selection procedures $\mathcal{U}$ that contains H and Sh as special cases. Let experiment selection procedures in $\mathcal{U}$ be those that choose $e \in \mathcal{E}$ at stage n to maximize $U(m_n(e))$ for some continuous function U, defined on $[0, 1]$, such

that (1) $U(0) = U(1) = 0$, (2) U attains a unique maximum in (0,1), and (3) there exist numbers $0 < k_0 < k_0' < \infty$ and $0 < k_1 < k_1' < \infty$ such that

$$\begin{aligned} k_0 x < U(x) < k_0' x \qquad & \text{for all } x \text{ sufficiently close to 0} \\ k_1(1-x) < U(x) < k_1'(1-x) \qquad & \text{for all } x \text{ sufficiently close to 1.} \end{aligned}$$

We take $\mathcal{U}$ to be the class of all such selection procedures. The halving algorithm, H, for example, is a member of $\mathcal{U}$ associated with the function $U(x) = 0.5 - |x - 0.5|$. This class of heuristics shares some nice properties.

Theorem 4.1. *Every experiment selection procedure in $\mathcal{U}$ is convergent.*

Another level of analysis is determining whether procedures in $\mathcal{U}$ attain optimal rates of convergence. When the underlying model is a lattice, they do.

Theorem 4.2. *If S is a lattice, every experiment selection procedure in $\mathcal{U}$ achieves the optimal rate of convergence for each state in S.*

When the underlying model is a poset but not a lattice, procedures in $\mathcal{U}$ may not always attain the fastest rate. Consider the following example.

Example 4.1. Consider the poset in Figure 4.1. It is assumed that s is the true state, and that $\pi_0(s)$ is close to 1. Suppose f and g are Bernoulli with respective parameters p_u and p_l, $p_u = 1 - p_l = .99$, that $\pi_0(3) = (.5) \cdot \pi_0(5)$, $\pi_0(4) = (.5) \cdot \pi_0(3)$, $\pi_0(6) = (.5) \cdot \pi_0(2)$, $\pi_0(7) = (.5) \cdot \pi_0(6)$, $\pi_0(2) = \pi_0(5)$, and that $\pi_0(A), \pi_0(B), \pi_0(C), \pi_0(a), \pi_0(b)$ are very small. It follows that $\pi_0(5) > \pi_0(3) + \pi_0(4)$ and $\pi_0(2) > \pi_0(6) + \pi_0(7)$. Thus, for any procedure in $\mathcal{U}$, the prior values can be chosen so that experiment A is the most attractive choice. Experiment A will continue to be selected until one more failure than success is observed, after which experiment B then becomes most attractive. By Theorem 4.1, this occurs with probability one. Similarly, once one more failure than success is seen for experiment B, experiment C becomes the best choice. It follows that experiment A will next be most attractive again and so on. Thus, procedures in $\mathcal{U}$ will cycle through experiments A, B and C. Yet, the optimal rate of convergence is obtained instead by administering the set $\{a, b\}$ in equal positive limiting proportion. Thus, procedures in $\mathcal{U}$ do not obtain the optimal rate in this situation.

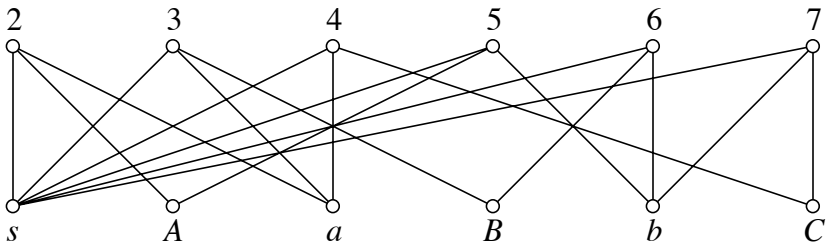

Figure 4.1

DeGroot (1962) describes a class of experiment selection rules that are based on measurable, concave functions on the space of all probability distributions on S. Shannon entropy gives an example of such a function. This class of procedures can be applied under the general conditions of Section 2. For procedures based on strictly concave functions, it can be established that they are convergent using the arguments of DeGroot (p. 406) and Theorem 4.1. See Ben-Bassat (1982) for other classes of experiment selection rules that can be employed under general experimental assumptions.

Appendix. Proofs.

Proof of Theorem 2.2. The rate of convergence is the rate at which $\pi_n(s)$ converges to one. This is the same as the rate at which $\dfrac{1-\pi_n(s)}{\pi_n(s)} = \sum_{j\neq s} \dfrac{\pi_n(j)}{\pi_n(s)}$ converges to zero.

$$\frac{1-\pi_n(s)}{\pi_n(s)} = \sum_{j\neq s} \frac{\pi_0(j)}{\pi_0(s)} \cdot \prod_{i=1}^{n} \frac{f(x_i|e_i,j)}{f(x_i|e_i,s)} = \sum_{j\neq s} \frac{\pi_0(j)}{\pi_0(s)} \exp\{-Z_j\},$$

where

$$Z_j = \sum_{i=1}^{n} \log \frac{f(x_i|e_i,s)}{f(x_i|e_i,j)}. \tag{18}$$

Let $n(e)$ be the (random) number of times that experiment e is used in the first n experiments. Then,

$$Z_j = \sum_{e=1}^{m} \sum_{i=1}^{n(e)} \log \frac{f(x_{e,i}|e,s)}{f(x_{e,i}|e,j)},$$

where $x_{e,1},\ldots,x_{e,n(e)}$ are the $n(e)$ observations among $x_1,\ldots,x_n$ that are taken using experiment e.

Let $Z = \min_{j\neq s} Z_j$. Then

$$\frac{1}{n}\log\frac{1-\pi_n(s)}{\pi_n(s)} = -\frac{1}{n}Z + \frac{1}{n}\log\sum_{j\neq s}\frac{\pi_0(j)}{\pi_0(s)}\exp\{-(Z_j - Z)\}.$$

The second term on the right converges a.s. to 0 as $n\to\infty$ since the sum is bounded above by $1/\pi_0(s)$ and below by $\min_j \pi_0(j)/\pi_0(s)$. The problem then is to choose the sequence of experiments to maximize the lim inf as $n\to\infty$ of

$$\frac{1}{n}Z = \min_{j\neq s} \sum_{e=1}^{m} p_{e,n} \frac{1}{n(e)} \sum_{i=1}^{n(e)} \log \frac{f(x_{e,i}|e,s)}{f(x_{e,i}|e,j)}$$

where $p_{e,n} = n(e)/n$ is the proportion of experiments allocated to experiment e among the first n experiments and $n(e)^{-1}\sum_{i=1}^{n(e)} \log(f(x_{e,i}|e,s)/f(x_{e,i}|e,j))$ is defined as 0 if $n(e)=0$. If experiment e is taken in limiting proportion $p_e^*(s)$, then

$$\frac{1}{n}Z \xrightarrow{a.s.} \min_{j\neq s} \sum_{e=1}^{m} p_e^*(s) K_{e,j}(s) = v^*(s).$$

So the value $v^*(s)$ is achieved in the limit. Moreover, this is the optimal rate because

$$\liminf_{n\to\infty} \frac{1}{n} Z = \liminf_{n\to\infty} \min_{j\neq s} \left[\sum_{e=1}^{m} p_{e,n} K_{e,j}(s) + \sum_{e=1}^{m} p_{e,n} \left(\frac{1}{n(e)} \sum_{i=1}^{n(e)} \log \frac{f(x_{e,i}|e,s)}{f(x_{e,i}|e,j)} - K_{e,j}(s) \right) \right]$$

and the second sum actually converges to zero a.s., and the minimum of the first sum is for all n at most $v^*(s)$.

Proof of Theorem 2.3. Let s denote the true state in S and let k be an element of S, $k \neq s$. From the definition of $v^*(k)$ and $p_e^*(k)$, we have from (7)

$$v^*(k) \leq \sum_{e\in\mathcal{E}} p_e^*(k) K_{e,s}(k).$$

Since S is identifiable throught $\mathcal{E}$, we have $v^*(k) > 0$, so that there exists an experiment $e \in \mathcal{E}$ such that $p_e^*(k) > 0$ and $K_{e,s}(k) > 0$, that is, e separates k and s. If R has $\hat{k}_n = k$ infinitely often, then $\pi_n(k)/\pi_n(s) \to 0$ a.s. as in the proof of Theorem 2.1. But when $\pi_n(k)/\pi_n(s) < 1$, we cannot have $\hat{k}_n = k$. Therefore, $\hat{k}_n = k$ only finitely often. Since k is an arbitrary element of S with $k \neq s$, we must have $\hat{k}_n = s$ for all n greater than some large random N. But then clearly R uses experiment e in limiting proportion $p_e^*(s)$, and Theorem 2.2 shows that R achieves the optimal rate.

Proof of Theorem 3.1. Using (10), the main constraints, (7), of the linear program become

$$v \leq K(f,g) \sum_{\substack{e\leq s \\ e\not\leq j}} p_e + K(g,f) \sum_{\substack{e\not\leq s \\ e\leq j}} p_e \qquad \text{for all } j \in S\backslash\{s\}. \tag{19}$$

We seek a probability vector, $\{p_e\}$ satisfying (6), to maximize the minimum of the right side of (19) over $j \in S\backslash\{s\}$.

(i) Suppose that $s = \hat{0}$. Then the first sum on the right of (19) is empty and we seek to maximize

$$K(g,f) \min_{j\in S\backslash\{s\}} \sum_{e\leq j} p_e. \tag{20}$$

The smallest values of this occur when j are in C_s, the covers of $\hat{0}$. Therefore the smallest terms are maximized by giving $p_e = 0$ for all e not in C_s. The smallest of the terms $K(g,f)p_e$ for $e \in C_s$ is maximized subject to $\sum p_e = 1$ when all the p_e are equal for $e \in C_s$. This common value is $1/|C_s|$, and the optimal rate (the maximum v) is $K(g,f)/|C_s|$.

(ii) Suppose s is not $\hat{0}$. Then there is a state $j < s$. For such a state, the second sum in (19) vanishes. The term p_e for $e < s$ does not occur in (19) unless accompanied in the sum with p_s, so the minimum over j can only be made larger if any mass given to p_e for

$e < s$ is transferred to p_s, i.e. we may take $p_e = 0$ for $e < s$. When this is done, all the inequalities (19) for $j < s$ reduce to

$$v \leq K(f,g)p_s. \tag{21}$$

For those j not comparable to s, i.e. $j \not\leq s$ and $s \not\leq j$, the right side of (19) is at least $K(f,g)p_s$. Therefore, these terms may be ignored. The remaining j are those such that $s < j$. For these, the first sum in (19) disappears and the inequality becomes

$$v \leq K(g,f) \sum_{\substack{e \not\leq s \\ e \leq j}} p_e. \tag{22}$$

For each such j, there is at least one cover c of s such that $c \leq j$. Therefore the minimum of the terms in (22) occurs for some cover of s. If $j = c$ is some cover of s, the sum in (22) is the sum over the set $E_c = \{e : e \leq c, e \not\leq s\}$. Since S is a lattice, these sets are disjoint by the property mentioned in the paragraph before the statement of the theorem. Therefore the minimum of the terms in (22) is maximized subject to the sum of the probabilities of these sets being $1 - p_s$, if and only if all the sets have the same probability, in which case the minimum value of (22) is $K(g,f)(1-p_s)/|C_s|$. The maximum of the minimum of this and (21) occurs when the two are equal, namely when $K(f,g)p_s = K(g,f)(1-p_s)/|C_s|$. This gives $p_s = K(g,f)/(|C_s|K(f,g) + K(g,f))$. From this, the optimal rate and the limiting proportions are easily found to be as stated in the theorem.

Proof of Theorem 4.1. Consider the experiment selection rule associated with an arbitrary function U such that $U(0) = U(1)$ and such that U is strictly increasing to a maximum on $(0,1)$ and strictly decreasing thereafter.

Suppose that $s \in S$ is the true state. Let $N \subseteq S$ be the (random) subset of states not separated from s i.o. By the argument of Theorem 3.1, with probability one the posterior probability of any state not in N converges to zero. Suppose an experiment type, $k \in \mathcal{E}$ does not separate s from any $j \in N$. Then $\uparrow k$ either (a) contains both s and N, or (b) does not contain s nor any element of N. In the former case, $m_n(k) \to 1$ and in the latter, $m_n(k) \to 0$. For all other k, we have $m_n(k) \not\to 0$ or 1, because $\uparrow k$ either (a) contains s and no element of N, or (b) does not contain s but does contain at least one element of N. Hence as the number of observations n gets large, since S is finite, $U(m_n(k))$ will be maximized by some k that separates s from N. There can be no upper bound to the number of times s is separated from N.

We precede the proof of Theorem 4.2 with two propositions the first of which is stated without proof.

Proposition 1. *Let A and B be positive numbers, and $a_1, a_2, \ldots$, and $b_1, b_2, \ldots$ be arbitrary sequences of numbers such that*

$$A_n = \frac{1}{n}\sum_{i=1}^{n} a_i \to A \qquad \text{and} \qquad B_n = \frac{1}{n}\sum_{i=1}^{n} b_i \to -B. \tag{23}$$

From these, given a number D, define a third sequence, $c_1, c_2, \ldots$, by

$$c_n = \begin{cases} \text{next unused } a_j \text{ if } \sum_{i=1}^{n-1} c_i \leq D \\ \text{next unused } b_j \text{ if } \sum_{i=1}^{n-1} c_i > D. \end{cases} \tag{24}$$

Then

$$C_n = \frac{1}{n}\sum_{i=1}^{n} c_i \to 0 \quad \text{and} \quad a(n)/n \to B/(A+B), \quad \text{as } n \to \infty \tag{25}$$

where $a(n)$ is the number of a_i used in C_n.

Proposition 2. *Let G_s denote the set of all minimal elements of K_s. For true state $\hat{0} < s < \hat{1}$ and $U \in \mathcal{U}$, only experiment types in $G_s \cup \{s\}$ are administered eventually a.s. Moreover, all $j \in G_s$ are administered in positive limiting proportion a.s.*

Proof. First note that $\pi_n(s)$ converges to 1 a.s. (cf. Theorem 4.1). Using a contradiction argument, it is straightforward to establish that for $U \in \mathcal{U}$, all $c \in C_s$ are separated from s in positive limiting proportion a.s. and that s is administered in positive limiting proportion a.s. as well. Moreover, for any $k \in \uparrow K_s$ there exists a $j \in G_s$ such that $\uparrow k \subseteq \uparrow j$. Similarly, $\uparrow s \subseteq \uparrow k$, $k \in \downarrow s$. It follows then that among experiment types in $\uparrow K_s$, eventually a.s. only experiment types in G_s will be administered by $U \in \mathcal{U}$, and that a.s. only experiment type s will be administered among those in $\downarrow s$.

Suppose some $z \in \{\uparrow K_s \cup \downarrow s\}^c$ is administered i.o. using $U \in \mathcal{U}$. Then $\pi_n(j)$ for $j \in \uparrow z \cap \{\uparrow s\}^c$ converges to 0 at a faster rate than $\pi_n(j)$ for $j \in \downarrow s \backslash \{s\}$. For $j \in \uparrow z \cap \uparrow s$, $\pi_n(j)$ converges to 0 at a faster rate than $\pi_n(c)$, where c is any cover of s such that $c \leq j$. Therefore $m_n(z)$ eventually remains much smaller than the largest of the $\pi_n(c)$ for all $c \in C_s$ and of the $\pi_n(j)$ for all $j \in \downarrow s \backslash \{s\}$. Hence eventually, $m_n(z)$ remains smaller than the largest of the $m_n(c)$ for all $c \in C_s$ and $(k_1/k_0') \cdot (1 - m_n(s))$. This contradicts the assumption that z is used i.o. by $U \in \mathcal{U}$.

Finally, suppose $j_1, j_2 \in G_s$, $j_1 \neq j_2$, separate s from the same cover in C_s, yet only j_1 is administered in positive limiting proportion. Note then that $\pi_n(j_2)$ is a dominant term, as j_2 is separated from s only by experiment type s in positive limiting proportion. Consider now posterior probability terms in $\uparrow j_1 \cap \uparrow {j_2}^c$. If such terms are in $\uparrow j_1 \cap \uparrow s^c$, these terms are not dominant. For terms in $\uparrow j_1 \cap \uparrow s$, note that these terms are in $\uparrow j_2$ as well, since j_1 and j_2 separate the same cover and S is a lattice. Hence, eventually a.s. experiment type j_2 will be more attractive than j_1.

Proof of Theorem 4.2. Let s denote the true state in S. The arguments to follow hold for any $U \in \mathcal{U}$, and almost surely. Note if $s = \hat{1}$, then by the arguments of Theorem 4.1 eventually only experiment type $\hat{1}$ will be used. Similarly, if $s = \hat{0}$, asymptotically only the experiment types in C_s will be administered, and in equal limiting proportion. Henceforth, assume that $\hat{0} < s < \hat{1}$.

Let $n_{ij} = |\{m \leq n : e_m \leq i \text{ and } e_m \not\leq j\}|$. Let $p_j(n) = n_j/n$ denote the proportion of times among the first n trials that experiment type $j \in S$ is used, and $p_{cs}(n) = n_{cs}/n$ denote the proportion of times among the first n trials that a cover $c \in C_s$ is separated from $s \in S$. For $c \in C_s$, let $G_{cs} = \{j \in G_s : j \leq c\}$. Note that $G_s = \bigcup_{c \in C_s} G_{cs}$, and that since S is a lattice, these sets are disjoint.

The proof will now proceed as follows. Two procedures will be defined, Procedure I and II. Referring to Theorem 3.1, it will be shown that both procedures are optimally convergent given s is the true state. Moreover, these procedures will provide upper and lower bounds for the limiting proportion that experiment types are administered by $U \in \mathcal{U}$.

Define

$$L_c = \{\uparrow c\} \cap \bigcap_{j \not\leq c, j \in G_s} \{\uparrow j\}^c \text{ for } c \in C_s, \text{ and } L_s = \{\uparrow s\}^c \cap \{\uparrow G_s\}^c.$$

Note L_c and L_s are non-empty, containing at least c and $\downarrow s \backslash \{s\}$, respectively. L_c is the subset of states in $\uparrow c$ that are separated from s only by the experiment types in G_{cs} among experiment types in G_s. L_s is the set of states that are separated from s only by experiment s among experiment types in $G_s \cup \{s\}$. Let

$$M_n^c = |S| \frac{k_0'}{k_1} \cdot \sup_{k \in L_c} \pi_n(k) \text{ for } c \in C_s, \text{ and } M_n^s = |S| \frac{k_1'}{k_0} \cdot \sup_{k \in L_s} \pi_n(k).$$

where k_0, k_1, k_0' and k_1' appear in the definition of $\mathcal{U}$. For some large n on, there exists $j_c \in L_c$ and $j_s \in L_s$ independent of n such that

$$\sup_{k \in L_c} \pi_n(k) = \pi_n(j_c), \text{ and } \sup_{k \in L_s} \pi_n(k) = \pi_n(j_s). \tag{26}$$

This follows because eventually, from some stage on, posterior probabilities of elements in L_c and L_s will be updated via Bayes rule with the same multiplier (cf. Proposition 2).

Let us now define Procedure I as follows:

$$\begin{cases} e_n = s \text{ if } M_n^s > \pi_n(c), \text{ all } c \in C_s; \\ \text{otherwise, find } c \in C_s \text{ with the largest corresponding } \pi_n(c)\text{-value,} \\ \text{and select } e_n \text{ by randomizing among } j \in G_{cs} \text{ with equal probability.} \end{cases}$$

Procedure I is convergent when s is the true state. Given $c \in C_s$, we now apply Proposition 1, first noting that Procedure I only administers experiment types in $G_s \cup \{s\}$, and that

$$\frac{\pi_n(c)}{\pi_n(j_s)} = \frac{\pi_0(c)}{\pi_0(j_s)} \cdot \exp\{\sum_{e_i = s} \log \frac{f(x_i)}{g(x_i)} + \sum_{e_i \in G_{cs}} - \log \frac{g(x_i)}{f(x_i)}\}, \ 1 \leq i \leq n.$$

Let

$$D_1 = \log[\sup_{j \in L_s} \frac{\pi_0(j)}{\pi_0(c)} \cdot |S| \frac{k_1'}{k_0}] \text{ and } D_2 = \log[\inf_{j \in L_s} \frac{\pi_0(j)}{\pi_0(c)} \cdot |S| \frac{k_1'}{k_0}].$$

By applying Proposition 1 with $D = D_1$ and then $D = D_2$, and noting that $D_2 \leq \log \frac{\pi_0(j_s)}{\pi_0(c)} \cdot |S| \frac{k_1'}{k_0} \leq D_1$, it follows that for $c \in C_s$

$$\frac{n_s}{n_s + n_{cs}} \to \frac{K(f,g)}{K(f,g) + K(g,f)} \text{ a.s.}$$

Hence, Procedure I is optimally convergent.

Among the terms in $\{\uparrow s\}^c$, the terms in L_s converge to 0 at the slowest rate since eventually they are only separated from s by experiment type s. In particular, then, eventually $|S| \cdot \pi_n(j_s) > 1 - m_n(s)$, where j_s is as in (26). Also note that for $j \leq c$, $c \in C_s$, $m_n(j) \geq \pi_n(c)$. Thus, eventually,

$$\pi_n(c) \geq M_n^s \quad \text{implies} \quad m_n(j) \geq (\frac{k_1'}{k_0}) \cdot (1 - m_n(s)) \text{ for all } j \in G_{cs}.$$

Hence, for large n, whenever separating a given $c \in C_s$ is more attractive than administering s according to Procedure I, it is more attractive for $U \in \mathcal{U}$ to do so as well. Following Proposition 1, this implies that for $U \in \mathcal{U}$ and any $c \in C_s$,

$$\limsup \frac{p_s(n)}{p_{cs}(n)} \leq \frac{K(f,g)}{K(g,f)} \text{ a.s.}$$

Indeed, it is straightforward to see that for $U \in \mathcal{U}$,

$$\limsup p_s(n)/p_{cs}(n) = K(f,g)/K(g,f) \text{ a.s.} \tag{27}$$

(or else $\frac{m_n(j)}{1-m_n(s)} \to 0$ a.s. for all $j \in G_{cs}$).

Consider now Procedure II:

$$\begin{cases} e_n = s \text{ if } \pi_n(\hat{0}) \geq M_n^c \text{ for all } c \in C_s; \\ \text{otherwise, find } c \in C_s \text{ with the largest corresponding } M_n^c\text{-value,} \\ \text{and randomize selection of } e_n \text{ among } j \in G_{cs} \text{ with equal probability.} \end{cases}$$

Procedure II is convergent if s is the true state. Further, by again applying Proposition 1, it can be shown that Procedure II also converges at the optimal rate when s is true.

Given $c \in C_s$, it will now be established that for large n

$$\pi_n(\hat{0}) \geq M_n^c \text{ for } c \in C_s \quad \text{implies} \quad 1 - m_n(s) \geq \frac{k_0'}{k_1} \cdot m_n(j) \text{ for all } j \in G_{cs}. \tag{28}$$

First note that if (28) holds at a given stage when n is large, then when it is more attractive for Procedure II to administer s than to separate $c \in C_s$ from s, it is also more attractive for $U \in \mathcal{U}$ to do so. Therefore, following Proposition 1, if (28) eventually holds for $U \in \mathcal{U}$,

$$\liminf p_s(n)/p_{cs}(n) \geq K(f,g)/K(g,f) \quad \text{a.s.,} \tag{29}$$

and the optimal rates of convergence are attained.

Given $c \in C_s$, the slowest converging terms in $\uparrow G_{cs}$ are either in L_c or $\uparrow G_{cs} \cap \{\uparrow s\}^c$. If the slowest converging terms are indeed in L_c, then (28) does eventually hold. Let

$$\delta_c(n) = \inf_{j \in G_{cs}} p_j(n).$$

The rate of convergence of the slowest converging terms in $\uparrow G_{cs} \cap \{\uparrow s\}^c$ is

$$\liminf[\delta_c(n)K(g,f) + p_s(n)K(f,g)].$$

Note that $\liminf \delta_c(n) > 0$ by Proposition 2. On the other hand, the rate of convergence for the slowest converging terms in L_{cs} is $\liminf p_{cs}(n)K(g,f)$.

Given $U \in \mathcal{U}$, $c \in C_s$, and $\gamma, \epsilon > 0$, there thus exists a large n such that with probability greater than $1 - \gamma/2$, (28) holds for all stages $n+k, k \geq 0$, if

$$\frac{p_s(n+k)}{p_{cs}(n+k)} > \frac{K(f,g)}{K(g,f)} - \epsilon.$$

Also, from (27), there exists a large n such that

$$\frac{p_s(n)}{p_{cs}(n)} = \frac{n_s}{n_{cs}} > \frac{K(f,g)}{K(g,f)} - \frac{\epsilon}{4}.$$

For such an n, let $\theta(n)$ be the smallest integer such that

$$\frac{n_s}{n_{cs} + \theta(n)} \leq \frac{K(f,g)}{K(g,f)} - \frac{\epsilon}{2}.$$

Following Proposition 1, n and hence $\theta(n)$ can be chosen large enough such that with probability greater than $1 - \gamma/2$, for each stage $n+k$, $k \geq 0$,

$$\frac{p_s(n+k)}{p_{cs}(n+k)} > \frac{K(f,g)}{K(g,f)} - \frac{\epsilon}{2} \tag{30}$$

when (28) holds at each stage. Thus, given sufficiently large n, (30) holds for all stages $n+k$, $k \geq 0$, with probability greater than $1 - \gamma$. Since $\gamma, \epsilon > 0$ were chosen arbitrarily, (29) is established.

Acknowledgements Special thanks to Drs. Kikumi and Maurice Tatsuoka, whose work serves as the basis of this research. This work was supported in part by NSF Grant SES-9810202.

References.

Ben-Bassat, M. (1982) Use of Distance Measures, Information Measures and Error Bounds in Feature Evaluation. In *Handbook of Statistics* Volume 2, P. P. Krishnaiah and L. N. Kanal (Eds.), pp. 773-791, Amsterdam: North-Holland.

Chernoff, H. (1972) *Sequential Analysis and Optimal Design*. Philadelphia: SIAM.

Davey, B. A. and Priestley, H. A. (1990) *Introduction to Lattices and Order*. Cambridge: Cambridge University Press.

Degroot, M. (1962) Uncertainty, Information and Sequential Experiments. *Annals of Mathematical Statistics*, **33**, pp. 404-419.

The Diabetes Control and Complications Research Group (1996) Effects of Intensive Diabetic Therapy on Neuropsychological Function in Adults in the Diabetes Control and Complications Trial. *Annals of Internal Medicine*, **124**, pp. 379-388.

Dorfman, R. (1943) The Detection of Defective Members of a Large Population. *Annals of Mathematical Statistics*, **14**, pp. 436-440.

Falmagne, J.-C. and Doignon, J.-P. (1988) A Class of Stochastic Procedures for the Assessment of Knowledge. *British Journal of Mathematical Psychology*, **41**, pp. 1-23.

Ferguson, T. (1967) *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

Gastwirth, J. L. and Johnson, W. O. (1994) Quality Control for Screening Tests: Potential Applications to HIV and Drug Use Detection. *Journal of the American Statistical Association*, **89**, pp. 972-981.

Gelfand, A., Smith, A. and Lee, T.-M. (1992) Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. *Journal of the American Statistical Association*, **87**, pp. 523-532.

Jaeger, J., Berns, S., Tigner, A., and Douglas, E. (1992) Remediation of neuropsychological deficits in psychiatric populations: rationale and methodological considerations. Psychopharmacology Bulletin, **28**, pp. 367-390.

Neveu, J. (1974) *Discrete Parameter Martingales*. Amsterdam: North Holland.

Sobel, M. and Groll, P. (1959) Group Testing to Eliminate Efficiently All Defectives in a Binomial Sample. *Bell System Technical Journal*, **38**, pp. 1179-1252.

Sobel, M. and Groll, P. (1966) Binomial Group-Testing with an Unknown Proportion of Defectives. *Technometrics*, **8**, pp. 631-656.

Stanley, R. (1986) *Enumerative Combinatorics*, Volume I. Monterey, CA: Wadsworth and Brooks/Cole.

Tatsuoka, C. (2002) Data Analytic Methods for Latent Partially Ordered Classification Models. *Applied Statistics*, **51**, pp. 337-350.

Tatsuoka, K. and Tatsuoka, M. (1987) Bug Distribution and Statistical Pattern Classification. *Psychometrika*, **52**, pp. 193-206.

Tatsuoka, K. (1990) Toward an Integration of Item-Response Theory and Cognitive Error Diagnosis. In *Diagnostic Monitoring of Skill and Knowledge Acquisition*, N. Frederiksen, R. Glaser, A. Lesgold and M. Shafto (Eds.), pp. 453-488, Hillsdale, N.J.: Lawrence Erlbaum.

Tatsuoka, K. (1995) Architecture of Knowledge Structures and Cognitive Diagnosis: A Statistical Pattern Classification Approach. In *Cognitively Diagnostic Assessments*, P. Nichols, S. Chipman and R. Brennan (Eds.), pp. 327-359, Hillsdale, N.J.: Lawrence Erlbaum.

Ungar, P. (1960) The Cutoff Point for Group Testing. *Communications in Pure and Applied Mathematics*, **13**, pp. 49-54.

Yao, Y. C. and Hwang, F. K. (1990) On Optimal Nested Group Testing Algorithms. *Journal of Statistical Planning and Inference*, **24**, pp. 167-178.