Statistica Sinica **11**(2001), 303-408

Celebrating the New Millennium*: *Editors' Invited Article - I

SEQUENTIAL ANALYSIS: SOME CLASSICAL PROBLEMS AND NEW CHALLENGES

Tze Leung Lai

Stanford University

Abstract: We give a brief review of the developments in several classical problems of sequential analysis and their applications to biomedicine, economics and engineering. Even though it can only focus on a limited number of topics, the review shows that sequential analysis is still a vibrant subject after six decades of continual development, with fresh ideas brought in from various fields of application and through interactions with other branches of statistics and probability. We conclude with some remarks on the opportunities and challenges ahead.

Key words and phrases: Change-point detection, multi-armed bandits, quality control, sequential point and interval estimation, sequential testing, stochastic approximation.

1. Introduction

In his brief historical account of the subject, Ghosh (1991) dates the rudiments of sequential analysis to the works of Huyghens, Bernoulli, DeMoivre and Laplace on the gambler's ruin problem. He traces the conception of the subject to the sequential sampling inspection procedures of Dodge and Romig (1929), to quality control charts introduced by Shewhart (1931) and to the two-stage designs of Thompson (1933). He then describes the period 1943-1950 as the birth and childhood of the subject, during which Wald and Barnard independently introduced the sequential probability ratio test (SPRT), Wald and Wolfowitz proved its optimality, and Haldane and Stein showed how sequential methods can be used to tackle some unsolved problems in point and interval estimation. The period from 1951 to 1990, described as "from adolescence to adulthood" by him, was marked by many important developments and breakthroughs in the subject. The last decade of the twentieth century, not covered in his account, also witnessed a number of major advances and new directions. In this paper, we review several classical problems in sequential analysis and consider some of the new directions in the last decade and new challenges in the twenty-first century. In particular, we show how these classical problems and new directions are connected to other branches of statistics and probability and to applications in other fields. We also discuss how these interactions and outreach can enrich and

broaden the subject and help meet the challenges ahead. Because of the broad scope of the subject and its vast and multifarious literature, our review can only focus on a limited number of topics that are chosen to relate sequential analysis to other fields and its past to the new challenges.

The first classical problem, which dates back to what Ghosh calls the "birth" of the subject, is the theory of sequential tests of hypotheses. How the problem has evolved from Wald's (1945) seminal paper on testing a simple null versus a simple alternative hypothesis to a relatively complete theory of sequential testing of composite hypotheses is summarized in Section 2. We also indicate the need to modify this theory for practical applications, particularly in the context of comparative clinical trials, and review in Section 2 the development of group sequential designs and statistical methods for their analysis during the past two decades.

Closely related to sequential testing theory is the theory of sequential detection. Section 3 reviews some major developments in sequential change-point detection and diagnosis beginning with the pioneering works of Shewhart (1931) and Page (1954) on quality control charts and culminating in the rich theory and widespread applications at present. It also discusses the connections between the theories of sequential testing and sequential change-point detection.

Another classical problem, which dates back to the birth and childhood years of sequential analysis, is sequential estimation. Section 4 gives a brief review of several different directions in the development of sequential estimation, from the more traditional fixed-accuracy/fixed-width/risk-efficient sequential estimates in the statistics literature to recursive estimators in signal processing and adaptive control in the engineering literature. It also reviews important developments in the long-standing problem of interval estimation following sequential tests, which is of basic importance in statistical inference in clinical trials that may be stopped early during interim analysis.

The theory of recursive estimation in the engineering literature originates from another classical sequential analysis problem, namely, stochastic approximation, introduced by Robbins and Monro (1951). Section 5 reviews not only important developments of stochastic approximation but also its subsequent interactions with stochastic adaptive control in the engineering literature. Another widely studied problem which also has significant impact on stochastic adaptive control is the "multi-armed bandit problem" introduced by Robbins (1952). Section 6 reviews important developments in this problem and their applications to engineering and economics, together with the closely related topic of adaptive treatment allocation in clinical trials.

Section 7 discusses some challenges and opportunities for sequential analysis in the twenty-first century. The experience of the twentieth century during

which the subject was born and grew to maturity sheds light on how to prepare for these opportunities and challenges. We conclude with some remarks on the interdisciplinary nature of the subject, which should therefore interact with other disciplines and other branches of statistics and probability for its healthy development. In view of the rich arsenal of techniques and concepts, methods and theories developed so far, some of which are reviewed in Sections 2-6, sequential analysis is ready to grow outwards and reach new horizons in the twenty-first century.

2. Sequential Tests of Hypotheses: Theory and Applications

Sequential analysis was born in response to demands for more efficient testing of anti-aircraft gunnery during World War II, culminating in Wald's development of the SPRT in 1943 (cf. Wallis (1980)). Let $X_1, X_2, \ldots$ be i.i.d. random variables with a common distribution P. To test the null hypothesis $H : P = P_0$ versus $K : P = P_1$, the SPRT stops sampling at stage

$$N = \inf\{n \geq 1 : R_n \geq A \text{ or } R_n \leq B\}, \tag{2.1}$$

where $A > 1 > B > 0$ are stopping boundaries and $R_n = \prod_{i=1}^{n}(f_1(X_i)/f_0(X_i))$ is the likelihood ratio, f_i being the density of P_i with respect to some common dominating measure ν, $i = 0, 1$. When stopping occurs, H or K is accepted according as $R_N \leq B$ or $R_N \geq A$. The choice of A and B is dictated by the error probabilities $\alpha = P_0\{R_N \geq A\}$ and $\beta = P_1\{R_N \leq B\}$. This simple test was shown by Wald and Wolfowitz (1948) to be the optimal solution of testing H versus K, in the sense that the SPRT minimizes both $E_0(T)$ and $E_1(T)$ among all tests whose sample size T has a finite expectation under both H and K, and whose error probabilities satisfy

$$P_0\{\text{Reject } H\} \leq \alpha \quad \text{and} \quad P_1\{\text{Reject } K\} \leq \beta. \tag{2.2}$$

Note the close analogy between Wald's SPRT and the classical Neyman-Pearson fixed sample size test of the simple null hypothesis H versus the simple alternative K, subject to the type I error constraint $P_0\{\text{Reject } H\} \leq \alpha$. Both tests involve the likelihood ratios R_n and are solutions to natural optimization problems. While the Neyman-Pearson optimization criterion is to maximize the power $P_1\{\text{Reject } H\}$ for a given sample size n and type I error bound α, the Wald-Wolfowitz criterion is to minimize both E_0T and E_1T under the type I and type II error constraints (2.2).

For fixed sample size tests, a first step to extend the Neyman-Pearson theory from simple to composite hypotheses is to consider one-sided composite hypotheses of the form $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ in the case of parametric families

with monotone likelihood ratio in a real parameter θ. In this case, the level α Neyman-Pearson test of $H : \theta = \theta_0$ versus $K : \theta = \theta_1 (> \theta_0)$ does not depend on the alternative θ_1 and has level α for testing the composite hypotheses H_0 versus H_1. Thus, the ability to reduce the composite hypotheses H_0 versus H_1 to the problem of simple hypotheses H versus K gives an optimal solution (in the sense of uniformly most powerful level α tests) in this case. In the sequential setting, however, we cannot reduce the optimality considerations for one-sided composite hypotheses to those for simple hypotheses even in the presence of monotone likelihood ratio. For example, let $X_1, X_2, \ldots$, be i.i.d. normal random variables with mean θ and variance 1. To test $H_0 : \theta \le \theta_0$ versus $H_1 : \theta \ge \theta_1 (> \theta_0)$ with type I and type II error probabilities not exceeding α and β, one can use the SPRT of $H : \theta = \theta_0$ versus $K : \theta = \theta_1$ with type I and type II error probabilities α and β. However, while this SPRT has minimum expected sample size at $\theta = \theta_0$ and at $\theta = \theta_1$ by the Wald-Wolfowitz theorem, its maximum expected sample size over θ can be considerably larger than the optimal fixed sample size. This led Kiefer and Weiss (1957) to consider the problem of minimizing the expected sample size at a given θ^* subject to error probability constraints at θ_0 and θ_1 in a one-parameter exponential family with natural parameter θ. Hoeffding (1960) derived a lower bound on $E_{\theta^*} T$ subject to error probability constraints at θ_0 and θ_1. Lorden (1976) showed that an asymptotic solution to the Kiefer-Weiss problem is a 2-SPRT with stopping rule of the form

$$\tilde{N} = \inf\{n : \prod_{i=1}^{n} (f_{\theta^*}(X_i)/f_{\theta_0}(X_i)) \ge A_0 \ \text{ or } \ \prod_{i=1}^{n} (f_{\theta^*}(X_i)/f_{\theta_1}(X_i)) \ge A_1\}. \quad (2.3)$$

For the special case of a normal family with mean θ, he also showed numerically that $E_{\theta^*} \tilde{N}$ is close to Hoeffding's lower bound. In this normal case, $\tilde{N}$ reduces to the triangular stopping boundary introduced by Anderson (1960), and has been shown by Lai (1973) to be an approximate solution to the optimal stopping problem associated with the Kiefer-Weiss problem. Making use of Hoeffding's (1960) lower bound on $E_{\theta^*} T$, Hall (1980) derived a family of tests, called "minimum probability ratio tests", that include Lorden's 2-SPRT as a special case.

Ideally the θ^* in (2.3), where we want to minimize the expected sample size, should be chosen to be the true parameter value θ that is unknown. For the problem of testing $H_0 : \theta \le \theta_0$ versus $H_1 : \theta \ge \theta_1 (> \theta_0)$ in the exponential family, replacing θ^* in (2.3) by its maximum likelihood estimate $\hat{\theta}_n$ at stage n leads to Schwarz's (1962) test which he derived as an asymptotic solution to the Bayes problem of testing H_0 versus H_1 with 0-1 loss and cost c per observation, as $c \to 0$ while θ_0 and θ_1 are fixed. For the case of a normal mean θ, Chernoff (1961, 1965) derived a different and considerably more complicated approximation to the Bayes test of $H_0' : \theta < \theta_0$ versus $H_1' : \theta > \theta_0$. In fact, setting $\theta_1 = \theta_0$

in Schwarz's test does not yield Chernoff's test. This disturbing discrepancy between the asymptotic approximations of Schwarz (assuming an indifference zone) and Chernoff (without an indifference zone separating the one-sided hypotheses) was resolved by Lai (1988), who gave a unified solution (to both problems) that uses a stopping rule of the form

$$\hat{N} = \inf\left\{n : \max\left[\sum_{i=1}^{n} \log \frac{f_{\hat{\theta}_n}(X_i)}{f_{\theta_0}(X_i)}, \sum_{i=1}^{n} \log \frac{f_{\hat{\theta}_n}(X_i)}{f_{\theta_1}(X_i)}\right] \geq g(cn)\right\} \tag{2.4}$$

for testing H_0 versus H_1, and setting $\theta_1 = \theta_0$ in (2.4) for the test of H_0' versus H_1'. The function g in (2.4) satisfies $g(t) \sim \log t^{-1}$ as $t \to 0$ and is the boundary of an associated optimal stopping problem for the Wiener process. By solving the latter problem numerically, Lai (1988) also gave a closed-form approximation to the function g.

This unified theory for composite hypotheses provides a bridge between asymptotically optimal sequential and fixed sample size tests. In the fixed sample size case, the Neyman-Pearson approach replaces the likelihood ratio by the generalized likelihood ratio (GLR), which is also used in (2.4) for the sequential test. Since the accuracy of $\hat{\theta}_n$ as an estimate of θ varies with n, (2.4) uses a time-varying boundary $g(cn)$ instead of the constant boundary in (2.3) (with $A_0 = A_1$) where θ is completely specified. Simulation studies and asymptotic analysis have shown that $\hat{N}$ is nearly optimal over a broad range of parameter values θ, performing almost as well as (2.3) that assumes θ to be known; see Lai (1988). This broad range covers both fixed alternatives, at which the expected sample size is of the order $O(|\log c|)$, and local alternatives θ approaching θ_0 (with $\theta_1 \to \theta_0$ also) as $c \to 0$, at which the expected sample size divided by $|\log c|$ tends to ∞. In other words, $\hat{N}$ can adapt to the unknown θ by learning it during the course of the experiment and incorporating the diminishing uncertainties in its value into the stopping boundary $g(cn)$. Lai and Zhang (1994) have extended these ideas to construct nearly optimal sequential GLR tests of one-sided hypotheses concerning some smooth scalar function of the parameter vector in multiparameter exponential families, with an indifference zone separating the null and alternative hypotheses and also without an indifference zone. Lai (1997) has provided further extension to a general class of loss functions and prior distributions, thereby unifying (2.4) with another type of sequential tests involving mixture likelihood ratios which were introduced by Robbins (1970) and whose asymptotic optimality properties under certain loss functions were subsequently established by Pollak (1978) and Lerche (1986). For $n \geq n_c$ with $n_c/|\log c| \to \infty$, the GLR statistics can be replaced by Rao-type score statistics. These score statistics can be extended to nonparametric/semiparametric models,

providing a complete analogue of the corresponding large-sample theory for fixed sample size tests.

Ghosh (1970) summarizes the extensive literature on sequential tests of composite hypotheses in multiparameter families that had grown steadily from 1945 to 1970. With the exception of Bartlett (1946), Cox (1963) and a few others, most of the authors during this period used invariance to reduce certain composite hypotheses to simple ones involving a single parameter for the distribution of the maximal invariant, instead of using GLR statistics or score statistics that have been widely used in fixed sample size tests. Because of this reduction, Wald's SPRT can again be applied, with R_n in (2.1) now given by $R_n = f_{1,n}(M_n)/f_{0,n}(M_n)$, where M_n is a maximal invariant (based on n observations) with respect to a group of transformations leaving the problem invariant, and $f_{i,n}$ is the density function of M_n under H_i, $i = 0, 1$. For these invariant SPRTs, $\{\log R_n, n \geq 1\}$ is no longer a random walk, and the arguments in the proof of the Wald-Wolfowitz theorem on the optimum character of Wald's SPRT are no longer applicable. By making use of approximations to $\log R_n$ and nonlinear renewal theory, Lai (1981) developed asymptotic approximations to the expected sample size of an invariant SPRT under prescribed probability constraints and showed that it is asymptotically minimal in the sense that for $i = 0, 1$, $E_i N$ differs from $\inf\{E_i T : T$ is the stopping time of an invariant sequential test with type I and type II error probabilities α and $\beta\}$ by at most $O(1)$ as $\alpha + \beta \to 0$ such that $\alpha \log \beta + \beta \log \alpha \to 0$. The $O(1)$ term is due to the overshoot $\log(R_N/A)$ or $\log(R_N/B)$ of the invariant SPRT, analogous to Wald's (1945) lower bound for the expected sample size of sequential tests, which is attained by the SPRT when overshoots of the SPRT are ignored.

The reduction of composite hypotheses to simple ones using invariance may require unduly strong restrictions on the composite hypotheses. For example, consider the problem of testing whether the mean μ of a normal distribution is 0 when the variance σ^2 is unknown. To be able to reduce the composite hypotheses on (μ, σ^2) to simple ones on $\theta = \mu/\sigma$ by scale invariance, one has to formulate the null hypothesis as $H : \theta = 0$ and pick an alternative hypothesis $K : \theta = \theta_1$. Although the invariant SPRT, which is the sequential t-test in this case (cf. Ghosh (1970)), is asymptotically optimal for testing H versus K, it is no longer asymptotically optimal for testing $H_0 : \theta \leq 0$ versus $H_1 : \theta \geq \theta_1$ (in the case $\theta_1 > 0$), similar to the case of known variance discussed above. On the other hand, it is relatively straightforward to use GLR statistics to test H_0 versus H_1, or even to test $H_0' : \theta = 0$ versus $H_1' : \theta \neq 0$ (without an indifference zone), in conjunction with a time-varying threshold of the form $g(cn)$ as in (2.4). Such sequential GLR tests have nearly optimal frequentist properties over a wide range of parameter values (ranging from θ near 0 to large θ) and are asymptotically

Bayes with respect to a large class of prior distributions; see Lai and Zhang (1994) and the simulation studies therein. The GLR statistics and the threshold $g(cn)$ basically "self-tune" the test to the unknown alternatives, irrespective of whether they are near 0 (local alternatives) or sufficiently far away from 0 (non-local alternatives). In the asymptotic theory, the non-local alternatives involve large deviation approximations to the boundary crossing probabilities, while the local alternatives involve moderate deviation approximations and functional central limit theorems.

Roughly speaking, functional central limit theorems are applicable only to contiguous alternatives that differ from the null by $O(n^{-1/2})$, while moderate deviation approximations are for alternatives further away from, but still within $o(1)$, of the null hypothesis, as the sample size n becomes infinite. When n is a stopping time (and therefore random), the $o(1)$ above refers to convergence in probability, while the O refers to boundedness in probability. Functional central limit theorems have been widely used in the analysis of nonparametric tests; see Sen (1981, 1991) for comprehensive overviews. The distinction between functional central limit theorems and moderate deviation approximations in the derivation and analysis of stopping rules will be discussed further near the end of Section 3 in the context of sequential change-point detection. Lemmas 4 and 9 of Lai (1988) illustrate such distinction in the context of the sequential GLR test (2.4).

Sequential life testing and acceptance sampling procedures, which were introduced in the decade after Wald's pioneering work on the SPRT, have been widely used by governments and industries; see Epstein and Sobel (1955), MIL-STD 781C (1977) and Basu's (1991) survey. The SPRT (2.1) was extended to handle sequential decision problems with 3 or more hypotheses by Sobel and Wald (1949), Armitage (1950) and Simons (1967). Sequential testing of 3 or more hypotheses has been applied to a variety of engineering problems, including target detection in multi-resolution radar, pattern recognition and machine learning, fault detection and isolation, cf. Marcus and Swerling (1962), Fu (1968) and Lai (2000). Closely related to the sequential multi-hypothesis testing problem are problems of sequential selection and ranking of 3 or more populations. Bechhofer, Kiefer and Sobel (1968) give a comprehensive treatment of the so-called "indifference zone approach" using vector-at-a-time sampling schemes. Gupta and Panchapakesan (1991) review subsequent developments and other approaches, including the "subset selection approach" and adaptive sampling schemes.

Hypothesis testing has been widely accepted by the biomedical community as a means of assessing the reproducibility of the results of an experiment. Within a few years after Wald's introduction of the SPRT, it was recognized that sequential hypothesis testing might provide a useful tool in clinical trials to test

new medical treatments. A number of papers appeared during the 1950s on modifications of the SPRT for the design of clinical trials, and an overview of these developments was given in Armitage's (1960) book, which was subsequently reviewed by Anscombe (1963). In his review, Anscombe introduced a decision-theoretic model for clinical trials to select the better of two treatments for treating a specified number N of patients. The trial phase involves pairwise allocation of treatments to n pairs of patients, after which the apparently superior treatment is given to the remaining $N - 2n$ patients. The pairwise treatment differences $Z_1, Z_2, \ldots, Z_N$ are assumed to be i.i.d. normal with mean δ and variance 1. If the magnitude $|\delta|$ is known, then the optimal fixed sample size can be shown by differentiation (assuming n to be a continuous variable) to be the solution of the equation $2g(|\delta|\sqrt{n}) = N/n$, where $g(0) = 0$ and $g(x) = 1 + \{2\Phi(x) - 1\}/x\varphi(x), \varphi$ and Φ being the standard normal density and distribution function, respectively. Since $|\delta|$ is unknown in Anscombe's model, Lai, Levin, Robbins and Siegmund (1980) proposed to estimate it during the course of the trial, yielding the stopping rule $T^* = \inf\{k : 2g(|S_k|/\sqrt{k}) \geq N/k\}$, where $S_k = Z_1 + \cdots + Z_k$. Anscombe (1963) proposed to put a normal prior distribution on δ and compute the Bayes solution to the corresponding optimal stopping problem. He did not carry out such computation, however, and argued heuristically that $T_A = \inf\{k : 1 - \Phi(|S_k|/\sqrt{k}) \leq k/N\}$ should provide a good approximation to the Bayes stopping rule. Subsequent computation of the Bayes rule T_B by Lai, Levin, Robbins and Siegmund (1980) and Chernoff and Petkau (1981) showed that this is actually not the case. However, the asymptotic results and simulation studies in Lai, Levin, Robbins and Siegmund (1980) show that T^*, T_A, T_B (and in fact a general class of rules including them as special cases) are all asymptotically optimal for large N and have similar performance as the optimal fixed sample size test that assumes known $|\delta|$, even for a patient horizon N as small as 100. Thus, suitably devised sequential procedures can self-tune themselves to unknown parameters that can be learned during the course of the trial.

In 1969, Armitage, McPherson and Rowe proposed a new alternative to the SPRT and its variants, called the "repeated significance test" (RST). The underlying motivation for the RST is that, since the strength of evidence in favor of a treatment from a clinical trial is conveniently indicated by the results of a conventional significance test, it is appealing to apply the significance test, with nominal significance level α, repeatedly during the trial. Noting that the overall significance level α^*, which is the probability that the nominal significance level is attained at some stage, is larger than α, they developed a recursive numerical algorithm to compute α^* in the case of testing a normal mean θ with known

variance σ^2, for which the RST of $H_0 : \theta = 0$ is of the form

$$T = \inf\{n \leq M : |S_n| \geq a\sigma\sqrt{n}\}, \tag{2.5}$$

rejecting H_0 if $T < M$ or if $T = M$ and $|S_M| \geq a\sigma\sqrt{M}$, where $S_n = X_1 + \cdots + X_n$. Haybittle (1971) proposed the following modification of the RST to increase its power. The stopping rule has the same form as (2.5) but the rejection region is modified to $T < M$ or $|S_M| \geq c\sigma\sqrt{M}$, where $a(\geq c)$ is so chosen that the overall significance level is equal to some prescribed number. In particular, $a = \infty$ gives the fixed sample size test while $a = c$ gives the RST.

In double blind multicenter clinical trials, it is not feasible to arrange for continuous examination of the data as they accumulate to perform the RST. This led Pocock (1977) to introduce a "group sequential" version of (2.5), in which the X_n represents an approximately normally distributed statistic of the data in the nth group (instead of the nth observation) and M represents the maximum number of groups. Instead of the square-root boundary $a\sigma\sqrt{n}$, O'Brien and Fleming (1979) proposed to use a constant stopping boundary in

$$T = \inf\{n \leq M : |S_n| \geq b\}, \tag{2.6}$$

which corresponds to the group-sequential version of Wald's SPRT.

While sequential analysis had an immediate impact on weapons testing when it was introduced during World War II to reduce the sample sizes of such tests (cf. Wallis (1980)), its refinements for testing new drugs and treatments received little attention from the biomedical community until the Beta-Blocker Heart Attack Trial (BHAT) that was terminated in October 1981, prior to its prescheduled end in June 1982. The main reason for this lack of interest is that the fixed sample size (i.e., the number of patients accrued) for a typical trial is too small to allow further reduction while still maintaining reasonable power at the alternatives of interest, as pointed out by Peto (1985). On the other hand, BHAT, which was a multicenter, double blind, randomized placebo-controlled trial to test the efficacy of long-term therapy with propranolol given to survivors of an acute myocardial infarction (MI), drew immediate attention to the benefits of sequential methods not because it reduced the number of patients but because it shortened a four-year study by 8 months, with positive results for a long-awaited treatment for MI patients. The trial started in June 1978 and was scheduled for 4 years, with all patients accrued within the first 27 months and with periodic reviews of the data by a Data and Safety Monitoring Board (DSMB) at 11, 16, 21, 28, 34, 40 and 48 months. These interim reviews of the data were incorporated into the trial design mainly for ethical reasons, so that patients would not be exposed to unsafe or ineffective treatment regimens. Actually the group sequential methods

available at that time were not quite adequate to perform the interim analyses. The DSMB used some informal arguments based on stochastic curtailment, together with a formal statistical test involving the O'Brien-Fleming boundary applied to the null variances of the time-sequential logrank statistics (instead of to the total sample size up to the nth group in (2.6)), whose joint asymptotic normality was just established by Tsiatis (1981); see BHAT (1982) and DeMets, Hardy, Friedman and Lan (1984).

The past two decades following the "success story" of BHAT witnessed not only steadily increasing use of group sequential designs in clinical trials to test the efficacy and safety of new drugs and treatments, but also major advances in the development of group sequential methods in clinical trials, beginning with the influential work of Lan and DeMets (1983) on using a "type I error spending function" to modify a fully sequential procedure that has a prescribed maximum sample size into a group sequential procedure when the group sizes are unequal and unknown at the beginning of the trial; see Lan and DeMets (1989), Jennison and Turnbull (1991, 2000) and Gu and Lai (1998) for reviews of these advances. In particular, inspired by the statistical issues raised by BHAT, a number of important and difficult problems concerning the design and analysis of clinical trials with failure-time endpoints and interim analyses have been resolved.

We now briefly describe these problems and related developments in time-sequential methods for survival analysis. Suppose that a clinical trial to compare times to failure between two treatment groups X and Y involves n patients who enter the trial serially, are randomly assigned to treatment X or Y and are then followed until they fail or withdraw from the study or until the study is terminated. Let $T_i' \geq 0$ denote the entry time and $X_i > 0$ the survival time (or time to failure) after entry of the ith subject in treatment group X and let T_j'' and Y_j denote the entry time and survival time after entry of the jth subject in treatment group Y. Thus the data at calendar time t consist of $(X_i(t), \delta_i'(t)), i = 1, \ldots, n'$, and $(Y_j(t), \delta_j''(t)), j = 1, \ldots, n''$, where

$$X_i(t) = X_i \wedge \xi_i' \wedge (t - T_i')^+, \quad Y_j(t) = Y_j \wedge \xi_j'' \wedge (t - T_j'')^+,$$
$$\delta_i'(t) = I_{\{X_i(t) = X_i\}}, \quad \delta_j''(t) = I_{\{Y_j(t) = Y_j\}},$$

and $\xi_i'(\xi_j'')$ denotes the withdrawal time, possibly infinite, of the ith (jth) subject in treatment group $X(Y)$. At a given calendar time t, one can compute, on the basis of the observed data from the two treatment groups, a rank statistic of the general form considered by Tsiatis (1982):

$$S_n(t) = \sum_{i=1}^{n'} \delta_i'(t) Q_n(t, X_i(t)) \left\{ 1 - \frac{m_{n,t}'(X_i(t))}{m_{n,t}'(X_i(t)) + m_{n,t}''(X_i(t))} \right\}$$
$$- \sum_{j=1}^{n''} \delta_j''(t) Q_n(t, Y_j(t)) \frac{m_{n,t}'(Y_j(t))}{m_{n,t}'(Y_j(t)) + m_{n,t}''(Y_j(t))}, \tag{2.7}$$

where $m'_{n,t}(s) = \sum_{i=1}^{n'} I_{\{X_i(t)\geq s\}}$, $m''_{n,t}(s) = \sum_{j=1}^{n''} I_{\{Y_j(t)\geq s\}}$, and $Q_n(t,s)$ is some weight function satisfying certain measurability assumptions. The case $Q_n \equiv 1$ corresponds to the logrank statistic. Letting $H_{n,t}$ denote a product-limit-type estimator of the common distribution function of the two treatment groups under the null hypothesis, based on $\{(X_i(t), \delta_i(t), Y_j(t), \delta_j(t)) : i \leq n', j \leq n''\}$, Prentice's (1978) generalization of the Wilcoxon statistic is the statistic (2.7) with $Q_n(t,s) = 1 - H_{n,t}(s)$, which was extended by Harrington and Fleming (1982) to the case $Q_n(t,s) = (1 - H_{n,t}(s))^\rho$ with $\rho \geq 0$.

Let F and G denote the distribution functions of X_i and Y_j, respectively. Assuming the $\{T'_i\}, \{T''_j\}, \{\xi'_i\}, \{\xi''_j\}$ to be i.i.d. sequences, Tsiatis (1982) showed that under the null hypothesis $H_0 : F = G, (S_n(t_1), \ldots, S_n(t_k))/\sqrt{n}$ has a limiting multivariate normal distribution for any k and $0 \leq t_1 < \cdots < t_k$, for a large class of two-sample rank statistics that includes the logrank statistics considered previously in Tsiatis (1981). Earlier, assuming a Lehmann (proportional hazards) family of the form $1 - G(s) = (1 - F(s))^{1-\theta}$, Jones and Whitehead (1979) considered the use of time-sequential logrank statistics $S_n(t)$ to test sequentially over time the one-sided null hypothesis $H'_0 : \theta \leq 0$. They suggested plotting $S_n(t)$ versus $V_n(t)$, where $V_n(t)$ is Mantel's (1966) estimate of the variance of $S_n(t)$ under $F = G$. They argued heuristically that $\{(V_n(t), S_n(t)), t \geq 0\}$ should behave approximately like $\{(v, W(v)), v \geq 0\}$, where $W(v)$ is the standard Wiener process under $\theta = 0$ and is a Wiener process with drift coefficient depending on θ under alternatives near 0. Using this Wiener process approximation, they suggested replacing $(v, W(v))$ in a sequential test for the sign of the drift of a Wiener process by $(V_n(t), S_n(t))$ to construct a corresponding sequential logrank test of H'_0, and considered in particular the case where the sequential test based on $(v, W(v))$ is an SPRT. Sellke and Siegmund (1983) established weak convergence of $\{S_n(t)/\sqrt{n}, t \geq 0\}$ to a zero-mean Gaussian process with independent increments under $F = G$ and general arrival and withdrawal patterns, thus providing a rigorous asymptotic justification of the heuristics of Jones and Whitehead (1979) under $H_0 : \theta = 0$. Gu and Lai (1991) later showed that $\{(V_n(t)/n, S_n(t)/\sqrt{n}), t \geq 0\}$ converges weakly to $\{(v, W(v)), v \geq 0\}$ under contiguous proportional hazards alternatives, where $W(v)$ is a Wiener process with $EW(v)/v = c$, thus giving a rigorous asymptotic justification of the heuristics of Jones and Whitehead under $H_1 : \theta = c/\sqrt{n}$.

For a general weight function of the form $Q_n(t,s) = \psi(H_{n,t}(s))$ in (2.7), Gu and Lai (1991) showed that $\{S_n(t)/\sqrt{n}, t \geq 0\}$ converges weakly to a Gaussian process with independent increments and variance function $V(t)$ under the null hypothesis and contiguous alternatives. The mean function of the limiting Gaussian process is 0 under the null hypothesis and is of the form $\mu_g(t)$ under

contiguous alternatives that satisfy

$$\int_0^{t^*} |\frac{d\Lambda_G}{d\Lambda_F} - 1| d\Lambda_F = O(\frac{1}{\sqrt{n}}),\ \sqrt{n}\{\frac{d\Lambda_G}{d\Lambda_F}(s) - 1\} \to g(s)$$

as $n \to \infty$, uniformly over closed subintervals of $\{s \in [0, t^*] : F(s) < 1\}$, where Λ_F and Λ_G are the cumulative hazard functions of F and G. In the case of the asymptotically optimal score function $\psi(\cdot) = g(F^{-1}(\cdot))$ for these alternatives, $\mu_g(t) = V(t)$ and therefore the Jones-Whitehead framework can be extended from the logrank score function to general ψ. In practice, the actual alternatives are unknown and μ_g need not even be monotone when ψ is not optimal for the actual alternatives, such as using logrank statistics for non-proportional hazards alternatives. This means that time-sequential tests based on $S_n(t)$ can achieve *both* savings in study duration and increase in power over the fixed-duration test based on $S_n(t^*)$, as shown by Gu and Lai (1991, 1998).

Lan and DeMets (1989) noted that there are two time scales in interim analysis of clinical trials with failure-time endpoints. One is "calendar time" t and the other is "information time" which is the estimate $V_n(t)$ of the variance of $S_n(t)/\sqrt{n}$ under the null hypothesis. There is no simple relationship between these two time scales and $V_n(t)$ is typically unknown before time t unless restrictive assumptions are made *a priori*. This had been a major difficulty in extending group sequential methods from immediate responses (e.g., Lan and DeMets (1983)) to failure-time outcomes. Gu and Lai (1998) recently resolved this difficulty by modifying and extending the Haybittle method described after (2.5) to time-sequential rank statistics, yielding simple but efficient time-sequential rank tests that can achieve both savings in study duration and increase in power over their nonsequential counterparts, not only when the score function used is not optimal for the actual alternatives, but also when there is noncompliance or crossover, which commonly occurs in practice. Gu and Lai (1999) also developed a simulation program to compute power and expected duration of these trials and to check the adequacy of the normal approximation to the type I error probability under various scenarios of baseline survival, censoring pattern, noncompliance, and accrual rate. The program gives the clinical trial designer four options for choosing the stopping boundary, including the boundary developed in Gu and Lai (1998), the earlier methods of Slud and Wei (1982) and Lan and DeMets (1983), and any other boundary specified by the user. The program also allows the user to choose the score function ψ in $Q_n(t, s) = \psi(H_{n,t}(s))$ from the beta family proposed by Self (1991), which includes the Harrington-Fleming (1982) class of statistics. This enables the clinical trialist to select the test statistic most sensitive to the anticipated kind of departures from the null hypothesis.

Gu and Lai (1999) also incorporated this power calculation program into another program that computes the sample size of a group sequential trial having a prescribed power at given baseline and alternative distributions.

It is widely recognized that tests of treatment effects based on the rank statistics (2.7) may lose substantial power when the effects of other covariates are strong. In nonsequential trials, a commonly used method to remedy this when logrank statistics are used is to assume the proportional hazards regression model and to use Cox's partial likelihood approach to adjust for other covariates. Tsiatis, Rosner and Tritchler (1985) and Gu and Ying (1995) have developed group sequential tests using this approach. A general asymptotic theory for time-sequential methods in proportional hazards regression models with applications to covariate adjustment is provided by Bilias, Gu and Ying (1997). Instead of relying on the proportional hazards model to adjust for concomitant variables, it is useful to have other methods for covariate adjustment, especially in situations where other score functions than the logrank are used in (2.7) to allow for the possibility of non-proportional hazards alternatives. Gu and Lai (1998) developed alternative covariate adjustment methods based on M-estimators in accelerated failure time models and established the associated asymptotic theory.

3. Sequential Change-point Detection in Quality Control and Stochastic Systems

The subject of statistical quality control is concerned with monitoring and evaluation of the quality of products from a continuous production process. Shewhart (1931) introduced the fundamental concept of a "state of statistical control", in which the behavior of some suitably chosen quality characteristic at time t has a given probability distribution. To detect significant departures from this state, he introduced a process inspection scheme that takes samples of fixed size at regular intervals of time and computes from the sample at time t a suitably chosen statistic X_t, which can be presented graphically in the form of a control chart.

Shewhart's control chart is a "single-sample" scheme whose decision depends solely on the current sample although the results of previous samples are available from the chart. To improve the sensitivity of the Shewhart charts, Page (1954) and Shiryaev (1963) modified Wald's theory of sequential hypothesis testing to develop the CUSUM and the Shiryaev-Roberts charts that have certain optimality properties. Their underlying statistical model, which will be denoted by $P^{(\nu)}$, is a sequence of independent random variables $X_1, X_2, \ldots$, where X_t denotes a statistic computed from the sample at monitoring time t, such that the X_t have a common specified distribution F_0 for $t < \nu$, representing Shewhart's "state of statistical control", and such that the X_t have another common distribution F_1

for $t \geq \nu$. We shall use P_0 to denote the alternative model of perpetual statistical control (corresponding to $\nu = \infty$). Assuming that F_0, F_1 have densities f_0 and f_1 with respect to some measure, Moustakides (1986) showed that Page's CUSUM scheme

$$N = \inf\{n : \max_{1 \leq k \leq n} \sum_{i=k}^{n} \log(f_1(X_i)/f_0(X_i)) \geq c_\gamma\} \tag{3.1}$$

is optimal in the following minimax sense: Let c_γ be so chosen that $E_0(N) = \gamma$ and let $\mathcal{F}_\gamma$ be the class of all monitoring schemes subject to the constraint $E_0(T) \geq \gamma$. Then (3.1) minimizes the worst-case expected delay $\sup_{\nu \geq 1} \operatorname{ess\,sup} E^{(\nu)}[(T - \nu + 1)^+ | X_1, \ldots, X_{\nu-1}]$ over all rules T that belong to $\mathcal{F}_\gamma$. Earlier Lorden (1971) showed that this minimax property holds asymptotically as $\gamma \to \infty$. Specifically he showed that $E_0(N) \geq \exp(c_\gamma)$ and that for $c_\gamma = \log \gamma$,

$$\begin{aligned} &\sup_{\nu \geq 1} \operatorname{ess\,sup} E^{(\nu)}[(N - \nu + 1)^+ | X_1, \ldots, X_{\nu-1}] \sim (\log \gamma)/I(f_1, f_0) \\ &\sim \inf_{T \in \mathcal{F}_\gamma} \{\sup_{\nu \geq 1} \operatorname{ess\,sup} E^{(\nu)}[(T - \nu + 1)^+ | X_1, \ldots, X_{\nu-1}]\}, \end{aligned} \tag{3.2}$$

where $I(f_1, f_0) = E_{f_1}\{\log(f_1(X_1)/f_0(X_1))\}$ denotes the Kullback-Leibler information number.

Note that the unknown change-point ν is estimated by maximum likelihood in Page's CUSUM scheme (3.1). Using a Bayesian approach, Shiryaev (1963, 1978) assumed a geometric prior distribution on ν ($P\{\nu = n\} = p(1-p)^{n-1}, n = 1, 2, \ldots$) and formulated the problem of optimal sequential change-point detection as an optimal stopping problem, with a loss of c for each observation taken after ν and a loss of 1 for a false alarm before ν. He showed that the Bayes rule is defined by the stopping time

$$N_q(\gamma) = \inf\{n \geq 1 : P(\nu \leq n | X_1, \ldots, X_n) \geq \gamma/(\gamma + p^{-1})\} = \inf\{n \geq 1 : R_{q,n} \geq \gamma\}, \tag{3.3}$$

where $q = 1 - p$ and $R_{q,n} = \sum_{k=1}^{n} \Pi_{i=k}^{n}\{q^{-1} f_1(X_i)/f_0(X_i)\}$. Note that $P(\nu \leq n | X_1, \ldots, X_n) = R_{q,n}/(R_{q,n} + p^{-1})$. Without assuming a specified prior distribution on ν, Roberts (1966) modified Shiryaev's rule to

$$N(\gamma) = \inf\{n \geq 1 : \lim_{q \to 1} R_{q,n} \geq \gamma\} = \inf\{n \geq 1 : \sum_{k=1}^{n} \prod_{i=k}^{n} (f_1(X_i)/f_0(X_i)) \geq \gamma\}, \tag{3.4}$$

which Pollak (1985) proved to be asymptotically Bayes risk efficient as $p \to 0$ and also asymptotically minimax as $\gamma \to \infty$.

As noted by Lorden (1971), Page's CUSUM scheme (3.1) corresponds to stopping when a one-sided SPRT with log-boundary based on $X_K, X_{K+1}, \ldots,$

rejects the null hypothesis $H_0 : f = f_0$, where K is the maximum likelihood estimate of ν. Thus, (3.1) can be expressed as

$$N = \min_{k\geq 1}(N_k + k - 1), \tag{3.5}$$

where N_k is the stopping time of the one-sided SPRT applied to $X_k, X_{k+1}, \ldots$. Instead of the stopping rule of the one-sided SPRT, one can use other stopping rules. Lorden (1971) showed that if $X_1, X_2, \ldots$ are i.i.d. and τ is a stopping time with respect to $X_1, X_2, \ldots$ such that $P(\tau < \infty) \leq \alpha$, then letting N_k be the stopping time obtained by applying τ to $X_k, X_{k+1}, \ldots$ and defining N by (3.5), $EN \geq 1/\alpha$ and N is a stopping time. Making use of Lorden's result with $\tau = m_\gamma$ if $\sum_{i=1}^{m_\gamma} \log(f_1(X_i)/f_0(X_i)) \geq \log\gamma$ and $\tau = \infty$ otherwise, Lai (1995) showed that the moving average scheme

$$N^* = \inf\{n : \sum_{i=n-m_\gamma+1}^{n} \log(f_1(X_i)/f_0(X_i)) \geq \log\gamma\} \tag{3.6}$$

satisfies both $E_0(N^*) \geq \gamma$ and the asymptotic minimax property (3.2) (with N replaced by N^*) if the fixed sample size m_γ of the Neyman-Pearson test in τ is so chosen that

$$m_\gamma \sim (\log\gamma)/I(f_1, f_0) \quad \text{and} \quad \{m_\gamma - (\log\gamma)/I(f_1, f_0)\}/(\log\gamma)^{\frac{1}{2}} \to \infty. \tag{3.7}$$

In fact, the moving average rule (3.6) is asymptotically as efficient as the CUSUM and the Shiryaev-Roberts rules as $\gamma \to \infty$ when the window size m_γ satisfies (3.7).

Moving averages of the type $\sum_{i=1}^{n} a_{n-i}X_i$ that put most (or all) weight on the current and immediate past observations are popular alternatives to Shewhart's $\bar{X}$-chart for detecting changes in the process mean. A commonly used performance measure of the Shewhart, CUSUM, Shiryaev-Roberts and moving average charts is the *average run length* (ARL), which is defined as $E_\theta T$ when the quality parameter remains at a fixed level θ. Roberts (1966) studied by simulation the ARL of moving average schemes of the form $N^* = \inf\{n : \sum_{i=1}^{n} a_{n-i}X_i \geq c\}$ for the special cases $a_i = p^i$ with $0 < p < 1$ (exponentially weighted moving averages, or EWMA) and $a_0 = \cdots = a_{m-1} = m^{-1}$, $a_i = 0$ for $i \geq m$ (finite moving averages). In the case of normally distributed X_i, Lai (1974) derived sharp upper and lower bounds and asymptotic approximations for $E_\theta N^*$. Böhm and Hackl (1990) extended these results to non-normal X_i. For EWMA charts, Crowder (1987) and Lucas and Saccuci (1990) developed numerical approximations to the ARL.

Topics of current interest in industrial quality control include use of variable sampling rates to achieve quicker detection by increasing the sampling rate when

the quality statistics X_t are near the control limits (cf. Assaf (1988), Reynolds, Amin and Arnold (1990)), multivariate control charts (cf. Mason, Champ, Tracy, Wierda and Young (1997)) and dependent observations (cf. Hunter (1990), Box and Ramirez (1992), Box and Luceño (1997)). Stoumbos, Reynolds, Ryan and Woodall (2000) give an overview of these topics and additional references in their discussion of the current status of statistical process control.

Sequential change-point detection involving multivariate and dependent observations is also an important topic in the engineering literature on fault detection and diagnosis, where X_t typically represents the output of a partially observable stochastic dynamical system at time t. Basseville and Nikiforov (1993) provide an overview of various detection algorithms in this literature. Although it is easy to extend the CUSUM rule (3.1) to non-independent observations by simply replacing $f_j(X_i)$ in (3.1) by the conditional density $f_j(X_i|X_1,\ldots,X_{i-1})$ for $j = 0, 1$, it has been an open problem concerning whether the asymptotic optimality property (3.2) of the CUSUM rule still holds. By using a change-of-measure argument and the strong law for log-likelihood ratio statistics, similar to that used by Lai (1981) to prove an asymptotic analogue of Hoeffding's (1960) lower bound for sequential tests based on dependent data, Lai (1998) recently proved that (3.2) holds quite generally, with $I(f_1, f_0)$ being the limit in the aforementioned strong law. Instead of the ARL constraint $E_0(T) \geq \gamma$, suppose one imposes a probability constraint of the form $\sup_{k\geq 1} P_0\{k \leq T < k + m\} \leq m/\gamma$ with $m = o(\gamma)$ but $m/\log\gamma \to \infty$. Then a similar change-of-measure argument gives an asymptotic lower bound for the detection delay of the form

$$E^{(k)}(T - k + 1)^+ \geq \{P_0(T \geq k)/I(f_1, f_0) + o(1)\}\log\gamma, \quad \text{uniformly in} \quad k \geq 1. \tag{3.8}$$

Alternatively, if one puts a prior distribution π on ν and imposes the false alarm constraint $P\{T < \nu\} = \sum_{k=1}^{\infty} \pi(k)P_0\{T < k\} \leq \gamma^{-1}$, then a similar argument can be used to derive an asymptotic lower bound for the expected detection delay $E(T-\nu+1)^+ = \sum_{k=1}^{\infty} \pi(k)E^{(k)}(T-k+1)^+$ under certain conditions on π. Moreover, CUSUM rules of the form (3.1) with $f_j(X_i)$ replaced by $f_j(X_i|X_1,\ldots,X_{i-1})$ for $j = 0, 1$ attain these asymptotic lower bounds subject to the ARL constraint $E_0(T) \geq \gamma$, or the probability constraint $\sup_k P_0(k \leq T < k + m) \leq m/\gamma$, or the Bayesian false alarm constraint $P(T < \nu) \leq \gamma^{-1}$ with a prior distribution π on ν; see Lai (1998).

Subject to the Bayesian false alarm constraint, minimization of $E(T - \nu)^+$ can be formulated as the optimal stopping problem of choosing a stopping rule T to minimize the expected loss $E[\lambda I_{\{T<\nu\}} + (T-\nu)I_{\{T\geq\nu\}}]$, where λ can be regarded as a Lagrange multiplier associated with the probability constraint and E denotes expectation with respect to the probability measure P under which the change-time ν has the prior distribution π. For the case of geometrically distributed ν

and independent X_t, Shiryaev (1978) showed that the optimal stopping rule has the form (3.3). Yakir (1994) generalized this result to finite-state Markov chains X_t. For more general prior distributions on ν or non-Markovian stochastic systems X_t, the optimal stopping problem associated with the Bayes detection rule becomes intractable. Instead of solving the optimal stopping problem directly, Lai's (1998) approach is to develop an asymptotic lower bound for the detection delay subject to the Bayesian false alarm constraint $P(T < \nu) \leq \gamma^{-1}$ and to show that the CUSUM rule with a suitably chosen threshold c_γ asymptotically attains this lower bound as $\gamma \to \infty$.

In practice, the post-change distributions are usually modeled by parametric families with unknown parameters, and the preceding theory that assumes completely specified $f_1(\cdot|x_1, \ldots, x_{i-1})$ is too restrictive. Nevertheless, the asymptotic lower bounds assuming known f_1 provide benchmarks that we try to attain even when unknown parameters are present. An obvious way to modify the CUSUM rule (3.1) for the case of $f_\theta(\cdot|x_1, \ldots, x_{i-1})$ with unknown post-change parameter θ is to estimate it by maximum likelihood, leading to the generalized likelihood ratio (GLR) rule

$$N_G = \inf \left\{ n : \max_{1 \leq k \leq n} \sup_{\theta \in \Theta} \sum_{i=k}^{n} \log \frac{f_\theta(X_i|X_1, \ldots, X_{i-1})}{f_{\theta_0}(X_i|X_1, \ldots, X_{i-1})} \geq c_\gamma \right\}. \tag{3.9}$$

For the problem of detecting shifts in the mean θ of independent normal observations with known variance, this idea was proposed by Barnard (1959), but the statistical properties of the procedure remained a long-standing problem that was recently solved by Siegmund and Venkatraman (1995), whose asymptotic approximations to the ARL of the GLR rule under θ_0 and under $\theta \neq \theta_0$ show that the GLR rule is asymptotically optimal in the sense of (3.2).

For practical implementation, the CUSUM rule (3.1) can be written in the recursive form $N = \inf\{n : \ell_n \geq c_\gamma\}$, where $\ell_n = \{\ell_{n-1} + \log(f_1(X_n)/f_0(X_n))\}^+$ with $\ell_0 = 0$. The GLR rule (3.9) does not have such convenient recursive forms and the memory requirements and number of computations at time n grow to infinity with n. A natural modification to get around this difficulty is to replace $\max_{1 \leq k \leq n}$ in (3.9) by $\max_{n-M \leq k \leq n}$. Such window-limited GLR rules were first introduced by Willsky and Jones (1976) in the context of detecting additive changes in linear state-space models. Consider the stochastic system described by the state-space representation of the observed signals y_t:

$$x_{t+1} = F_t x_t + G_t u_t + w_t \tag{3.10a}$$

$$y_t = H_t x_t + J_t u_t + \epsilon_t \tag{3.10b}$$

in which the unobservable state vector x_t, the observable input vector u_t and the measurement vector y_t have dimensions p, q and r, respectively, and w_t, ϵ_t are

independent Gaussian vectors with zero means and $\mathrm{Cov}(w_t) = Q_t, \mathrm{Cov}(\epsilon_t) = R_t$. The Kalman filter provides a recursive algorithm to compute the conditional expectation $\hat{x}_{t|t-1}$ of the state x_t given the past observations $y_{t-1}, u_{t-1}, y_{t-2}, u_{t-2}, \ldots$ The innovations $e_t := y_t - H_t\hat{x}_{t|t-1} - J_t u_t$ are independent zero-mean Gaussian vectors with $\mathrm{Cov}(e_t) = V_t$ given recursively by the Riccati equations. If at an unknown time τ the system undergoes some additive change in the sense that $u_t' I_{\{t\geq\tau\}}$ and/or $u_t'' I_{\{t\geq\tau\}}$ are added to the right hand side of (3.10a) and/or (3.10b), then the innovations e_t are still independent Gaussian vectors with covariance matrices V_t, but their means $m_t = E(e_t)$ for $t \geq \tau$ are of the form $m_t = \rho(t,\tau)\theta$ instead of the baseline values $m_t = 0$ for $t < \tau$, where $\rho(t,k)$ is a matrix that can be computed recursively and θ is an unknown parameter vector; see Willsky and Jones (1976) who proposed the window-limited GLR detector of the form

$$\begin{aligned} N_W &= \inf\Big\{n : \max_{n-M\leq k\leq n-\tilde{M}} \sup_\theta \sum_{i=k}^n \log[f(V_i^{-1/2}(e_i - \rho(i,k)\theta))/f(V_i^{-1/2}e_i)] \geq c_\gamma\Big\} \\ &= \inf\Big\{n : \max_{n-M\leq k\leq n-\tilde{M}} \Big(\sum_{i=k}^n \rho^T(i,k)V_i^{-1}e_i\Big)^T \Big(\sum_{i=k}^n \rho^T(i,k)V_i^{-1}\rho(i,k)\Big)^{-1} \\ &\quad \times \Big(\sum_{i=k}^n \rho^T(i,k)V_i^{-1}e_i\Big)/2 \geq c_\gamma\Big\}, \end{aligned} \tag{3.11}$$

where f denotes the standard r-dimensional normal density function.

Although window-limited GLR rules of the type (3.11) have found widespread applications in fault detection of navigation and other control systems and in signal processing and tracking of maneuvering targets (cf. Basseville and Nikiforov (1993)), how to choose $M, \tilde{M}$ and c_γ appropriately has remained a difficult open problem that was recently addressed by Lai (1995) and Lai and Shan (1999). Lai (1995) began by considering the simpler situation of detecting changes in the mean θ of independent normal observations $X_1, X_2, \ldots$ from a known baseline value $\theta = 0$. Here the window-limited GLR rule has the form

$$N_W = \inf\{n : \max_{n-M\leq k\leq n} (X_k + \cdots + X_n)^2/2(n-k+1) \geq c_\gamma\}, \tag{3.12}$$

and the methods of Siegmund and Venkatraman (1995) to analyze the GLR rule (3.9) in this independent normal case can be extended to (3.12). In particular, if we choose $M \sim \gamma$, then we have $E_0 N_W \sim E_0 N_G \sim K c_\gamma^{-1/2} e^{c_\gamma}$ as $c_\gamma \to \infty$, where an explicit formula for K is given in Siegmund and Venkatraman (1995). Therefore, choosing $c_\gamma = \log\gamma + \frac{1}{2}\log\log\gamma - \log K + o(1)$ gives $E_0 N_W \sim E_0 N_G \sim \gamma$. With this choice of c_γ, we also have $E_\theta N_W \sim E_\theta N_G \sim \min\{\gamma, (2\log\gamma)/\theta^2\}$ uniformly in $0 < |\theta| \leq (\log\gamma)^{\frac{1}{2}-\epsilon}$ for every $\epsilon > 0$. The choice $M = \gamma$ for the

window size in (3.12) requires computation of $\gamma+1$ quantities $(X_{n-i}+\cdots+X_n)^2/(i+1), i=0,\ldots,\gamma$, at every stage $n>\gamma$, and it is desirable to reduce the computational burden for large γ by using a smaller window size. To develop efficient detection schemes that involve $O(\log\gamma)$ computations at every stage n, Lai (1995) used an idea similar to that in the theory of group sequential tests, which is to replace $\max_{0\le n-k\le M}$ in (3.12) by $\max_{n-k+1\in\mathcal{N}}$ where $\mathcal{N}=\{1,\ldots,M\}\cup\{[b^jM]:1\le j\le J\}$, with $M\sim a\log\gamma, b>1$ and $J=\min\{j:[b^jM]\ge\gamma\}\sim(\log\gamma)/(\log b)$. Specifically, replacing N_W by

$$\tilde{N}_W=\inf\{n:\max_{k:n-k+1\in\mathcal{N}}(X_k+\cdots+X_n)^2/2(n-k+1)\ge c_\gamma\},\tag{3.13}$$

Lai (1995) showed that $E_0(\tilde{N}_W)\sim\tilde{K}c_\gamma^{-1/2}e^{c_\gamma}\sim\gamma$ if $c_\gamma=\log\gamma+\frac{1}{2}\log\log\gamma-\log\tilde{K}+o(1)$, and that $E_\theta(\tilde{N}_W)\sim(2\log\gamma)/\theta^2$ if $|\theta|>\sqrt{2/a}$ while $E_\theta(\tilde{N}_W)\le(1+o(1))\min\{\gamma,(2b\log\gamma)/\theta^2\}$ uniformly in $0<|\theta|\le\sqrt{2/a}$. Hence, choosing b close to 1 (say $b=1.1$), there is little loss of efficiency in reducing the computational complexity of N_G by its window-limited modification (3.13).

This idea was subsequently extended by Lai and Shan (1999) to address the long-standing problem concerning how the window size and the threshold should be chosen in the Willsky-Jones rule (3.11). As pointed out by Basseville and Benveniste (1983), the main difficulty of this problem lies in the coupling effect between the threshold and window size on the performance of the rule. The basic idea of Lai and Shan (1999) is to decouple the effects of the threshold and window size. A threshold of order $\log\gamma$ is needed to ensure a false alarm duration of γ, as in the simpler problem of detecting changes in a normal mean. With the threshold thus chosen to control the false alarm rate, the choice of the windows is targeted towards making the rule as efficient as possible for detecting the unknown change. Putting a complexity constraint of $O(\log\gamma)$ on the number of elements of the window, the Willsky-Jones window in (3.11) is enlarged to the form $\tilde{\mathcal{N}}=\{\tilde{M},\ldots,M\}\cup\{[b^jM]:1\le j\le J\}$ as in (3.11). Here we need a minimal delay $\tilde{M}\ge\dim(\theta)$ to avoid difficulties with GLR statistics when $n-k<\dim(\theta)$. Under certain stability assumptions on the Kalman filter, such window-limited GLR rules with $\max_{n-M\le k\le n-\tilde{M}}$ in (3.11) replaced by $\max_{k:n-k+1\in\tilde{\mathcal{N}}}$ can be shown to be asymptotically optimal for detecting changes with $I(\theta,0)>a^{-1}$, and to be within b times the asymptotic lower bound for expected delay in detecting smaller changes. Moreover, for these window-limited GLR rules T, $\sup_{k\ge1}P_0(k\le T<k+m)\sim P_0(T\le m)\sim m/E_0(T)$, as $E_0(T)\sim\gamma\to\infty$ and $m/\log\gamma\to\infty$ but $\log m=o(\log\gamma)$. Hence to determine the threshold c_γ, the ARL constraint $E_0(T)\doteq\gamma$ can be replaced by the probability constraint $P_0(T\le m)\doteq m/\gamma$, which is much more tractable since simulating $P_0(T\le m)$ involves far fewer random variables (no more than m in each simulation run)

than directly simulating $E_0(T)$. Importance sampling methods have also been developed by Lai and Shan (1999) for the Monte Carlo evaluation of $P_0(T \le m)$. Making use of saddlepoint approximations for Markov additive processes, Chan and Lai (2000a) have developed asymptotic formulas for $P_0(T \le m)$ when T is a window-limited rule applied to the GLR or other functions of scan statistics based on Markov-dependent observations. Siegmund and Yakir (2000) have used another approach to derive asymptotic approximations for GLR rules.

Another topic in the engineering literature on sequential detection of parameter changes of signals and systems is concerned with detection rules based on non-likelihood-based detection statistics, which are often used in applications. By using weak convergence theory, the CUSUM rule has been extended to non-likelihood-based detection statistics by Benveniste, Basseville and Moustakides (1987), whose "asymptotic local approach" can be summarized as follows. Suppose the detection statistics Y_i are such that for every fixed μ,

$$\{\gamma^{-1/2}\sum_{i=1}^{[\gamma t]} Y_i, t \ge 0\} \quad \text{converges weakly under} \quad P_{\mu/\sqrt{\gamma}} \quad \text{to} \quad \{W_\mu(t), t \ge 0\} \tag{3.14}$$

as $\gamma \to \infty$, where $\{W_\mu(t), t \ge 0\}$ is a multivariate Gaussian process with independent increments such that $EW_\mu(t) = \mu t$ and $\mathrm{Cov}(W_\mu(t)) = tV$. The baseline probability measure is P_0 that corresponds to $\mu = 0$. Let $S_{n,k} = \sum_{i=k}^n Y_i$ and $\lambda = \lambda_1/\sqrt{\gamma}$. Consider the CUSUM rule

$$T_\gamma = \inf\{n : \max_{1\le k\le n} [\lambda^T S_{n,k} - (n-k+1)\lambda^T V\lambda/2] \ge c\}. \tag{3.15}$$

Then as $\gamma \to \infty$, the weak convergence property (3.14) implies that for fixed $c, T_\gamma/\gamma$ converges in distribution under $P_{\mu/\sqrt{\gamma}}$ to $\tau_\mu(c) = \inf\{t : \max_{0\le u\le t} [\lambda_1^T(W_\mu(t) - W_\mu(u)) - (t-u)\lambda_1^T V\lambda_1/2] \ge c\}$, and therefore $E_{\mu/\sqrt{\gamma}}(T_\gamma) \sim \gamma E\tau_\mu(c)$. How should (3.15) be modified when λ is not specified in advance? Maximizing the CUSUM statistics $\lambda^T S_{n,k} - (n-k+1)\lambda^T V\lambda/2$ over λ yields $S_{n,k}^T V^{-1} S_{n,k}/2(n-k+1)$, which leads to the following window-limited rule proposed by Zhang, Basseville and Benveniste (1994):

$$\tilde{T}_\gamma = \inf\{n > b_1\gamma : \max_{n-b_2\gamma\le k\le n-b_1\gamma} S_{n,k}^T V^{-1} S_{n,k}/2(n-k+1) \ge c\}, \tag{3.16}$$

with $b_2 > b_1 > 0$. For fixed c and μ, (3.14) implies that $\tilde{T}_\gamma/\gamma$ converges in distribution under $P_{\mu/\sqrt{\gamma}}$ to $\tilde{\tau}_\mu = \inf\{t > b_1 : \max_{t-b_2\le u\le t-b_1}(W_\mu(t) - W_\mu(u))^T V^{-1}(W_\mu(t) - W_\mu(u))/2(t-u) \ge c\}$. The preceding asymptotic approach has been called "local" because it is based on weak convergence of T_γ/γ or $\tilde{T}_\gamma/\gamma$ under P_0 to the same limiting distribution as that in the canonical setting of

independent standard normal $V^{-1/2}Y_t$, when λ is of the order $\lambda_1/\sqrt{\gamma}$ in (3.15) or when the window is of the form $b_1\gamma \le n-k \le b_2\gamma$ in (3.16). Such choice of window size (or λ) makes $\tilde{T}_\gamma$ (or T_γ) very inefficient for detecting changes that are considerably larger than the $O(\gamma^{-1/2})$ order of magnitude for the "local" changes.

Lai and Shan (1999) and Chan and Lai (2000b) have used another approach based on moderate deviations theory instead of weak convergence approximations to extend the window-limited GLR rule (3.11) to non-likelihood-based detection statistics. The detection rule has the general form

$$N_W^* = \inf\{n > \tilde{M} : \max_{k:n-k+1\in N_{\tilde{M}}} \Big(\sum_{i=k}^n Y_i\Big)^T V_{n,k}^{-1}\Big(\sum_{i=k}^n Y_i\Big)/2 \ge c\}, \tag{3.17}$$

where $c \sim \log\gamma$ (instead of bounded c in the asymptotic local approach), $\mathcal{N}_{\tilde{M}} = \{\tilde{M}j : j \in \mathcal{N}\}$, $\mathcal{N} = \{1,\dots,M\} \cup \{[b^j M] : 1 \le j \le J\}, b > 1, M \sim ac, J = \min\{j : [b^j M\tilde{M}] \ge \gamma\}$, $\tilde{M}/c \to \infty$ (to ensure that the sums $\sum_{i=k}^n Y_i$ become approximately Gaussian in the moderate deviations sense), and $V_{n,k}$ is an estimate of the asymptotic covariance matrix of $S_{n,k}$ under P_0. Note that $V^{-1}/(n-k+1)$ in (3.16) corresponds to the inverse of $V_{n,k} = (n-k+1)V$. In many applications V is unknown and needs to be estimated. Even when V is known, using $V_{n,k}$ instead of $(n-k+1)V$ offers the flexibility of making adjustments for non-normality in treating (3.17) under P_0 as if it were a window-limited GLR rule with independent standard normal $V^{-1/2}Y_t$, where the Y_t are actually dependent random vectors with unknown (and possibly non-normal) distributions. Note that grouping the observations into batches as in (3.17) arises naturally in conventional quality control applications, in which samples of size s are taken at regular intervals of time, so $\tilde{M}$ is either s or some multiple thereof.

Another direction of research in the engineering literature is on-line fault isolation (or change diagnosis), whose objective is to determine, upon detection of change in a system, which one in a set of J possible changes has actually occurred (cf. Basseville and Nikiforov (1993), Nikiforov (1995), Patton, Frank and Clark (1989)). By making use of sequential testing theory for $J+1$ hypotheses together with the basic ideas in sequential change-point detection described above, Lai (2000) recently developed window-limited detection-isolation rules that are not too demanding in computational and memory requirements and yet are nearly optimal under several performance criteria.

4. Point and Interval Estimation in Sequential Experiments

In this section we first consider the topic of bounded-risk and asymptotically risk-efficient point estimates and fixed-width confidence intervals that are

possible only in sequential designs. We then consider the difficult but important problem of confidence intervals following sequential tests, especially those in group sequential trials where this long-standing problem is of particular practical importance. The third topic considered is recursive estimation in signal processing and adaptive control in the engineering literature.

4.1. Bounded-risk estimators, fixed-width confidence intervals and asymptotically risk-efficient sequential estimation

In the same year that Wald's pioneering paper on sequential testing appeared, Haldane (1945) published a seminal paper on bounded-risk sequential estimation. He was motivated by a biological application involving estimating the small probability p of some attribute. Noting that the usual sample proportion based on n observations has variance $p(1-p)/n$, he introduced a sequential estimate that stops sampling at the first time N when $m(\geq 2)$ successes occur and used the estimate $\hat{p}_m = (m-1)/(N-1)$. He showed that $\hat{p}_m$ is unbiased and has variance

$$\frac{p^2(1-p)}{m}\Big\{1 + \frac{2(1-p)}{m+1} + \frac{3!(1-p)^2}{(m+1)(m+2)} + O(m^{-3})\Big\}. \tag{4.1}$$

Thus, $\hat{p}_m$ has bounded quadratic risk $E\{(\hat{p}_m - p)/p\}^2 = O(m^{-1})$ uniformly in $0 < p \leq \frac{1}{2}$.

About thirty years later, Borisov and Konev (1977) gave another application of sequential designs to yield a bounded-risk unbiased estimator in the AR(1) model $y_n = \beta y_{n-1} + \epsilon_n$ with i.i.d. errors ϵ_n such that $E\epsilon_n = 0$ and $E\epsilon_n^2 = \sigma^2$. The y_n have a limiting distribution when $|\beta| < 1$, behave like a random walk when $|\beta| = 1$, and exhibit exponential growth when $|\beta| > 1$. In spite of this, they showed that it is possible to construct a sequential estimate $\hat{\beta}_c$ that is unbiased and whose variance is uniformly bounded by σ^2/c over all $\beta \in \mathbf{R}$. Konev and Lai (1995) generalized the construction to the stochastic regression model $y_n = \beta x_n + \epsilon_n$, in which x_n is measurable with respect to the σ-field $\mathcal{F}_{n-1}$ generated by $\{\epsilon_1, \ldots, \epsilon_{n-1}\}$ and $\{\epsilon_n, \mathcal{F}_n, n \geq 1\}$ is a martingale difference sequence such that

$$\sup_n E(\epsilon_n^2 | \mathcal{F}_{n-1}) \leq \sigma^2 \quad \text{a.s. and} \quad \sum_{n=1}^{\infty} x_n^2 = \infty \text{ a.s.} \tag{4.2}$$

Let $N_c = \inf\{n \geq 1 : \sum_{i=1}^n x_i^2 \geq c\}$. Then $N_c < \infty$ a.s. by (4.2) and $\sum_{i \leq N_c} x_i^2 \geq c > \sum_{i<N_c} x_i^2$, so we can define $\theta_c \in (0,1]$ uniquely by the equation

$$\sum_{i<N_c} x_i^2 + \theta_c x_{N_c}^2 = c. \tag{4.3}$$

Let $\hat{\beta}_c$ be the modified least squares estimator

$$\hat{\beta}_c = \Big(\sum_{i<N_c} x_i y_i + \theta_c x_{N_c} y_{N_c} \Big)/c. \tag{4.4}$$

Konev and Lai (1995) showed that $\hat{\beta}_c$ is unbiased and $\mathrm{Var}(\hat{\beta}_c) \le \sigma^2/c$. They also extended this to construct bounded-risk estimators in the multiple stochastic regression model $y_n = \beta^T \mathbf{x}_n + \epsilon_n$, in which $\mathbf{x}_n$ is a $p \times 1$ $\mathcal{F}_{n-1}$-measurable random vector.

It is well known that the usual least squares estimator $\tilde{\beta}_n$ of the autoregressive parameter based on a sample of fixed size n from an AR(1) model is asymptotically normal if $|\beta| < 1$, but its limiting distribution after studentization is highly non-Gaussian if $|\beta| = 1$. This causes considerable difficulty in constructing large-sample confidence intervals in non-explosive but possibly nonstationary AR(1) models. Lai and Siegmund (1983) circumvented this difficulty by using the stopping rule $N_c = \inf\{n \ge 1 : \sum_{i=1}^n y_{i-1}^2 \ge c\}$. They proved the following uniform asymptotic normality property of $\tilde{\beta}_{N_c}$ as $c \to \infty$:

$$\sup_{|\beta| \le 1,\ x \in \mathbf{R}} |P\{\sqrt{c}\ (\tilde{\beta}_{N_c} - \beta)/\sigma \le x\} - \Phi(x)| \to 0, \tag{4.5}$$

in which Φ denotes the standard normal distribution. Since $\hat{\sigma}_n^2 := n^{-1} \sum_{i=1}^n (y_i - \tilde{\beta}_n y_{i-1})^2$ is a consistent estimate of σ^2, $\tilde{\beta}_{N_c} \pm c^{-1/2} \hat{\sigma}_{N_c} \Phi^{-1}(1-\alpha)$ is an approximate $(1-2\alpha)$-level confidence interval for β in non-explosive AR(1) models. Moreover, letting

$$N(d) = \inf \Big\{ n \ge 1 : (\hat{\sigma}_n \vee n^{-1/2}) \Phi^{-1}(1-\alpha) \Big(\sum_{i=1}^n y_{i-1}^2 \Big)^{-1/2} \le d \Big\}, \tag{4.6}$$

$\tilde{\beta}_{N(d)} \pm d$ is an approximate $(1-2\alpha)$-level confidence interval for β, with fixed width $2d \to 0$.

Fixed-width confidence intervals based on stopping rules of the type (4.6) were first proposed by Chow and Robbins (1965) for a population mean when the variance σ^2 is unknown. Letting $\bar{Y}_n$ denote the sample mean and $\hat{\sigma}_n$ the sample variance based on a sample of size n, define

$$\tilde{N}(d) = \inf\{n \ge n_0 : (\hat{\sigma}_n \vee n^{-1/2}) \Phi^{-1}(1-\alpha) \le d\sqrt{n}\}. \tag{4.7}$$

The Chow-Robbins approximate $(1-2\alpha)$-level confidence interval for the population mean is $\bar{Y}_{\tilde{N}(d)} \pm d$. When the Y_i are normal, Stein (1945, 1949) introduced a two-stage procedure, giving an exact confidence interval with width $2d$. Its first stage uses a sample of fixed size m to estimate σ and the sample size for the second stage is based on the estimate $\hat{\sigma}_m$. Let $n(\sigma, d)$ be the fixed sample size

such that $\bar{Y}_{n(\sigma,d)} \pm d$ has coverage probability $1-2\alpha$ when σ is known. If $m \to \infty$ but $m = o(n(\sigma, d))$ as $d \to 0$, then the expected sample size of Stein's two-stage procedure is of a larger order of magnitude than $n(\sigma, d)$ (cf. Cox (1952)). On the other hand, for the Chow-Robbins procedure, Woodroofe (1977) showed that for $n_0 \geq 4, E(\tilde{N}(d)) - n(\sigma, d)$ approaches a finite limit as $d \to 0$. Moreover, for $n_0 \geq 7$, the coverage probability of $\bar{Y}_{\tilde{N}(d)} \pm d$ is $1-2\alpha+O(d^2)$. Hall (1981) showed that a suitably chosen three-stage procedure can give a fixed-width interval with similar second-order properties in comparison with the fully sequential procedure. For the parametric problem of fixed-width interval estimation of the mean of a normal distribution with unknown variance, Woodroofe (1986b) developed an asymptotic lower bound for the expected sample size of any sequential fixed-width procedure that has the prescribed coverage probability as the width $2d$ approaches 0, and showed that this bound is attained by a fully sequential procedure. For nonparametric problems, Jurečková (1991) gives a review of extensions of the Chow-Robbins approach to the multivariate setting and to U-statistics and R- and M-estimators.

Closely related to the Chow-Robbins theory of fixed-width confidence intervals is the theory of asymptotically risk-efficient sequential estimation, introduced by Robbins (1959) for estimating a normal mean μ with cost c per observation. If the variance σ^2 is known, the optimal fixed sample size that minimizes $R_n := E(\bar{X}_n - \mu)^2 + cn = \sigma^2/n + cn$ is n_c, which is the smallest positive integer $\geq \sigma/\sqrt{c}$, and $R_{n_c} = 2\sigma\sqrt{c} + O(c)$ as $c \to 0$. Without assuming σ to be known, Robbins (1959) proposed to replace n_c by the stopping rule

$$T_c = \inf\{n \geq n_0 : n^2 \geq \hat{\sigma}_n^2/c\}. \tag{4.8}$$

Woodroofe (1977) showed that $(R_{T_c} - R_{n_c})/c$ converges to a finite limit as $c \to 0$ if $n_0 \geq 4$, where the risk of a stopping rule T is defined by $R_T = E\{(\bar{Y}_T - \mu)^2 + cT\}$. Without assuming normality, Chow and Martinsek (1982) showed under certain assumptions on n_0 and the moments of Y_1 the "bounded regret" property $(R_{T_c} - R_{n_c})/c = O(1)$, by using certain asymptotic expansions for the moments of the stopping time and of the randomly stopped sample mean. A comprehensive theory of these asymptotic expansions has been developed by Aras and Woodroofe (1993). Extending this theory to U-statistics, de la Peña and Lai (1997) have established the bounded regret property for sequential estimators based on U-statistics.

Closely linked to the theory of asymptotically risk-efficient estimation is the asymptotic theory of sequential Bayes and empirical Bayes estimators. For a given prior, the Bayes rule involves an optimal stopping problem to determine the stopping rule, whose explicit solution is difficult in most situations. In the case of

a one-parameter exponential family, Bickel and Yahav (1968) derived simple stopping rules whose Bayes risks are asymptotically equivalent to that of the Bayes rule as the cost c per observation approaches 0. Woodroofe (1981) subsequently showed that such rules are asymptotically non-deficient in the sense that their Bayes risks are within $o(c)$ of that of the Bayes rule. Martinsek (1987) extended Woodroofe's result to the problem of sequential empirical Bayes estimation of the mean of a normal or exponential distribution where certain hyperparameters of the conjugate prior are unknown and have to be estimated from auxiliary data. Besides the Bayes risk, similar asymptotic techniques can also be applied to analyze the frequentist risks of sequential Bayes and empirical Bayes estimators. In particular, for the problem of estimating the mean vector of a p-variate normal distribution with covariance matrix $\sigma^2 V$, in which V is known but σ is unknown, Ghosh, Nickerson and Sen (1987) developed asymptotic expansions of the frequentist risks of certain empirical Bayes estimators (the James-Stein shrinkage estimators) when the stopping rule T_c is of the form (4.8) modified for the present multivariate problem. These asymptotic expansions show that the shrinkage estimators outperform the sample mean $\bar{X}_{T_c}$ for $p \geq 3$. Ghosh, Mukhopadhyay and Sen (1997) give an introduction to the developments in this problem and other topics in sequential estimation.

4.2. Confidence intervals following sequential tests

Analysis of the data at the conclusion of a clinical trial typically involves not only testing of the null hypothesis but also providing estimates of the parameters associated with the primary and secondary endpoints. The use of a stopping rule whose distribution depends on these parameters introduces substantial difficulties in constructing valid confidence intervals. Siegmund (1978) developed a method, based on ordering the sample space in a certain way, to construct confidence intervals for the mean of a normal population with known variance following a repeated significance test. Tsiatis, Rosner and Mehta (1984) extended Siegmund's method to the group sequential tests of Pocock (1977) and O'Brien and Fleming (1979). Alternative orderings of the sample space were subsequently introduced by Chang and O'Brien (1986), Rosner and Tsiatis (1988), Chang (1989) and Emerson and Fleming (1990).

For samples of fixed size, an important methodology for constructing confidence intervals without distributional assumptions is Efron's (1981, 1987) bootstrap, which can be extended as follows to the case where the sample size is determined from the data by a stopping rule T. Let $X_1, X_2, \ldots$ be i.i.d. random variables with a common unknown distribution function F. Given a randomly stopped sample $X_1, \ldots, X_T$, let F_T denote the empirical distribution function

that puts probability mass $1/T$ at each of the sample values X_i. Let $X_1^*, X_2^*, \ldots$ be i.i.d. random variables with common distribution function F_T, and let T^* denote the corresponding stopping time for the sequence $\{X_i^*\}$. The "bootstrap sample" $\{X_1^*, \ldots, X_{T^*}^*\}$ can be used to tackle a wide range of problems in sequential analysis, and it is natural to ask whether known methods and results in the bootstrap literature dealing with fixed sample sizes generalize to the sequential setting. To study this problem, Chuang and Lai (1998) began by considering the simple example where the X_i are i.i.d. with unknown mean μ and known variance 1. If the X_i are known to be normal, then Rosner and Tsiatis (1988) developed the following method to construct from $(T, \bar{X}_T)$ an exact $(1-2\alpha)$-level confidence interval for μ, when T is the stopping rule of a group sequential test. For each value of μ, one can find by the recursive numerical integration algorithm of Armitage, McPherson and Rowe (1969) the quantiles $u_\alpha(\mu)$ and $u_{1-\alpha}(\mu)$ that satisfy

$$P_\mu\{(S_T - \mu T)/\sqrt{T} < u_\alpha(\mu)\} = \alpha = P_\mu\{(S_T - \mu T)/\sqrt{T} > u_{1-\alpha}(\mu)\}. \tag{4.9}$$

Hence the confidence region $\{\mu : u_\alpha(\mu) \le (S_T - \mu T)/\sqrt{T} \le u_{1-\alpha}(\mu)\}$ has coverage probability $1-2\alpha$. Letting $\bar{X}_T = S_T/T$, this confidence region reduces to an interval whose end-points are found by intersecting the line $\sqrt{T}(\bar{X}_T - \mu)$ with the curves $u_\alpha(\mu)$ and $u_{1-\alpha}(\mu)$ if there is only one intersection with each curve, which is the case commonly encountered in practice. In particular, if T is nonrandom (corresponding to a fixed sample size), then $(S_T - \mu T)/\sqrt{T}$ is standard normal with α-quantile z_α, and therefore (4.9) implies that $u_\alpha(\mu)$ and $u_{1-\alpha}(\mu)$ are horizontal lines, yielding the classical confidence interval $\bar{X}_T \pm z_\alpha/\sqrt{T}$. One way of relaxing the assumption of normally distributed X_i with mean μ and variance 1 is to assume that $X_i - \mu$ has some unknown distribution G that has mean 0 and variance 1. After stopping we can estimate G by the empirical distribution $\widehat{G}_T$ of $(X_i - \bar{X}_T)/\widehat{\sigma}_T$, $1 \le i \le T$, where $\widehat{\sigma}_T^2 = T^{-1}\sum_{i=1}^T (X_i - \bar{X}_T)^2$. Let $\epsilon_1, \epsilon_2, \ldots$ be i.i.d. with distribution $\widehat{G}_T$ and let $X_i' = \mu + \epsilon_i$. Let T' be the stopping rule T applied to $X_1', X_2', \ldots$ (instead of to $X_1, X_2, \ldots$). In analogy with (4.9), define the quantiles $\widehat{u}_\alpha(\mu)$ and $\widehat{u}_{1-\alpha}(\mu)$ of the distribution of $(\sum_{i=1}^{T'} \epsilon_i)/\sqrt{T'}$ given $\widehat{G}_T$. An approximate $1-2\alpha$ confidence set is

$$\{\mu : \widehat{u}_\alpha(\mu) < \sqrt{T}(\bar{X}_T - \mu) < \widehat{u}_{1-\alpha}(\mu)\}. \tag{4.10}$$

For every fixed μ, the quantiles $\widehat{u}_\alpha(\mu)$ and $\widehat{u}_{1-\alpha}(\mu)$ in (4.10) can be computed by simulation. Chuang and Lai (1998) developed an algorithm to compute (4.10) and used Edgeworth expansions to establish its second-order accuracy. They carried out simulation studies and found (4.10) to compare favorably with the

exact confidence interval when the X_i are normal. They also extended the method to more complex situations involving nuisance parameters.

Chuang and Lai (2000a) subsequently developed a relatively complete theory of the hybrid resampling approach. As they point out, there are three issues that must be addressed for the practical implementation of hybrid resampling methods. First, one must choose a root $R(\mathbf{X}, \theta)$, where $\mathbf{X}$ denotes the vector of observations and θ is the unknown parameter of interest; e.g., $R(\mathbf{X}, \mu) = \sqrt{T}(\bar{X}_T - \mu)$ in (4.10). Second, one needs to find a suitable resampling family $\{\widehat{F}_\theta, \theta \in \Theta\}$, where Θ denotes the set of all possible values of θ. For example, in (4.10) we take $\widehat{F}_\mu$ as the distribution of $\mu + \epsilon_i$ with the ϵ_i generated from $\widehat{G}_T$. Finally, an "implicit" hybrid region of the form $\{\theta : \widehat{u}_\alpha(\theta) < R(\mathbf{X}, \theta) < \widehat{u}_{1-\alpha}(\theta)\}$ (such as (4.10) above) has to be inverted into an "explicit" confidence interval in practice. These issues are addressed in Chuang and Lai (2000a), who also provide a synthesis and refinement of a variety of basic results from the bootstrap literature (since the bootstrap confidence interval is a special case of the hybrid confidence set with $\widehat{F}_\theta \equiv \widehat{F}$, the empirical distribution function).

Another approach to the construction of confidence intervals following sequential tests in exponential families has been developed by Woodroofe (1986a, 1992), Woodroofe and Coad (1997) and Weng and Woodroofe (2000). This approach makes use of "very weak" asymptotic expansions that provide confidence intervals I whose integrated coverage errors $\int P_\theta(\theta \notin I)\xi(\theta)d\theta$ differ from the nominal value 2α by $o(a^{-1})$ for a large class of smooth probability densities ξ, where $a(\to \infty)$ denotes the boundary of the stopping rule T_a. This average coverage accuracy differs from the usual sense of second-order accuracy used in the bootstrap literature where Edgeworth expansions are applied to the coverage errors $P_\theta(\theta \notin I)$ for all parameter values θ (cf. Hall (1992), Efron and Tibshirani (1993)). On the other hand, the very weak expansions are computationally much more appealing than the Edgeworth-type expansions (which involve difficult fluctuation-theoretic quantities) for randomly stopped sample sums in Woodroofe and Keener (1987), Woodroofe (1988) and Lai and Wang (1994). However, as pointed out by Chuang and Lai (2000a, b), these Edgeworth-type expansions can be implemented indirectly by simulation via the bootstrap or hybrid resampling, which has the additional advantage of not requiring parametric assumptions on the underlying distribution. Resampling methods are particularly attractive in clinical trials with multiple endpoints, where the stopping rule may be based on a primary outcome variable of interest and the estimation may be related to a secondary variable. It is often difficult to come up with tractable and realistic statistical models between the primary and secondary variables, although some progress has been made in this direction; see e.g., Whitehead (1986,

1997) and Yakir (1997). Instead of explicit modeling, the hybrid resampling approach uses nonparametric estimates of the joint distribution of the primary and secondary variables to adjust for early stopping in the construction of confidence intervals.

4.3. Recursive estimation in signal processing and adaptive control

The sequential estimation problems in Section 4.1 are actually sequential design problems in which a conventional estimate, based on a sample whose size is determined by a suitably chosen stopping rule, achieves certain properties such as bounded risk or asymptotic risk efficiency that are unattainable by samples of fixed size. This is in sharp contrast with sequential (or on-line) estimation in engineering, where methods for fast updating of parameter or state estimates as new data arrive in real time are known as "recursive identification" in the control systems literature, and as "adaptive algorithms" in signal processing. Prototypical of these recursive procedures is the well-known Kalman filter, which expresses in a recursive form the conditional expectation of the state x_t given current and past observations $y_t, u_t, y_{t-1}, u_{t-1}, \ldots$ for the filtering problem (or given the past observations $y_{t-1}, u_{t-1}, \ldots$ for the prediction problem) in the linear state-space model (3.10). Applying the Kalman filter to the matrix parameter θ (as the unobservable state) in the stochastic regression model

$$Y_n = \theta X_n + w_n, \tag{4.11}$$

the least squares estimate $\hat{\theta}_n = (\sum_{i=1}^n Y_i X_i^T)(\sum_{i=1}^n X_i X_i^T)^{-1}$ can be represented recursively as

$$\hat{\theta}_n = \hat{\theta}_{n-1} + (Y_n - \hat{\theta}_{n-1} X_n) X_n^T P_n, \tag{4.12a}$$

$$P_n = P_{n-1} - P_{n-1} X_n X_n^T P_{n-1} / (1 + X_n^T P_{n-1} X_n). \tag{4.12b}$$

The Y_n in the multivariate regression model (4.11) are $k \times 1$ vectors, θ is a $k \times h$ matrix of unknown parameters, X_n is an $\mathcal{F}_{n-1}$-measurable $h \times 1$ vector of regressors and the unobservable disturbances w_n form a martingale difference sequence with respect to an increasing sequence of σ-fields $\mathcal{F}_n$. Note that (4.12a) updates the previous estimate $\hat{\theta}_{n-1}$ by using the prediction error $Y_n - \hat{\theta}_{n-1} X_n$, while (4.12b) computes $P_n = (\sum_{i=1}^n X_i X_i^T)^{-1}$ recursively without having to invert matrices, an obvious advantage when h is large or when $\sum_{i=1}^n X_i X_i^T$ is ill-conditioned.

A special case of (4.11) is the ARX model (autoregressive model with exogenous inputs) $A(q^{-1}) Y_n = B(q^{-1}) U_{n-d} + w_n$, where $A(q^{-1}) = I + A_1 q^{-1} + \cdots + A_\ell q^{-\ell}$ and $B(q^{-1}) = B_1 + \cdots + B_r q^{-(r-1)}$ are matrix polynomials in the unit delay operator q^{-1} (defined by $q^{-1} x_n = x_{n-1}$) and $d \geq 1$ represents the delay. Here $X_n^T = (Y_{n-1}^T, \ldots, Y_{n-\ell}^T, U_{n-d}^T, \ldots, U_{n-d-r-1}^T)$. While the least squares

estimate of $\theta = (A_1, \ldots, A_\ell,\ B_1, \ldots B_r)$ has a simple recursive form (4.12) for on-line updating, a major topic in the literature is the development of similar recursive algorithms for the ARMAX model (in which MA stands for "moving average" disturbances):

$$A(q^{-1})Y_n = B(q^{-1})U_{n-d} + C(q^{-1})w_n, \tag{4.13}$$

in which $C(q^{-1}) = I + C_1 q^{-1} + \cdots + C_\nu q^{-\nu}$. Many recursive algorithms have been proposed in the literature for the ARMAX model (cf. Ljung and Söderström (1983)). However, little was known about their statistical properties until the 1980s, when convergence results were established for the stochastic gradient algorithm by Goodwin, Ramadge and Caines (1981) and for extended least squares algorithms by Lai and Wei (1986). These recursive estimates, however, are asymptotically less efficient than the off-line maximum likelihood estimates. Lai and Ying (1991) showed how parallel recursive algorithms can be developed to attain full asymptotic efficiency. The basic idea is to use an extended least squares algorithm AML, which is consistent but inefficient, to monitor a recursive maximum likelihood algorithm (RML2) that is obtained by linearizing the likelihood function in some neighborhood of $\hat{\theta}_{n-1}$. Thus, AML and RML2 are run in parallel.

The stochastic gradient algorithm in Goodwin, Ramadge and Caines (1981) is a special case of stochastic approximation, which will be considered in the next section. Stochastic approximation offers a simple approach to construct and to analyze recursive estimators; see Nevel'son and Has'minskii (1973), Fabian (1978) and Ljung and Söderström (1983).

5. Stochastic Approximation and Sequential Optimization

Consider the regression model

$$y_i = M(x_i) + \epsilon_i \quad (i = 1, 2, \ldots) \tag{5.1}$$

where y_i denotes the response at the design level x_i, M is an unknown regression function, and ϵ_i represents unobservable noise (error). In the deterministic case (where $\epsilon_i = 0$ for all i), Newton's method for finding the root θ of a smooth function M is a sequential scheme defined by the recursion

$$x_{n+1} = x_n - y_n/M'(x_n). \tag{5.2}$$

When errors ϵ_i are present, using Newton's method (5.2) entails that

$$x_{n+1} = x_n - M(x_n)/M'(x_n) - \epsilon_n/M'(x_n). \tag{5.3}$$

Hence, if x_n should converge to θ, so that $M(x_n) \to 0$ and $M'(x_n) \to M'(\theta)$ (assuming M to be smooth and to have a unique root θ such that $M'(\theta) \neq 0$),

then (5.3) implies that $\epsilon_n \to 0$, which is not possible for many kinds of random errors ϵ_i (e.g., when the ϵ_i are i.i.d. with mean 0 and variance $\sigma^2 > 0$). To dampen the effect of the errors ϵ_i, Robbins and Monro (1951) replaced $1/M'(x_n)$ in (5.2) by constants that converge to 0. Specifically, assuming that

$$\inf_{\epsilon \le |x-\theta| \le \epsilon^{-1}} \{M(x)/(x-\theta)\} > 0 \quad \text{for all} \quad 0 < \epsilon < 1, \tag{5.4}$$

the Robbins-Monro scheme is defined by the recursion

$$x_{n+1} = x_n - \alpha_n y_n \quad (x_1 = \text{ initial guess of } \theta), \tag{5.5}$$

where α_n are positive constants such that

$$\sum_1^\infty \alpha_n^2 < \infty, \quad \sum_1^\infty \alpha_n = \infty. \tag{5.6}$$

It is well known that many stochastic models of random noise ϵ_n (e.g., i.i.d. random variables with mean 0 and finite variance, L_2-bounded martingale difference sequences) have the property that

$$\sum_1^\infty \alpha_n \epsilon_n \text{ converges a.s. for all contants } \alpha_n \text{ such that } \sum_1^\infty \alpha_n^2 < \infty. \tag{5.7}$$

Assuming that $\{\epsilon_n\}$ satisfies (5.7) and that M satisfies (5.4) and

$$|M(x)| \le c|x| + d \text{ for some } c, d \text{ and all } x, \tag{5.8}$$

Blum (1954) showed that the Robbins-Monro scheme (5.5) indeed converges a.s. to θ. Earlier, Robbins and Monro (1951) showed that the scheme converges to θ in L_2 under additional assumptions. It should be noted that although Blum's assumptions on $\{\epsilon_n\}$ were stronger than (5.7), his proof in fact consisted of showing that the scheme converges if (5.7) obtains.

Noting that maximization of a smooth unimodal regression function $M : \mathbf{R} \to \mathbf{R}$ is equivalent to solving the equation $M'(x) = 0$, Kiefer and Wolfowitz (1952) proposed the following recursive maximization scheme

$$x_{n+1} = x_n + \alpha_n \triangle(x_n), \tag{5.9}$$

where at the nth stage observations $y_n^{(1)} = M(x_n^{(1)}) + \epsilon_n^{(1)}$ and $y_n^{(2)} = M(x_n^{(2)}) + \epsilon_n^{(2)}$ are taken at the design levels $x_n^{(1)} = x_n - c_n$ and $x_n^{(2)} = x_n + c_n$, respectively, α_n and c_n are positive constants such that

$$\sum_1^\infty (\alpha_n/c_n)^2 < \infty, \ \sum_1^\infty \alpha_n = \infty, \ c_n \to 0, \tag{5.10}$$

and $\triangle(x_n) = (y_n^{(2)} - y_n^{(1)})/2c_n$ is an estimate of $M'(x_n)$. Blum (1954) proved a.s. convergence of the Kiefer-Wolfowitz scheme to the maximizer of M under certain assumptions on $M, \epsilon_n^{(1)}$ and $\epsilon_n^{(2)}$.

Beginning with the seminal papers of Robbins and Monro (RM) and Kiefer and Wolfowitz (KW), there is a vast literature on stochastic approximation (SA) schemes of the type (5.5) and (5.9) in statistics and engineering. In particular, for the case of i.i.d. ϵ_i with mean 0 and variance σ^2 in (5.1), it has been shown by Sacks (1958) that an asymptotically optimal choice of α_n in the RM scheme (5.5) is $\alpha_n \sim (n\beta)^{-1}$, for which $\sqrt{n}(x_n - \theta)$ has a limiting normal distribution with mean 0 and variance σ^2/β^2, assuming that $\beta := M'(\theta) > 0$. This led Venter (1967) and Lai and Robbins (1979) to develop adaptive SA schemes of the form

$$x_{n+1} = x_n - y_n/nb_n, \tag{5.11}$$

in which b_n is an estimate of β based on the current and past observations. It is relatively easy to estimate β consistently, by using $b_n = \zeta_n \vee \{\xi_n \wedge n^{-1} \sum_{j=1}^{n-1} (y_j^{(2)} - y_j^{(1)})/(2c_j)\}$, with $\zeta_n \to 0$ and $\xi_n \to \infty$ and with suitably chosen c_j, as in the KW scheme. Venter took $y_n = (y_n^{(1)} + y_n^{(2)})/2$ in (5.11) as an estimate of the unobserved response at x_n, and chose c_j of the order $j^{-\gamma}$ for some $1/4 < \gamma < 1/2$ to meet the goal of attaining the variance σ^2/β^2 in the limiting distribution of $\sqrt{n}(x_n - \theta)$. Motivated by adaptive control applications in engineering (as reviewed in Section 4.3) and economics (as in Anderson and Taylor (1976)), Lai and Robbins (1979) considered the efficiency of the final estimate x_N of θ as well as the "regret" (or cost) $\sum_{i=1}^N (x_i - \theta)^2$ of the design. The rationale is that if θ were known, then the inputs for the finite-horizon regulation problem of minimizing $E(\sum_{i=1}^N y_i^2)$ should be set at $x_i \equiv \theta$. While Venter's design has regret of the algebraic order of constant times $N^{1-2\gamma}$, Lai and Robbins (1979) showed that it is possible to have both asymptotically minimal regret and efficient final estimate, i.e.,

$$\sum_{i=1}^N (x_i - \theta)^2 \sim (\sigma^2/\beta^2) \log N \text{ a.s. and } \sqrt{N}(x_N - \theta) \Rightarrow N(0, \sigma^2/\beta^2) \tag{5.12}$$

as $N \to \infty$, by using a modified least squares estimate in (5.11) or by certain modifications of Venter's design. Wei (1987) generalized (5.12) to the multivariate case in which x_n, y_n, θ and $M(\theta)$ are $p \times 1$ vectors and $1/b_n$ is replaced by a $p \times p$ matrix A_n that estimates the inverse of the Jacobian matrix $\partial M/\partial\theta$ using a modified Venter-type design.

Asymptotic normality of the KW scheme (5.9) has also been established by Sacks (1958). However, instead of the usual $n^{-1/2}$ rate, one has the $n^{-1/3}$

rate for the choices $\alpha_n = a/n$ and $c_n = cn^{-1/6}$, assuming M to be three times continuously differentiable in some neighborhood of θ. The reason for the slower rate is that the estimate $\triangle(x)$ of $M'(x)$ has a bias of the order $O(c_n^2)$ when $M'''(x) \neq 0$. This slower rate is common to nonparametric regression and density estimation problems, where it is known that the rate of convergence can be improved by making use of additional smoothness of M. Fabian (1967, 1971) showed how to redefine $\triangle(x_n)$ in (5.9) when M is $(s+1)$-times continuously differentiable in some neighborhood of θ for even integers s, so that letting $c_n = cn^{-1/(2s+2)}, n^{s/(2s+2)}(x_n - \theta)$ has a limiting normal distribution.

In the engineering literature, SA schemes are usually applied to multivariate problems and dynamical systems, instead of to (static) regression functions considered in the statistics literature. Besides the dynamics in the SA recursion, the dynamics of the underlying stochastic system also plays a basic role in the convergence analysis. Ljung (1977) developed the so-called ODE method that has been widely used in such convergence analysis in the engineering literature; it studies the convergence of SA or other recursive algorithms in stochastic dynamical systems via the stability analysis of an associated ordinary differential equation (ODE) that defines the "asymptotic paths" of the recursive scheme; see Kushner and Clark (1978) and Benveniste, Metivier and Priouret (1987). Moreover, a wide variety of KW-type algorithms have been developed for constrained or unconstrained optimization of objective functions on-line in the presence of noise. For $M : \mathbf{R}^p \to \mathbf{R}$, Spall (1992) introduced "simultaneous perturbation" SA schemes that take only 2 (instead of $2p$) measurements to estimate a smoothed gradient approximation to $\nabla M(\theta)$ at every stage; see also Spall and Cristion (1994). For other recent developments of SA in the engineering literature, see Kushner and Yin (1997). Ruppert (1991) gives a brief survey of the statistics literature on SA up to 1990.

6. Adaptive Treatment Allocation and the Multi-Armed Bandit Problem

The pioneering paper of Robbins (1952) introduced the multi-armed bandit problem and stochastic approximation as two new problems in sequential design of experiments, "different from those usually met in statistical literature". The latter problem deals with the continuous case where the design points can be chosen sequentially from $\mathbf{R}$ or a finite interval to maximize $E(\sum_{i=1}^N y_i)$, in which y_i is given by the regression model (5.1), while the former problem deals with the finite case where the design chooses sequentially which of k populations to sample from so that $E(\sum_{i=1}^N y_i)$ is maximized. Actually Robbins considers only the case $k = 2$, but his results in fact apply to any k. He cites Wald's (1950)

book on statistical decision theory as "the first significant contribution to the theory of sequential design", and comments that although it "states the problem in full generality" it does not provide concrete solutions to these two particular sequential design problems. Both problems have subsequently been extensively studied and undergone far-reaching generalizations, and have become fundamental problems in the field of stochastic adaptive control; see Kumar (1985) and Benveniste, Metivier and Priouret (1987).

The "multi-armed bandit problem" derives its name from an imagined slot machine with $k \geq 2$ arms. When an arm is pulled, the player wins a random reward. For each arm j, there is an unknown probability distribution Π_j of the reward, and the player's problem is to choose N successive pulls on the k arms so as to maximize the total expected reward. The problem is prototypical of a general class of adaptive control problems in which there is a fundamental dilemma between "information" (such as the need to learn from all populations about their parameter values) and "control" (such as the objective of sampling only from the best population), cf. Kumar (1985). Suppose Π_j has finite mean $\mu(\theta_j)$ and density function $f(x;\theta_j)$ with respect to some nondegenerate measure ν, where $f(\cdot;\cdot)$ is known and the θ_j are unknown parameters belonging to some set Θ. An adaptive allocation rule consists of a sequence of random variables $\phi_1, \phi_2, \ldots$ taking values in $\{1, \ldots, k\}$ such that the event $\{\phi_t = j\}$ ("X_{t+1} is sampled from Π_j") belongs to the σ-field generated by $\phi_0, X_1, \phi_1, \ldots, X_{t-1}, \phi_{t-1}, X_t$. Let $\theta = (\theta_1, \ldots, \theta_k)$. If θ were known, then we would sample from the population $\Pi_{j(\theta)}$ with the largest mean; i.e., $\mu_\theta^* := \max_{1 \leq j \leq k} \mu(\theta_j) = \mu(\theta_{j(\theta)})$. In ignorance of θ, the problem is to sample $X_1, X_2, \ldots$ sequentially from the k populations to maximize $E_\theta(\sum_{i=1}^N X_i)$, or equivalently to minimize the regret

$$R_N(\theta) = N\mu_\theta^* - E_\theta\left(\sum_{i=1}^N X_i\right) = \sum_{j:\mu(\theta_j)<\mu_\theta^*} (\mu_\theta^* - \mu(\theta_j)) E_\theta T_N(j), \tag{6.1}$$

where $T_N(j) = \sum_{t=1}^N I_{\{\phi_{t-1}=j\}}$ is the total number of observations from Π_j up to stage N.

Lai and Robbins (1985) showed how to construct sampling rules for which $R_N(\theta) = O(\log N)$ at every θ. These rules are called "uniformly good". They also developed asymptotic lower bounds for the regret $R_N(\theta)$ of uniformly good rules and showed that the rules constructed actually attain these asymptotic lower bounds and are therefore asymptotically efficient. Specifically, they showed that under certain regularity conditions

$$\liminf_{N\to\infty} R_N(\theta)/\log N \geq \sum_{j:\mu(\theta_j)<\mu_\theta^*} (\mu_\theta^* - \mu(\theta_j))/I(\theta_j, \theta^*), \tag{6.2}$$

for uniformly good rules, where $\theta^* = \theta_{j(\theta)}$ and $I(\lambda, \lambda')$ is the Kullback-Leibler information number. Their result was subsequently generalized by Anantharam, Varaiya and Walrand (1987) to the multi-armed bandit problem in which each Π_j represents an aperiodic, irreducible Markov chain on a finite state space S so that the successive observations from Π_j are no longer independent but are governed by a Markov transition density $p(x, y; \theta_j)$. This extension was motivated by the more general problem of adaptive control of finite-state Markov chains with a finite control set, whose development up to the mid-1980s has been surveyed by Kumar (1985).

Let $\{X_n, n \geq 0\}$ be a controlled Markov chain on state space S, with a parametric family of transition density functions $p(x, y; u, \theta)$ with respect to some measure M on S, where u belongs to a control set U and θ is an unknown parameter taking values in a compact metric space Θ. Thus the transition probability measure under control action u and parameter θ is given by $P_\theta^u(X_{n+1} \in A | X_n = x) = \int_A p(x, y; u, \theta) dM(y)$. The initial distribution of X_0 under P_θ^u is also assumed to be absolutely continuous with respect to M. Let $G = \{g_1, \ldots, g_k\}$ be a finite set of stationary control laws $g_j : S \to U$ such that for every $g \in G$, the transition probability function $P_\theta^{g(x)}(x, A)$ is irreducible with respect to some maximal irreducibility measure and has stationary distribution π_θ^g. Let $r(X_t, u_t)$ represent the one-step reward at time t, where $r : S \times U \to \mathbf{R}$, and define the long-run average reward

$$\mu_\theta(g) = \int r(x, g(x)) d\pi_\theta^g(x), \tag{6.3}$$

assumed to be finite. If θ were known, then one would use the stationary control law $g_{j(\theta)}$ such that

$$\mu_\theta^* := \max_{g \in G} \mu_\theta(g) = \mu_\theta(g_{j(\theta)}). \tag{6.4}$$

Suppose there is no switching cost among the (typically equivalent) optimal stationary control laws that attain the maximum in (6.4) and a cost $a(\theta)$ for each switch from one $g \in G$ to another $g' \in G$ when g and g' are not both optimal. An *adaptive control rule* ϕ is a sequence of random variables $\phi_1, \phi_2, \ldots$ taking values in G such that $\{\phi_t = g\} \in \mathcal{F}_t$ for all $g \in G$ and $t \geq 0$, where $\mathcal{F}_t$ is the σ-field generated by $X_0, \phi_0, \ldots, X_{t-1}, \phi_{t-1}, X_t$. We can generalize (6.1) to controlled Markov chains by letting

$$R_N(\theta) = \sum_{g \in G : \mu_\theta(g) < \mu_\theta^*} (\mu_\theta^* - \mu_\theta(g)) E_\theta T_N(g), \quad \text{with} \quad T_N(g) = \sum_{t=0}^{N-1} I_{\{\phi_t = g\}}. \tag{6.5}$$

In view of the additional switching cost $a(\theta)$ for each switch between two control laws in G, not both optimal, we define the overall regret to be $R_n(\theta)+a(\theta)S_n(\theta)$, where

$$S_N(\theta) = E_\theta\left(\sum_{t=1}^{N} I_{\{\phi_t\neq\phi_{t-1},\min(\mu_\theta(\phi_t),\mu_\theta(\phi_{t-1}))<\mu_\theta^*\}}\right).$$

An adaptive control rule ϕ is said to be *uniformly good* if

$$R_N(\theta) = O(\log N) \quad \text{and} \quad S_N(\theta) = o(\log N) \quad \text{for every} \quad \theta\in\Theta. \tag{6.6}$$

Graves and Lai (1997) extended the asymptotic lower bound (6.2) for the regret $R_N(\theta)$ in multi-armed bandits to a similar lower bound for the regret in (6.5) for controlled Markov chains. Moreover, by making use of sequential testing theory, they constructed uniformly good adaptive control rules that attain this asymptotic lower bound.

Besides control engineering, the theory of multi-armed bandits also has an extensive literature in economics. In particular, it has been applied to pricing under demand uncertainty, decision making in labor markets, general search problems and resource allocation (cf. Rothschild (1974), Mortensen (1985), Banks and Sundaram (1992), Brezzi and Lai (2000)). Unlike the formulation above, the formulation of adaptive allocation problems in the economics literature involves a discount factor that relates future rewards to their present values. Moreover, an economic agent typically incorporates his prior beliefs about the unknown parameters into his choice of actions. Suppose an agent chooses actions sequentially from a finite set $\{a_1,\ldots,a_k\}$ such that the reward $r(a_j)$ of action a_j has a probability distribution depending on an unknown parameter θ_j which has a prior distribution $\Pi^{(j)}$. The agent's objective is to maximize the total discounted reward

$$\int\cdots\int E_{\theta_1,\ldots,\theta_k}\left\{\sum_{t=0}^{\infty}\beta^t r(u_{t+1})\right\} d\Pi^{(1)}(\theta_1)\ldots d\Pi^{(k)}(\theta_k), \tag{6.7}$$

where $0<\beta<1$ is a discount factor and u_t denotes the action chosen by the agent at time t. The optimal solution to this problem, commonly called the "discounted multi-armed bandit problem", was shown by Gittins and Jones (1974) and Gittins (1979) to be the "index rule" that chooses at each stage the action with the largest "dynamic allocation index" (DAI).

The DAI is a complicated functional of the posterior distribution $\Pi_n^{(j)}$ given the rewards $Y_{j,1},\ldots,Y_{j,T_n(j)}$ of action a_j up to stage n. Bounds and approximations to the DAI have been developed by Brezzi and Lai (2000) and Chang and Lai (1987). Making use of these bounds, Brezzi and Lai (2000) gave a simple proof of the incompleteness of learning from endogenous data by an optimizing

economic agent. Specifically, they showed that with positive probability the index rule uses the optimal action only finitely often and that it can estimate consistently only one of the θ_j, generalizing Rothschild's (1974) "incomplete learning theorem" for Bernoulli two-armed bandits. Moreover, the DAI can be written as an "upper confidence bound" of the form $\mu_{j,n} + \sqrt{v_{j,n}}\ \psi(\beta, \Pi_n^{(j)})$, where $\mu_{j,n}$ and $v_{j,n}$ are the mean and variance of the posterior distribution $\Pi_n^{(j)}$ and ψ is a nonnegative function of β and $\Pi_n^{(j)}$. When the $Y_{j,i}$ are normal with mean θ_j and variance 1 and the prior distribution $\Pi^{(j)}$ is normal, Chang and Lai (1987) showed that $\psi(\beta, \Pi_n^{(j)})$ can be expressed as

$$\tilde{\psi}(s) = \{2\log s - \log\log s - \log 16\pi + o(1)\}^{1/2} \tag{6.8}$$

as $\beta \to 1$ and $s := v_{j,n}/(1-\beta) \to \infty$.

There is also a similar asymptotic theory of the finite-horizon bandit problem in which the agent's objective is to maximize the total reward

$$\int \cdots \int E_{\theta_1,\ldots,\theta_k} \left\{ \sum_{t=0}^{N-1} r(u_{t+1}) \right\} d\Pi(\theta_1, \ldots, \theta_k), \tag{6.9}$$

where Π is a prior distribution of the vector $(\theta_1, \ldots, \theta_k)$. Even when the θ_i are independent under Π (so that Π is a product of marginal distributions as in (6.7)), the optimal rule that maximizes (6.9) does not reduce to an index rule. In principle, one can use dynamic programming to maximize (6.9). In the case of $k = 2$ Bernoulli populations with independent Beta priors for their parameters, Fabius and van Zwet (1970) and Berry (1972) studied the dynamic programming equations analytically and obtained several qualitative results concerning the optimal rule. Lai (1987) showed that although index-type rules do not provide exact solutions to the optimization problem (6.9), they are asymptotically optimal as $N \to \infty$, and have nearly optimal performance from both the Bayesian and frequentist viewpoints for moderate and small values of N.

The starting point in Lai's approximation to the optimal rule is to consider the normal case. Suppose that an experimenter can choose at each stage $n(\le N)$ between sampling from a normal population with known variance 1 but unknown mean μ and sampling from another normal population with known mean 0. Assuming a normal prior distribution $N(\mu_0, v)$ on μ, the optimal rule that maximizes the expected sum of N observations samples from the first population (with unknown mean) until stage $T^* = \inf\{n \le N : \hat{\mu}_n + a_{n,N} \le 0\}$ and then takes the remaining observations from the second population (with known mean 0), where $\hat{\mu}_n$ is the posterior mean based on observations $Y_1, \ldots, Y_n$ from the first population and $a_{n,N}$ are positive constants that can be determined by backward induction. Writing $t = n/N$, $w(t) = (Y_1 + \cdots + Y_n)/\sqrt{N}$, and treating $0 < t \le 1$

as a continuous variable, Lai (1987) approximates $a_{n,N}$ by $\sqrt{v_n}\ h(n/N)$, where v_n is the posterior variance of μ and

$$h(t) = \begin{cases} \{2\log t^{-1} - \log\log t^{-1} - \log 16\pi \\ \quad +0.99\exp(-0.038t^{-1/2})\}^{1/2} & \text{if } 0 < t \le 0.01, \\ -1.58\sqrt{t} + 1.53 + 0.07t^{-1/2} & \text{if } 0.01 < t \le 0.28, \\ -0.576t^{3/2} + 0.299t^{1/2} + 0.403t^{-1/2} & \text{if } 0.28 < t \le 0.86, \\ t^{-1}(1-t)^{1/2}\{0.639 - 0.403(t^{-1}-1)\} & \text{if } 0.86 < t \le 1. \end{cases} \tag{6.10}$$

This function h is obtained by evaluating numerically the boundary of the corresponding optimal stopping problem for Brownian motion, first studied by Chernoff and Ray (1965), and then developing some simple closed-form approximation to the boundary. Although it differs from the function $\tilde{\psi}$ in (6.8) because of the difference between the finite-horizon criterion and the discounted criterion, note that $h(t) = \tilde{\psi}(t^{-1}) + o(1)$ as $t \to 0$.

More generally, without assuming a prior distribution on the unknown parameters, suppose $Y_{j,1}, Y_{j,2}, \ldots$, are independent random variables from a one-parameter exponential family with density function $f_{\theta_j}(y) = \exp\{\theta_j y - \omega(\theta_j)\}$ with respect to some dominating measure. Then $\mu(\theta) = E_\theta Y_1 = \omega'(\theta)$ is increasing in θ since $\mathrm{Var}(Y_{j,1}) = \omega''(\theta_j)$, and the Kullback-Leibler information number is $I(\theta,\lambda) = E_\theta[\log(f_\theta(Y)/f_\lambda(Y))] = (\theta-\lambda)\mu(\theta) - (\omega(\theta)-\omega(\lambda))$. Assuming that all θ_j lie in some open interval Γ such that $\inf_{\theta\in\Gamma}\omega''(\theta) > 0$ and $\sup_{\theta\in\Gamma}\omega''(\theta) < \infty$, and letting $\hat{\theta}_{j,n}$ be the maximum likelihood estimate of θ_j based on $Y_{j,1}, \ldots, Y_{j,n}$, Lai (1987) considered an upper confidence bound for $\mu(\theta_j)$ of the form $\mu(\theta^*_{j,n})$, where

$$\theta^*_{j,n} = \inf\{\theta \in \Gamma :\ \theta \ge \hat{\theta}_{j,n}\ ,\ 2nI(\hat{\theta}_{j,n},\theta) \ge h^2(n/N)\}. \tag{6.11}$$

Note that $nI(\hat{\theta}_{j,n},\theta_0)$ is the GLR statistic for testing $\theta = \theta_0$, so the above upper confidence bound is tantamount to the usual construction of confidence limits by inverting an equivalent test. Lai (1987) showed that this upper confidence bound rule is uniformly good and attains the lower bound (6.2) not only at fixed $(\theta_1, \ldots, \theta_k)$ as $N \to \infty$ (so that the rule is asymptotically optimal from the frequentist viewpoint), but also uniformly over a wide range of parameter configurations, which can be integrated to show that the rule is asymptotically Bayes with respect to a large class of prior distributions Π for $(\theta_1, \ldots, \theta_k)$. There is also an analogous asymptotic theory for the discounted multi-armed bandit problem as $\beta \to 1$, as shown by Chang and Lai (1987).

Adaptive treatment allocation has also been studied in the statistics literature in the context of clinical trials. Bather (1985) and Basu, Bose and Ghosh

(1991) survey many important developments and basic results in this direction. Unfortunately, except for the allocation schemes to achieve balance following Efron's (1971) seminal paper on the biased coin design, these adaptive treatment allocation methods, which are mostly targeted towards minimizing the number of patients receiving the inferior treatment(s), "(have) found no application whatsoever in the actual conduct of trials", as noted by Armitage (1985), because of the practical difficulties in implementing them and because the objectives of clinical trials are much more complex than the simple criterion of minimizing the expected number of patients receiving the inferior treatment, subject to either a given patient horizon or a prescribed type I error probability.

7. Future Opportunities and Challenges

Sequential analysis has been developing steadily but at a somewhat uneven pace during the past six decades. There is now a rich arsenal of techniques and concepts, methods and theories, that will provide a strong foundation for further advances and breakthroughs. The subject is still vibrant after six decades of continual development, with many important unsolved problems and with new interesting problems brought in from other fields.

Indeed, new directions and new problems in sequential analysis have been inspired by applications to engineering, economics and medicine, as the preceding sections have shown. There are basic sequential statistical problems in these fields, which have their own specialists and journals. For sequential analysis to remain a vibrant statistical subject, it has to grow not only inwards in the form of further methodological advances and breakthroughs, but also outwards through active involvement in the biomedical, socio-economic and engineering sciences. Integrating its internal and external growth is a new challenge that will provide it with increasing opportunities and visible roles in the changing world of science, technology and socio-economic activities.

In his critique of previous work on adaptive treatment allocation in clinical trials, Armitage (1985) suggests "that statisticians concerned with the development of optimization models and those concerned directly in clinical trials should meet to discuss the feasibility of these methods for various sorts of trials" and "that members of the two groups should work in collaboration on specific trials so as to foster closer understanding and to explore the possibilities in a realistic setting". As we have pointed out in Sections 2 and 4, the relatively small sample size and the multi-center environment, together with the multiplicity of study objectives, make conventional sequential testing procedures and adaptive treatment allocation procedures unattractive to the clinical trials community. To develop methods that are usable in practice, modifications and refinements are

needed. Considerable work in this direction has been done in sequential testing, as reviewed in Section 2. This illustrates the need for collaborative, interdisciplinary research to bridge the gap between statistical methods and substantive applications, which is another challenge for sequential analysis in the twenty-first century.

The difficulties in using adaptive treatment allocation rules in clinical trials pointed out by Armitage (1985) and the discussants of his paper disappear in the applications to control engineering and economics reviewed in Section 6. A challenge for sequential analysis, therefore, is to identify the right "clientele" for its variety of tools and methods. We next give an example of an emerging clientele from financial economics in which we show how sequential analysis techniques can be applied to the valuation of American options.

To begin with, a call (put) option gives the holder the right to buy (sell) the underlying asset by a certain date, known as the "expiration date" or "maturity", at a certain price, known as the exercise (or strike) price. European options can be exercised only on the expiration date. In the case of a complete market in which the price of the underlying asset can be modeled by geometric Brownian motion, there is the celebrated Black-Scholes formula for pricing European options. In contrast, an American option can be exercised at any time up to the expiration date. Except for American calls written on non-dividend-paying stocks that reduce to European calls, American option valuation does not have explicit formulas, and an active area of research in the past two decades is the development of fast and accurate numerical methods for the valuation of option books and implementation of dynamic hedging strategies. AitSahlia and Lai (1999) recently made use of ideas similar to those used in developing the approximation (6.10) for the multi-armed bandit problem to come up with a simple and accurate method to compute the values and early exercise boundaries of American options. A key idea underlying the method is the reduction of American option valuation to a *single* optimal stopping problem for standard Brownian motion, indexed by one parameter in the absence of dividends, and by two parameters in the presence of a dividend rate. Unlike the commonly used Cox-Ross-Rubinstein (1979) method that is based on approximating the underlying geometric Brownian motion by a binomial tree with the given spot price as the root node, AitSahlia and Lai (1999) first use a change of variables to reduce the optimal stopping problem to a canonical form involving a standard Brownian motion and then solve the optimal stopping problem by using the numerical method developed by Chernoff and Petkau (1986). Motivated by applications to sequential analysis, in which the stopping boundary associated with the optimal sequential procedure is of primary interest, Chernoff and Petkau (1986)

developed a continuity correction method to address the issue of approximating Brownian motion by a discrete-time and discrete-state Bernoulli random walk. After computing the optimal exercise boundary, the American option prices for different maturities and spot prices are readily computed by using the following decomposition formula of Carr, Jarrow and Myneni (1992):

$$\text{American option price} = \text{European option price} + \text{Early exercise premium},$$

in which the early exercise premium is a one-dimensional integral whose integrand is an explicit function of the optimal exercise boundary. Numerical results show that the optimal exercise boundary can be well approximated, in the canonical space-time scale, by a linear spline with a few knots. This has led to two fast and accurate approximations to the prices and hedge parameters of American options proposed in AitSahlia and Lai (1999), which show marked improvements in both speed and accuracy over previous approximations in the literature.

Sequential analysis has both contributed to and benefited from developments in many areas of statistics and probability. For example, the work of Wald and Wolfowitz (1948) on the optimality of the SPRT marked the beginning of optimal stopping theory, while the papers of Wald (1950) and Arrow, Blackwell and Girshick (1951) on sequential decision theory marked the beginning of the theory of stochastic control and dynamic programming. On the other hand, renewal theory and martingale theory have provided powerful tools to analyze the SPRT and randomly stopped likelihood ratio statistics. Extension of the renewal-theoretic tools to handle nonlinear stopping boundaries and nonlinear functions of sample sums led to the development of nonlinear renewal theory by Woodroofe (1976), Lai and Siegmund (1977, 1979) and Zhang (1988), which in turn has made nonlinear problems in sequential analysis much more tractable. Concerning the construction of confidence intervals following sequential tests in Section 4.2, the resampling methods of Chuang and Lai benefited from the comprehensive bootstrap theory developed in the past two decades, while Woodroofe's asymptotic expansions benefited from Stein's (1981) identity in the theory of multivariate normal mean estimation. Strengthening and increasing such interactions with other branches of statistics and probability is another challenge that will be important for sequential analysis to enhance its growth and impact in the field of statistics.

Acknowledgement

This research was supported by the National Science Foundation, the National Security Agency and the Center for Advanced Study in the Behavioral Sciences.

References

AitSahlia, F. and Lai, T. L. (1999). A canonical optimal stopping problem for American options and its numerical solution. *J. Comput. Finance* **3**, 33-52.

Anantharam, V., Varaiya, P. and Walrand, J. (1987). Asymptotically efficient allocation rules for multi-armed bandit problems with multiple plays: II. Markovian rewards. *IEEE Trans. Automat. Contr.* **32**, 968-982.

Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size. *Ann. Math. Statist.* **31**, 165-197.

Anderson, T. W. and Taylor, J. (1976). Some experimental results on the statistical problem of least squares estimates in control problems. *Econometrica* **44**, 1289-1302.

Anscombe, F. J. (1963). Sequential medical trials. *J. Amer. Statist. Assoc.* **58**, 365-384.

Aras, R. and Woodroofe, M. (1993). Asymptotic expansions for moments of a randomly stopped average. *Ann. Statist.* **21**, 503-519.

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses and its relation to discriminant function analysis. *J. Roy. Statist. Soc. Ser. B* **12**, 137-144.

Armitage, P. (1960). *Sequential Medical Trials.* Blackwell, Oxford.

Armitage, P. (1985). The search for optimality in clinical trials. *Internat. Statist. Rev.* **53**, 15-24.

Armitage, P., McPherson, C. K. and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A* **132**, 235-244.

Arrow, K. J., Blackwell, D. and Girshick, M. A. (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica* **17,** 213-244.

Assaf, D. (1988). A dynamic sampling approach for detecting a change in distribution. *Ann. Statist.* **16**, 236-253.

Banks, J. S. and Sundaram, R. K. (1992). Denumerable-armed bandits. *Econometrica* **60**, 1071-1096.

Barnard, G. A. (1959). Control charts and stochastic processes (with discussion). *J. Roy. Statist. Soc. Ser. B* **21**, 239-271.

Bartlett, M. S. (1946). The large sample theory of sequential tests. *Proc. Cambridge Phil. Soc.* **42**, 239-244.

Basseville, M. and Benveniste, A. (1983). Design and comparative study of some sequential jump detection algorithms in digital signals. *IEEE Trans. Acoustics, Speech Signal Processing* **31**, 521-535.

Basseville, M. and Nikiforov, I. V. (1993). *Detection of Abrupt Changes: Theory and Applications.* Englewood Cliffs, Prentice Hall.

Basu, A., Bose, A. and Ghosh, J. K. (1991). Sequential design and allocation rules. In *Handbook of Sequential Analysis* (Edited by B. K. Ghosh and P. K. Sen), 475-501. Dekker, New York.

Basu, A. P. (1991). Sequential methods in reliability and life testing. In *Handbook of Sequential Analysis* (Edited by B. K. Ghosh and P. K. Sen), 581-592. Dekker, New York.

Bather, J. A. (1985). On the allocation of treatments in sequential medical trials. *Internat. Statist. Rev.* **53**, 1-13.

Bechhofer, R. E., Kiefer, J. and Sobel, M. (1968). *Sequential Identification and Ranking Procedures.* University of Chicago Press.

Benveniste, A., Basseville, M. and Moustakides, G. (1987). The asymptotic local approach to change detection and model validation. *IEEE Trans. Automatic Control* **32**, 583-592.

Benveniste, A., Metivier, M. and Priouret, P. (1987). *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag, New York.

Berry, D. A. (1972). A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43**, 871-897.

BHAT (β-Blocker Heart Attack Trial Research Group) (1982). A randomized trial of propranolol in patients with acute myocardial infarction. *J. Amer. Med. Assoc.* **147**, 1707-1714.

Bickel, P. and Yahav, J. (1968). Asymptotically optimal Bayes and minimax procedures in sequential estimation. *Ann. Math. Statist.* **39**, 442-456.

Bilias, Y., Gu, M. and Ying, Z. (1997). Towards a general asymptotic theory for Cox model with staggered entry. *Ann. Statist.* **25**, 662-682.

Blum, J. (1954). Approximation methods which converge with probability one. *Ann. Math. Statist.* **25**, 382-386.

Böhm, W. and Hackl, P. (1990). Improved bounds for average run length of control charts based on finite weighted sums. *Ann. Statist.* **18**, 1895-1899.

Borisov, V. Z. and Konev, V. V. (1977). On sequential parameter estimation in discrete time process. *Automat. Remote Contr.* **38**, 58-64.

Box, G. E. P. and Luceño, A. (1997). *Statistical Control by Monitoring and Feedback Adjustment.* Wiley, New York.

Box, G. E. P. and Ramirez, J. (1992). Cumulative score charts. *Qual. Reliab. Engng.* **8**, 17-27.

Brezzi, M. and Lai, T. L. (2000). Incomplete learning from endogenous data in dynamic allocation. *Econometrica* **68**, 1511-1516.

Carr, P., Jarrow, R. and Myneni, R. (1992). Alternative characterizations of American put options. *Math. Finance* **2**, 87-106.

Chan, H. P. and Lai, T. L. (2000a). Saddlepoint approximations for Markov random walks and nonlinear boundary crossing probabilities for scan statistics. Technical Report, Department of Statistics, Stanford University.

Chan, H. P. and Lai, T. L. (2000b). Maxima of Gaussian random fields and moderate deviation approximations to boundary crossing probabilities of scan statistics. Technical Report, Department of Statistics, Stanford Univversity.

Chang, F. and Lai, T. L. (1987). Optimal stopping and dynamic allocation. *Adv. Appl. Probab.* **19**, 829-853.

Chang, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45**, 247-254.

Chang, M. N. and O'Brien, P. C. (1986). Confidence intervals following group sequential tests. *Contr. Clin. Trials* **7**, 18-26.

Chernoff, H. (1961). Sequential tests for the mean of a normal distribution. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **1**, 79-91. University of California Press.

Chernoff, H. (1965). Sequential tests for the mean of a normal distribution III (Small t). *Ann. Math. Statist.* **36**, 28-54.

Chernoff, H. and Petkau, A. J. (1981). Sequential medical trials involving paired data. *Biometrika* **68**, 119-132.

Chernoff, H. and Petkau, A. J. (1986). Numerical solutions for Bayes sequential decision problems. *SIAM J. Sci. Statist. Comput.* **7,** 46-59.

Chernoff, H. and Ray, S. N. (1965). A Bayes sequential sampling inspection. *Ann. Math. Statist.* **36**, 1387-1407.

Chow, Y. S. and Martinsek, A. (1982). Bounded regret of a sequential procedure for estimation of the mean. *Ann. Statist.* **10**, 904-914.

Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed width sequential confidence intervals for the mean. *Ann. Math. Statist.* **36,** 457-462.

Chuang, C. S. and Lai, T. L. (1998). Resampling methods for confidence intervals in group sequential trials. *Biometrika* **85**, 317-352.

Chuang, C. S. and Lai, T. L. (2000a). Hybrid resampling methods for confidence intervals (with discussion). *Statist. Sinica* **10**, 1-50.

Chuang, C. S. and Lai, T. L. (2000b). Bootstrap and hybrid resampling for confidence intervals in sequential analysis. Technical Report, Department of Statistics, Stanford University.

Cox, D. R. (1952). Estimation by double sampling. *Biometrika* **39**, 217-227.

Cox, D. R. (1963). Large sample sequential tests for composite hypotheses. *Sankhyā Ser. A* **25**, 5-12.

Cox, J., Ross, S. and Rubinstein, M. (1979). Option pricing: a simplified approach. *J. Financial Econ.* **7**, 229-264.

Crowder, S. V. (1987). A simple method for studying the run-length distribution of exponentially weighted moving average charts. *Technometrics* **29**, 401-408.

de la Peña, V. and Lai, T. L. (1997). Moments of randomly stopped U-statistics. *Ann. Probab.* **25**, 2055-2081.

DeMets, D. L., Hardy, R., Friedman, L. M. and Lan, G. K. K. (1984). Statistical aspects of early termination in the Beta-Blocker Heart Attack Trial. *Contr. Clin. Trials* **5**, 362-372.

Dodge, H. F. and Romig, H. G. (1929). A method of sampling inspection. *Bell Syst. Tech. J.* **8**, 613-631.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika* **58**, 403-417.

Efron, B. (1981). Nonparametric standard errors and confidence intervals (with discussion). *Canad. J. Statist.* **9**, 139-172.

Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *J. Amer. Statist. Assoc.* **82**, 171-200.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap.* Chapman Hall, New York.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77**, 875-892.

Epstein, B. and Sobel, M. (1955). Sequential life tests in the exponential case. *Ann. Math. Statist.* **25**, 373-381.

Fabian, V. (1967). Stochastic approximation of minima with improved asymptotic speed. *Ann. Math. Statist.* **38**, 191-200.

Fabian, V. (1971). Stochastic approximation. In *Optimizing Methods in Statistics* (Edited by J. Rustagi), 439-470. Academic Press, New York.

Fabian, V. (1978). On asymptotically efficient recursive estimation. *Ann. Statist.* **6**, 854-866.

Fabius, J. and van Zwet, W. R. (1970). Some remarks on the two-armed bandit. *Ann. Math. Statist.* **41**, 1906-1916.

Fu, K. S. (1968). *Sequential Methods in Pattern Recognition and Machine Learning.* Academic Press, New York.

Ghosh, B. K. (1970). *Sequential Tests of Statistical Hypotheses.* Addison-Wesley, Cambridge, MA.

Ghosh, B. K. (1991). A brief history of sequential analysis. In *Handbook of Sequential Analysis* (Edited by B. K. Ghosh and P. K. Sen), 1-19. Dekker, New York.

Ghosh, M., Nickerson, D. M. and Sen, P. K. (1987). Sequential shrinkage estimation. *Ann. Statist.* **15**, 817-829.

Ghosh, M., Mukhopadhyay, N. and Sen, P. K. (1997). *Sequential Estimation.* Wiley, New York.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Statist. Soc. Ser. B* **41**, 148-177.

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (Edited by J. Gani et al.), 241-266. North Holland, Amsterdam.

Goodwin, G. C., Ramadge, P. J. and Caines, P. E. (1981). Discrete time stochastic adaptive control. *SIAM J. Contr. Optimiz.* **19**, 829-853.

Graves, T. L. and Lai, T. L. (1997). Asymptotically efficient adaptive choice of control laws in controlled Markov chains. *SIAM J. Contr. Optimiz.* **35**, 715-743.

Gu, M. G. and Lai, T. L. (1991). Weak convergence of time-sequential rank statistics with applications to sequential testing in clinical trials. *Ann. Statist.* **19**, 1403-1433.

Gu, M. G. and Lai, T. L. (1998). Repeated significance testing with censored rank statistics in interim analysis of clinical trials. *Statist. Sinica* **8**, 411-423.

Gu, M. G. and Lai, T. L. (1999). Determination of power and sample size in the design of clinical trials with failure-time endpoints and interim analyses. *Contr. Clin. Trials* **20**, 423-438.

Gu, M. G. and Ying, Z. (1995). Group sequential methods for survival data using partial likelihood score processes with covariate adjustment. *Statist. Sinica* **5**, 793-804.

Gupta, S. S. and Panchapakesan, S. (1991). Sequential ranking and selection procedures. In *Handbook of Sequential Analysis* (Edited by B. K. Ghosh and P. K. Sen), 363-380. Dekker, New York.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika* **33**, 222-225.

Hall, P. (1981). Asymptotic theory of triple sampling for sequential estimation of a mean. *Ann. Statist.* **9**, 1229-1238.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansions*. Wiley, New York.

Hall, W. J. (1980). Sequential minimum probability ratio tests. In *Asymptotic Theory of Statistical Tests and Estimation* (Edited by I. M. Chakravarti), 325-350. Academic Press, New York.

Harrington, D. P. and Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553-566.

Haybittle, J. L. (1971). Repeated assessments of results in clinical trials of cancer treatment. *Brit. J. Radiol.* **44**, 793-797.

Hoeffding, W. (1960). Lower bounds for the expected sample size and average risk of a sequential procedure. *Ann. Math. Statist.* **31**, 352-368.

Hunter, J. S. (1990). Discussion of "Exponentially weighted moving average control schemes" by Lucas and Saccuci. *Technometrics* **32**, 21-22.

Jennison, C. and Turnbull, B. W. (1991). Group sequential tests and repeated confidence intervals. In *Handbook of Sequential Analysis* (Edited by B. K. Ghosh and P. K. Sen), 283-311. Dekker, New York.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall/CRC, Boca Raton and London.

Jones, D. and Whitehead, J. (1979). Sequential forms of the logrank and modified Wilcoxon tests for censored data. *Biometrika* **66**, 105-113.

Jurečková, J. (1991). Confidence sets and intervals. In *Handbook of Sequential Analysis* (Edited by B. K. Ghosh and P. K. Sen), 269-281. Dekker, New York.

Kiefer, J. and Weiss, L. (1957). Some properties of generalized sequential probability ratio tests. *Ann. Math. Statist.* **28**, 57-74.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.* **23**, 462-466.

Konev, V. and Lai, T. L. (1995). Estimators with prescribed precision in stochastic regression models. *Sequential Anal.* **14**, 179-192.

Kumar, P. R. (1985). A survey of some results in stochastic adaptive control. *SIAM J. Contr. Optimiz.* **23**, 329-380.

Kushner, H. J. and Clark, D. S. (1978). *Stochastic Approximation for Constrained and Unconstrained Systems*. Springer-Verlag, New York.

Kushner, H. J. and Yin, G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York.

Lai, T. L. (1973). Optimal stopping and sequential tests which minimize the maximum expected sample size. *Ann. Statist.* **1**, 659-673.

Lai, T. L. (1974). Control charts based on weighted sums. *Ann. Statist.* **2**, 134-147.

Lai, T. L. (1981). Asymptotic optimality of invariant sequential probability ratio tests. *Ann. Statist.* **9**, 318-333.

Lai, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15**, 1091-1114.

Lai, T. L. (1988). Nearly optimal sequential tests of composite hypotheses. *Ann. Statist.* **16**, 856-886.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **57**, 613-658.

Lai, T. L. (1997). On optimal stopping problems in sequential hypothesis testing. *Statist. Sinica* **7**, 33-51.

Lai, T. L. (1998). Information bounds and quick detection of parameter changes in stochastic systems. *IEEE Trans. Information Theory* **44**, 2917-2929.

Lai, T. L. (2000). Sequential multiple hypothesis testing and efficient fault detection-isolation in stochastic systems. *IEEE Trans. Information Theory* **46**, 595-608.

Lai, T. L., Levin, B., Robbins, H. and Siegmund, D. (1980). Sequential medical trials. *Proc. Natl. Acad. Sci. USA* **79**, 3135-3138.

Lai, T. L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* **7**, 1196-1221.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6**, 4-22.

Lai, T. L. and Shan, J. Z. (1999). Efficient recursive algorithms for detection of abrupt changes in signals and systems. *IEEE Trans. Automat. Contr.* **44**, 952-966.

Lai, T. L. and Siegmund, D. (1977, 1979). A nonlinear renewal theory with applications to sequential analysis: I. *Ann. Statist.* **5**, 946-954; II. *Ann. Statist.* **7**, 60-76.

Lai, T. L. and Siegmund, D. (1983). Fixed accuracy estimation of an autoregressive parameter. *Ann. Statist.* **11**, 478-485.

Lai, T. L. and Wang, J. Q. (1994). Asymptotic expansions of stopped random walks and first passage times. *Ann. Probab.* **22**, 1957-1992.

Lai, T. L. and Wei, C. Z. (1986). Extended least squares and their applications to adaptive control and prediction in linear systems. *IEEE Trans. Automat. Contr.* **31**, 898-906.

Lai, T. L. and Ying, Z. (1991). Recursive identification and adaptive prediction in linear stochastic systems. *SIAM J. Contr. Optimiz.* **29**, 1061-1090.

Lai, T. L. and Zhang, L. (1994). A modification of Schwarz's sequential likelihood ratio tests in multivariate sequential analysis. *Sequential Anal.* **13**, 79-96.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

Lan, K. K. G. and DeMets, D. L. (1989). Group sequential procedures: calendar versus information time. *Statist. Med.* **8**, 1191-1198.

Lerche, H. R. (1986). The shape of Bayes tests of power one. *Ann. Statist.* **14**, 1030-1048.

Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Contr.* **22**, 551-575.

Ljung, L. and Söderström, T. (1983). *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, MA.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* **42**, 1897-1908.

Lorden, G. (1976). 2-SPRT's and the modified Kiefer-Weiss problem of minimizing an expected sample size. *Ann. Statist.* **4**, 281-291.

Lucas, J. M. and Saccuci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics* **32**, 1-12.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in their consideration. *Cancer Chemotherapy Rep.* **50**, 163-170.

Marcus, M. B. and Swerling, P. (1962). Sequential detection in radar with multiple resolution elements. *IRE Trans. Inform. Theory* **IT-8**, 237-245.

Martinsek, A. (1987). Empirical Bayes methods in sequential estimation. *Sequential Anal.* **6**, 119-137.

Mason, R. L., Champ, C. W., Tracy, N. D., Wierda, S. J. and Young, J. C. (1997). Assessment of multivariate process control techniques. *J. Qual. Tech.* **29**, 140-143.

MIL-STD 781C (1977). *Reliability Design Qualification and Production Acceptance Tests: Exponential Distribution*. U.S. Department of Defense, Washington, D.C.

Mortensen, D. (1985). Job search and labor market analysis. *Handbook of Labor Economics* **2**, 849-919.

Moustakides, G. V. (1986). Optimal stopping for detecting changes in distribution. *Ann. Statist.* **14**, 1379-1387.

Nevel'son, M. B. and Has'minskii, R. Z. (1973). *Stochastic Approximation and Recursive Estimation*. Amer. Math. Soc. Transl.

Nikiforov, I. V. (1995). A generalized change detection problem. *IEEE Trans. Inform. Theory* **41**, 171-181.

O'Brien, P. C. and Fleming, T. R. (1979). a multiple testing procedure for clinical trials. *Biometrics* **35**, 549-556.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41**, 100-114.

Patton, R. J., Frank, P. M. and Clark, R. N. (1989). *Fault Diagnosis in Dynamic Systems: Theory and Applications*. Prentice Hall, Englewood Cliffs, NJ.

Peto, R. (1985). Discussion of papers by J. A. Bather and P. Armitage. *Internat. Statist. Rev.* **53**, 31-34.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-199.

Pollak, M. (1978). Optimality and almost optimality of mixture stopping rules. *Ann. Statist.* **6**, 910-916.

Pollak, M. (1985). Optimal detection of a change in distribution. *Ann. Statist.* **13**, 206-227.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika* **65**, 167-179.

Reynolds, M. R., Amin, R. W. and Arnold, J. C. (1990). CUSUM charts with variable sampling intervals. *Technometrics* **32**, 371-384.

Roberts, S. W. (1966). A comparison of some control chart procedures. *Technometrics* **8**, 411-430.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58**, 527-535.

Robbins, H. (1959). Sequential estimation of the mean of a normal population. In *Probability and Statistics (Harold Cramér Volume)*, 235-245. Almquist & Wiksell, Stockholm.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *Ann. Math. Statist.* **41**, 1397-1409.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Ann. Math. Statist.* **22**, 400-407.

Rosner, G. L. and Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: a comparison of methods. *Biometrika* **75**, 723-729.

Rothschild, M. (1974). A two-armed bandit theory of market pricing. *J. Economic Theory* **9**, 185-202.

Ruppert, D. (1991). Stochastic approximation. In *Handbook of Sequential Analysis* (Edited by B. K. Ghosh and P. K. Sen), 503-529. Dekker, New York.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *Ann. Math. Statist.* **29**, 375-405.

Schwarz, G. (1962). Asymptotic shapes of Bayes sequential testing regions. *Ann. Math. Statist.* **33**, 224-236.

Self, S. G. (1991). An adaptive weighted logrank test with application to cancer prevention and screening trials. *Biometrics* **47**, 975-986.

Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70**, 315-326.

Sen, P. K. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference.* Wiley, New York.

Sen, P. K. (1991). Nonparametric methods in sequential analysis. In *Handbook of Sequential Analysis* (Edited by B. K. Ghosh and P. K. Sen), 331-362. Dekker, New York.

Shewhart, W. A. (1931). *Economic Control of Manufactured Products.* Van Nostrand Reinhold, New York.

Siegmund, D. and Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistics for sequential detection of a change-point. *Ann. Statist.* **23**, 255-271.

Siegmund, D. and Yakir, B. (2000a). Tail probabilities for the null distribution of scanning statistics. *Bernoulli* **6**, 191-213.

Simons, G. (1967). Lower bounds for average sample number of sequential multihypothesis tests. *Ann. Math. Statist.* **38**, 1343-1364.

Slud, E. V. and Wei, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *J. Amer. Statist. Assoc.* **77**, 862-868.

Sobel, M. and Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Ann. Math. Statist.* **20**, 502-522.

Spall, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Contr.* **37**, 332-341.

Spall, J. C. and Cristion, J. A. (1994). Nonlinear adaptive control using neural networks: estimation with a smoothed form of simultaneous perturbation gradient approximation. *Statist. Sinica* **4**, 1-27.

Stein, C. (1945). A two sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.* **16**, 243-258.

Stein, C. (1949). Some problems in sequential estimation. *Econometrica* **17**, 77-78.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9**, 1131-1151.

Stoumbos, Z. G., Reynolds, M. R., Ryan, T. P. and Woodall, W. H. (2000). The state of statistical process control as we proceed into the 21st century. *J. Amer. Statist. Assoc.* **95**, 992-998.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25**, 285-294.

Tsiatis, A. A. (1981). The asymptotic joint distribution of the efficient scores test for the proportional hazards model calculated over time. *Biometrika* **68**, 311-315.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Amer. Statist. Assoc.* **77**, 855-861.

Tsiatis, A. A., Rosner, G. L. and Tritchler, D. L. (1985). Group sequential tests with censored survival data adjusting for covariates. *Biometrika* **72**, 365-373.

Venter, J. H. (1967). An extension of the Robbins-Monro procedure. *Ann. Math. Statist.* **38**, 181-190.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Statist.* **16**, 117-186.

Wald, A. (1950). *Statistical Decision Functions.* Wiley, New York.

Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statist.* **19**, 326-339.

Wallis, A. W. (1980). The Statistical Research Group. *J. Amer. Statist. Assoc.* **75**, 320-334.

Wei, C. Z. (1987a). Multivariate adaptive stochastic approximation. *Ann. Statist.* **15**, 1115-1130.

Wetherill, G. B. (1963). Sequential estimation of quantal response curves (with discussion). *J. Roy. Statist. Soc. Ser. B* **25**, 1-48.

Weng, R. C. and Woodroofe, M. (2000). Integrable expansions for posterior distributions for multiparameter exponential families with applications to sequential confidence intervals. *Statist. Sinica* **10**, 693-713.

Willsky, A. S. and Jones, H. G. (1976). A generalized likelihood ratio approach to detection and estimation of jumps in linear systems. *IEEE Trans. Automat. Contr.* **21**, 108-112.

Whitehead, J. (1986). Supplementary analysis at the conclusion of a sequential clinical trial. *Biometrics* **42**, 461-471.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials.* rev. 2nd edition. Wiley, New York.

Woodroofe, M. (1976). A renewal theorem for curved boundaries and moments of first passage times. *Ann. Probab.* **4**, 67-80.

Woodroofe, M. (1977). Second order approximations for sequential point and interval estimation. *Ann. Statist.* **5**, 984-995.

Woodroofe, M. (1981). A.P.O. rules are asymptotically non-deficient for estimation with squared error loss. *Z. Wahrsch. Verw. Gebiete* **58**, 331-341.

Woodroofe, M. (1986a). Very weak expansions for sequential confidence levels. *Ann. Statist.* **14**, 1049-1067.

Woodroofe, M. (1986b). Asymptotic optimality in sequential interval estimation. *Adv. Appl. Math.* **7**, 70-79.

Woodroofe, M. (1988). Asymptotic expansions for first passage times. *Stoch. Process. Appl.* **28**, 301-315.

Woodroofe, M. (1992). Estimation after sequential testing: A simple approach for a truncated sequential probability ratio test. *Biometrika* **79**, 347-353.

Woodroofe, M. and Coad, D. S. (1997). Corrected confidence sets for sequentially designed experiments. *Statist. Sinica* **1**, 53-74.

Woodroofe, M. and Keener, R. (1987). Asymptotic expansion in boundary crossing problems. *Ann. Probab.* **15**, 102-114.

Yakir, B. (1994). Optimal detection of a change in distribution when the observations form a Markov chain with a finite state space. In *Change-Point Problems* (Edited by E. Carlstein, H. Muller and D. Siegmund), 364-358. Inst. Math. Statist., Hayward, CA.

Yakir, B. (1997). On the distribution of a concomitant statistic in a sequential trial. *Sequential Anal.* **16**, 287-294.

Zhang, C. H. (1988). A nonlinear renewal theory. *Ann. Probab.* **16**, 793-825.

Zhang, Q., Basseville, M. and Benveniste, A. (1994). Early warning of slight changes in systems and plants with application to condition based maintenance. *Automatica* **30**, 95-114.

Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
E-mail: lait@stat.stanford.edu

(Received March 2000; accepted November 2000)

COMMENTS

Herman Chernoff

Harvard University

Sequential Experimental Design

This paper by Professor Lai constitutes a *tour de force*, introducing and describing a vast literature in substantial detail and in so limited a space. One area which I consider slightly neglected in this presentation, because of my own major interest in it, is that of sequential experimental design, and I would like to expand on this.

First, let me describe my general philosophy about Statistics. I believe that Wald's decision theory formulation completed the stages taken by Fisher and Neyman and Pearson to describe the mathematical formulation of problems of statistical inference. It built on the Neyman-Pearson introduction of alternative hypotheses by adding cost considerations. Making alternative hypotheses and cost considerations explicit clarified many of the issues that used to cause confusion. Also, I feel that a statistical problem is not well understood unless one can describe it from the point of view of a Bayesian decision problem. The very formulation of the relevant possible alternatives requires some subjective consideration based on previous experience. Having understood the problem, it is often not necessary to analyze it from a Bayesian or decision theoretic point of view.

The Robbins (1952) formulation of the problem of experimental design introduced the two armed bandit problem and publicized the Robbins-Monro (1951) stochastic approximation method. As Professor Lai points out, the two armed bandit problem raises the issue "Is it ultimately economical to sacrifice, by pulling

the apparently poorer arm, in the hope of thereby getting more useful information?"

To review this special problem, we wish to maximize the expected number of successes with n pulls distributed sequentially among the two arms with specified success probabilities p_1 and p_2, where $p_1 > p_2$ but it is not known which arm belongs to p_1, each arm equally likely to be the better arm. It was widely conjectured that the optimal strategy was to always select the arm for which the posterior probability of being the better arm was greater. This conjecture baffled a substantial number of mathematical statisticians until Feldman (1962) derived a proof in his doctoral dissertation.

In the meantime the Robbins paper illuminated the more general issue of sequential experimentation which is at the very heart of inductive inference and scientific progress. How should one use past experience to help select the next experiment to perform, or to decide to stop experimentation? Previously, sequential analysis problems had been formulated in terms of when to stop repeating a given experiment, and no serious thought was given to deciding among alternative experiments.

For the single experimental scheme applied to estimation, there is relatively little to gain from sophisticated sequential analysis. As Chow and Robbins (1965) and others later pointed out, optimality results there involved higher order improvements over rather natural naive procedures. The Robbins-Monro method represented a significant variation and innovation which has led to an immense literature. But for testing hypotheses, sequential analysis led to substantial gains. Would sequential analysis be equally productive in the presence of a choice of experimentation?

Interestingly enough, the Feldman proof of the conjecture suggested that the two armed bandit problem would *not* serve as a useful example to pursue this question, because it consisted of optimizing by maximizing the short run pay off.

It was this situation that partly was responsible for my attack, Chernoff (1959), on the problem of deciding among a finite number of alternative composite hypotheses when sequential experimentation was allowed and there were several possible "elementary experiments" available to be performed at each stage. My results, complemented by those of some of my students, Bessler (1960a, b) and Albert (1961), clearly demonstrated the role of a minimax game involving Kullback-Leibler information numbers where the experimenter tries to select the experiment to maximize the information, treating nature as an opponent trying to select an alternative to minimize that information. These results were different in that the choice of experiments were not as limited as in the Robbins paper, and the problem attacked was one of discriminating among a finite number of hypotheses. As in the simple testing problem pioneered by Wald, the gains in using sequential methods were substantial first order effects.

Backward Induction

The Arrow, Blackwell, Girshick (1949) approach to the optimizing problem of sequential analysis involved a nontrivial backward induction argument to bypass a measure theoretic difficulty in the original Wald-Wolfowitz (1948) proof. As Professor Lai pointed out, this backward induction approach was the basis for the future development of dynamic programming. In principle the solution of sequential problems in a Bayesian framework involves such a backward induction argument, and an optimal solution cannot bypass such an argument. An interesting aspect of many asymptotic results is that asymptotic optimality is often achieved by methods that seem to ignore this requirement of using backward induction. I would like to offer an explanation.

Consider a sequential problem such as testing whether the mean of a normal distribution with variance one is positive or negative. Suppose that an observation may be taken every hour at a cost of 1 per observation, and the cost of making the wrong decision is equal to the absolute value of the unknown mean which has a normal prior with mean 0 and standard deviation 1 million. The optimal stopping procedure could possibly lead to sampling a long time. Suppose now that the experimenter is told that the experiment is to be truncated, and a decision must be reached by the end of the year. In principle this raises a different backward induction problem with a different stopping rule. On the other hand, with this prior, it is most likely that the unknown mean will be quite large in magnitude, and the experimenter will receive overwhelming evidence in the first few observations, and the effect of a slightly different stopping cost, imposed by truncation at the end of the year, will be practically irrelevant. It is obvious that this type of reasoning is not feasible in all sequential or dynamic programming problems, but in the many where it is relevant, there is room for asymptotically optimal procedures which do not invoke the need to use backward induction calculations explicitly.

Composite Hypotheses

The Kiefer-Weiss (1957) attack on the problem of testing composite hypotheses was innovative but, in my opinion, misguided. As Professor Lai pointed out, they formulated the problem of minimizing the expected sample size for $\theta = \theta_0$ (indifference zone) subject to specified error probabilities at $\theta = \theta_1$ and $\theta = \theta_2$ (alternative hypotheses). This optimization task represents a well defined mathematical problem, but one which does not fit well into a decision theoretic framework. Why should one neglect the costs incurred in the expected sample sizes under the alternative hypotheses? The Schwarz (1962) approach addressed the real problem by attacking a proper decision theoretic problem, incidentally presenting some elegant asymptotic results.

There is a hidden subtext, on which I will not elaborate here. It involves the remarkable Wald-Wolfowitz optimality result that the sequential probability ratio test *simultaneously minimizes both expected sample sizes for given error probabilities*, which would ordinarily have seemed too ambitious a project to attempt.

Clinical Trials and Ethical Problems

As Armitage (1985) pointed out, the objectives of clinical trials are too complex to be easily formulated as a simple sequential problem. Nevertheless, sequential theory represents a useful method of evaluating how well one accomplishes some of the goals of such a trial. One aspect that has evaded proper consideration is the ethical issue.

Given the doctor's ethical requirement to provide what he considers the best possible treatment to his patient, how can he participate in a randomized study where he may be consciously applying a technique that he does not believe to be the best? The medical profession has struggled with this issue and has adopted justifications which I regard as weak rationalizations that do not confront the issue directly. One such has been that as long as we are not sure which is the better treatment we are entitled to treat them as equally good. This argument has been bolstered somewhat by providing informed consent to patients selected to be given the experimental treatment. Petkau and I (1985) once proposed the use of an ethical cost proportional to the current estimate of the difference in expected treatment outcomes. This represents a somewhat naive, but direct attack on the issue.

Group Testing

The increase of computer power has made it much easier to evaluate proposed sequential strategies by direct computing effort. Nevertheless, theory still plays an important role. In situations where there are many factors involved, it is difficult to present the outcomes of the computations on individual strategies in a suitable form for comparison without the dimension reducing effects of relatively simple theoretical results.

As an example consider the problem of testing for the sign of the mean of a normal distribution with known variance. The original problem has five parameters of concern besides the unknown mean. These are the cost of sampling, the (linear) cost of the wrong decision, the variance of the distribution, and the mean and variance of the normal prior distribution. Embedding this problem in the continuous time version involving a Wiener process with unknown drift, Chernoff (1961), provided two major advantages. One could consider the original discrete

time problem as a variation of the continuous time problem where stopping was only allowed at specified times.

One advantage was the ability to use the continuous technology of partial differential equations, and the other was that one could apply several normalizations to reduce the number of effective parameters. The outcome was that the solution could be represented by a single stopping set in a two dimensional graph representing the path of a Wiener process. The five parameters simply determined the starting point of the process.

Incidentally this simplification has relevance to the problem of group testing. In group testing the set of times of possible stopping (measured, for example, in the number of observations between possible stopping times) is further restricted. But there is a simple approximate correction (see Chernoff (1965) and Chernoff and Petkau (1986)) to relate the solution of the continuous time problem to the original discrete time problem. That correction is easily modified for the group testing problem.

With the insights from such theoretical analyses, it is easy to determine plausible strategies worth evaluating further.

Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.
E-mail: chernoff@stat.harvard.edu

COMMENTS

Cheng-Der Fuh and Inchi Hu

Academia Sinica, Taipei and Hong Kong University of Science and Technology

Professor Lai's paper constitutes a comprehensive review of recent developments in sequential analysis and of some challenges and opportunities ahead. The review focuses on several classical problems and new horizons which highlight the interdisciplinary nature of the subject.

In the development of sequential change-point detection in dynamic systems, the window-limited detection rules introduced in this paper are first-order asymptotically optimal (up to order $o(\log \gamma)$ of the average delay to detection). The properties of second-order (up to $O(1)$) asymptotically optimal detection rules can also be considered in the stochastic dynamic system (3.10a, b). By representing the likelihood function of a finite state hidden Markov model as a Markovian

iterated random function, Fuh (2000) established the asymptotic optimality of the SPRT and Cusum in hidden Markov models. It appears that the second-order asymptotic optimality of the detection rules can be investigated along this line.

An idea quite popular in adaptive control and stochastic approximation is the certainty equivalence principle (CEP). The stochastic approximation can be viewed as a control scheme following this principle if one assumes the slope at the regression root is known (cf. Lai and Robbins (1979)). However, strictly applying this principle meets difficulty in establishing consistency of the parameter estimator, which is necessary for the efficiency of control schemes. Some technical conditions need to be imposed for obtaining parameter consistency (cf. Lai and Robbins (1981)). In Hu (1997, 1998), it was shown that if one employs the Bayes estimates in connection with CEP, one can obtain a control scheme with consistent parameter estimates in simple linear regression models. It seems that with more information on the convergence rate, these results can be generalized to nonparametric regression models as in the framework of Robbins and Monroe (1951). It would be interesting to see whether similar results can be obtained under the frequentist setting for general stochastic control models.

The multi-armed bandit problem was originally introduced by Robbins (1952). Lai and Robbins (1985) established a lower bound on the regret (cf. (6.1) and (6.2)) and they called any rule achieving the lower bound an asymptotically efficient control rule. Recently, by adopting the approach of Lai and Robbins, Fuh and Hu (2000) considered a stochastic scheduling problem with order constraints. The scheduling problem is motivated by stratified multistage computerized adaptive tests. They first transformed the problem to an irreversible correlated multi-armed bandit problem with Markovian rewards and then constructed a lower bound of the regret for any uniformly good control rule, characterized by a deterministic constraint minimization problem. In ignorance of the parameter value, they constructed a class of efficient control rules, which achieve the lower bound, based on the theory of sequential testing. They also generalized the results to partial order constraint cases. It is known that the asymptotically efficient control rule approximates the celebrated Gittins' index rule, cf. Chang and Lai (1987) and Brezzi and Lai (2000). The extention of the preceding results to correlated multi-armed bandit problem with partial order constraints seems to be an issue which deserves further investigation.

Concerning the valuation of option price in financial economics, AitSahlia and Lai (1999) applied the theory of optimal stopping to first reduce the problem to a canonical form involving a standard Brownian motion. Employing a numerical method to evaluate the optimal stopping boundary described by (6.10), they provided an approximation of the American option price. This approach is

based on the classical Black-Scholes model, which is inadequate to describe empirical data. A natural extension of this model is the stochastic volatility model or the jump diffusion model. The stochastic volatility model, with its intrinsic model structure, can describes the phenomena of volatility smile; while the jump diffusion model explains this phenomena by incorporating an exogenous jump process into the geometric Brownian motion. An analytic approximation of the American option, barrier option and look-back option to these two models would be an interesting task.

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.
E-mail: stcheng@stat.sinica.edu.tw

Department of Information and Systems Management, Hong Kong University of Science and technology, Clear Water Bay, Kowloon, Hong Kong.
E-mail: imichu@usthk.ust.hk

COMMENTS

B. K. Ghosh

Lehigh University

Professor Lai is an eminent and prolific researcher in sequential statistical analysis and numerous related areas. Over the last thirty years his research has extended from sophisticated asymptotic analysis of sequential methods to their applications in quality control, clinical trials, recursive estimation and, more recently, financial markets. This paper gives a comprehensive review of practically all subfields of sequential analysis and clearly exhibits Professor Lai's command of, and research contributions to, the general subject. The reader gets a clear idea about the nature of the problems, the solutions available at the moment, and what is yet to be done.

I have only one comment, which is not a criticism but rather a point of information. Practically all useful results in sequential analysis since 1970 are asymptotic in nature. It is not at all clear to me how adequate they are, i.e., how close the approximate formulas are to their exact counterparts in various problems of sequential analysis. In fact, a logical topic among Professor Lai's future challenges could be simulation studies and numerical calculations for verifying the adequacy of every asymptotic result. Appleby and Freund (1962) and

Chernoff and Petkau (1986) are two prototypes I have in mind. I like to believe that such undertakings are now feasible in view of recent advances in simulation techniques, numerical analysis, and personal computers. Perhaps Professor Lai can amplify on this in his rejoinder.

Department of Mathematics, Lehigh University, Bethlehem, PA 18015, U.S.A.
E-mail: bkg0@lehigh.edu

COMMENTS

M. G. Gu

The Chinese University of Hong Kong

First I want to thank Professor Lai for a thorough and illuminating review of the subject of sequential analysis. This is a tremendous task since sequential analysis is composed of many different areas of applications with their own subject headings in their respective fields. Maybe the word "analysis" does not serve well for many sequential procedures and turns away potential users of methods contained in this interesting and vast field.

I agree completely that "for sequential analysis to remain a vibrant statistical subject, it has to grow not only inwards ... but also outward through active involvement in the biomedical, socio-economic and engineering sciences". In the following, I point out another possible application for sequential analysis that is related to the computer intensive Markov chain Monte Carlo (MCMC) methods in spatial statistics.

Consider the following problem. Suppose that image x_0 is sampled from the Gibbs distribution $f(x;\theta)$ of the form

$$f(x;\theta) = [c(\theta)]^{-1}\exp[-U(\theta,x)], \tag{1}$$

where $c(\theta)$ is an unknown normalizing factor and $U(\theta,x)$ is a known function. Finding the maximum likelihood estimate of θ based on the observation x_0, one has to maximize the function $\log f(x_0,\theta) = -\log c(\theta) - U(\theta,x_0)$. Equivalently one has to find the zero of the function $h(\theta)$, where

$$h(\theta) = -[\nabla_\theta c(\theta)]/c(\theta) - \nabla_\theta U(\theta,x_0) \tag{2}$$

and ∇_θ is the gradient operator with respect to the argument θ. Using the identity $E_\theta[\nabla_\theta \log f(X,\theta)] = 0$ with X having density f given in (1), we obtain that

$$E_\theta[\nabla_\theta U(\theta, X)] = -[\nabla_\theta c(\theta)]/c(\theta).$$

Therefore, the function h defined in (2) can be approximated by

$$\hat{h}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta U(\theta, X_i) - \nabla_\theta U(\theta, x_0) \tag{3}$$

if $X_1, X_2, \ldots, X_n$ is a sample from the Gibbs density in (1). Based on this and the recent advances in the field of MCMC and stochastic approximation (see Section 5 of the paper for an introduction), a procedure for finding the zero $\hat{\theta}$ of $h(\theta)$ is given in Gu and Zhu (2001). The procedure was applied successfully to find the MLE for some spatial models with moderate dimension of the image vector X.

The procedure is briefly described in the following. Suppose that $\theta^{(k)}$ is the current estimate of $\hat{\theta}$, then we first simulate $X_1, X_2, \ldots, X_n$ from the Gibbs density $f(x; \theta^{(k)})$ (this usually involves iterating the Gibbs sampler or the Metropolis-Hastings algorithm), and then update the parameter by

$$\theta^{(k+1)} = \theta^{(k)} + \lambda_k \hat{h}(\theta^{(k)}), \tag{4}$$

where $\hat{h}(\theta^{(k)})$ is defined by (3) and λ_k is a diminishing (vector) constant of the order $O(1/k^\alpha)$ for some α larger than 0 but not exceeding 1.

Similar to many iterative procedures for application, we have the problem of when to stop iteration in (4) and declare $\theta^{(k)}$ is a good enough estimate of $\hat{\theta}$. At first sight, we might think that, since simulations are done by a computer, one should do as much iteration as possible. But in reality, especially when we deal with larger size problems, we usually do not have a clue as to how many iterations of (4) are good enough. We note that each simulated X_i is a high dimensional vector representing an image which itself can only be simulated through MCMC methods. Our experience is that if the model is complex, such as the Very-Soft-Core model for spatial patterns, a few hours of simulation on a top-of-the-line WorkStation for a medium size problem usually does not produce enough iteration to guarantee convergence. So an automatic convergence criterion-stopping criterion is needed here. This is intrinsically a sequential search problem. There are certain similarities between this problem and the adaptive treatment allocation and the fixed-width confidence interval problems discussed in the paper. We find no standard procedure for the stopping problem in the literature.

The approach adopted in Gu and Zhu (2001) for the stopping criterion is to base it on the relative size of the value $h(\theta^{(k)})$. Monte Carlo error has to

be considered, since the function h is unknown and has to be estimated. For details, we refer the reader to Section 3 of Gu and Zhu (2001). We note that the approach there is only a first attempt to deal with an important and complicated sequential problem.

Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong.
E-mail: minggao@cuhk.edu.hk

COMMENTS

Gary Lorden

Caltech
Pasadena

Professor Lai's survey makes an important contribution to the field of sequential analysis, as we should expect from a scholar whose research contributions over the last three decades of the subject's history have been wide-ranging and masterful.

As an aid to organizing my remarks, I will characterize the elements of what I think a mathematical statistician would ideally like to do in a piece of work in sequential analysis.

1. *Formulation.* Define a statistical problem that is worth solving for the sake of applications, and identify measures of performance that are useful and appropriate for comparing statistical procedures for solving the problem.

2. *Prescription.* Propose a class of statistical procedures to solve the problem.

3. *Performance Description.* Give formulas (or algorithms) for approximating the operating characteristics of the proposed procedures, i.e., their performance as described by the identified measures.

4. *Assessment.* Show that the performance of the proposed procedures is better than other proposals (good), is nearly optimal (better), or is exactly optimal (best).

The canonical example is the work of Wald and Wolfowitz on the Sequential Probability Ratio Test. In response to the needs of certain applications, Wald considered the general problem of testing two simple hypotheses, identifying the error probabilities and the expected sample sizes as measures of performance. He proposed the class of SPRT's with critical values A and B, and gave quite useful approximate formulas for the error probabilities and expected sample sizes in

terms of A and B and the readily computable information numbers. Finally, Wald and Wolfowitz proved that for any choices of A and B, the SPRT is optimal in an appropriate exact sense, somewhat analogous to the Neyman-Pearson optimality of any fixed-sample likelihood ratio test, but even more startling in the fact that an SPRT simultaneously optimizes both expected sample sizes.

How wonderful it would have been if the post-Wald history of research in sequential analysis could have continued the pattern of formulating and solving important problems so completely in one or two papers! But of course, the problems got harder, and the development of new formulations, methods, and mathematical results about them typically stretched over many years and many authors. Still, as Professor Lai skillfully explains, major developments have continued, and fruitful areas of application have been found and exploited, both as inspiration for new research and as sources of new "customers" for sequential analysis.

By way of echoing Lai's description of the progress over the last fifty years, I would assert that it teaches important lessons about the roles of the four elements in my idealized list, especially the lesson that progress in the first three elements has had larger impact than progress in assessment, even though the ideal of optimality (or "statistical efficiency"), and the mathematical challenges of demonstrating it, continue to have fundamental appeal and have inspired much fruitful effort. As one example, two of the three post-Wald papers that inspired me to write my mathematics Ph.D. thesis in the field of sequential analysis — Chernoff's (1959) paper laying the foundations of asymptotic optimality theory for sequential testing and design of experiments, and Kiefer and Sacks' (1963) quite general formulation of first-order asymptotic optimality theorems in the same vein, do not even appear in Lai's bibliography! (The third paper, Schwarz's (1962) beautiful work on "asymptotic shapes", does appear.) This observation is intended not at all as a criticism of Professor Lai's choices, but as evidence for my assertion that optimality theory has not been very influential in the development of sequential analysis in the last forty years.

Another case in point concerns the Kiefer-Weiss problem. As Lai's discussion makes clear, this problem — to minimize the *maximum* expected sample size rather than the two expected sample sizes the SPRT optimizes, was a natural outgrowth of the realization that the SPRT minimizes the expected sample size when the parameter θ is far from the "worst case". In that sense it represented progress in the element of "formulation", and as Lai explains, my proposal of the 2-SPRT (Lorden (1976)), sharpened previous work of Anderson (1960). The 2-SPRT was shown to minimize the expected sample size at any given θ to within $o(1)$ as the error probabilities go to zero. For the well-known problem of testing the mean of a normal distribution with equal error probability specifications at

$\theta = \pm\delta$, considerations of symmetry reduce the problem of attaining minimax expected sample size to this "modified" Kiefer-Weiss problem of minimizing for $\theta = 0$. Without any reliance on symmetry, my student Huffman (1983) showed how to construct 2-SPRT's for testing in one-dimensional exponential families so as to attain the minimax expected sample size to within $o(\sqrt{EN})$; and Dragalin and Novikov (1987) refined Huffman's results to achieve the minimax expected sample size to within O(1). Since their work includes useful approximations to expected sample sizes, it would seem that the chain of work from Kiefer and Weiss to Dragalin and Novikov completes my "ideal" of progress from formulation through prescription to performance description and proof of (near) optimality. Why, then, isn't this work well-known and widely used? I submit that it is largely because, as Lai effectively points out in his Section 2, there exist generally preferable prescriptions like Lai's sequential versions of the Generalized Likelihood Ratio (GLR) test, which (at least for large samples) "self-tune" to approximately minimize the expected sample size at whatever is the true θ! Thus, the problem of "nearly minimizing" the expected sample size for every θ is simply a better formulation than the Kiefer-Weiss problem, and statistical procedures which perform reasonably well in that context (first exemplified in Schwarz (1962)) are simply more attractive than even exact Kiefer-Weiss solutions would be.

The same theme regarding the importance of formulating the right problem and proposing procedures that address it appears in Lai's excellent discussion of sequential analysis for clinical trials. He makes an effective case for the view that progress in the development of influential sequential methods for these problems depends on finding formulations and prescriptions that address the complex needs of practitioners. The cited work of Gu and Lai (1998, 1999) appears to contain the right elements for maximizing the impact and usefulness of sequential analysis in this field: formulating problems with flexible features and an array of performance measures and descriptions that address the multiple needs of users. Moreover, their performance descriptions have two aspects — a "theoretical" method that gives reasonable approximations and aids understanding, as well as a computer program that can make a variety of calculations (and simulations) to accurately portray the consequences of choosing different features of the statistical procedure.

The richness and scope of Lai's description of the past, present, and future of sequential analysis will undoubtedly stimulate much discussion and will also, I expect, encourage an acceleration of progress and stimulate increased interest in this branch of statistics. In an effort not to abandon totally the ideal of brevity, I will limit my discussion at this point to the following additional remarks.

1. For the "internal" growth of the field of sequential analysis, I agree that strengthening ties with other branches of probability and statistics is critical. An

exciting example is the recent progress (Chuang and Lai (1998, 2000a, 2000b)) in the exploitation of bootstrap methods (and modifications called "hybrid" methods) to solve important problems of constructing confidence intervals after termination of a group sequential trial. Another example is the development of what might be called "semi-Bayes" formulations, e.g., the problem of minimizing a given weighted average of the expected sample sizes of a test, subject to the classical bounds on the probabilities of Type I and Type II error.

2. In the twenty-first century the impact of readily available computer power on problem formulations and proposals of sequential methods is likely to accelerate. As the popularity of the bootstrap and other computer-intensive methods illustrates, the amount of computation that it is feasible to perform in applying a statistical procedure to data is astronomically larger than it was during the "adolescence" of sequential analysis. Moreover, to meet the need to give performance descriptions of sequential procedures, it should be anticipated that relatively simple computer calculations, particularly intelligently designed simulations, will supplant and perhaps replace many of the intricate formulas that have been developed to yield performance descriptions, e.g., those that make corrections for "excess over the boundary".

3. Given the exciting and colorful panorama of progress sketched by Professor Lai, the question that intrudes on one's sense of optimism is — why isn't sequential analysis more popular? It continues to be something of a "niche" subject within the field of statistics, not well-represented in the usual statistics texts and courses, with an often frustrating slowness of "technology transfer" from researchers to practitioners. There are many explanatory variables – the subject is hard, understanding it requires delicate probability theory and a high level of mathematical sophistication, using it in practice requires a high level of discipline, etc. But one factor has always tended to limit its appeal: one-at-a-time sampling is in many applications impractical or at least unattractive. Sampling in multiple stages — particularly in two or three stages (as in Stein (1945) and Hall (1981))– can be an appealing compromise between fixed-sample procedures and fully sequential procedures, with nearly as good performance as the latter, provided that the sample size in each stage can be chosen flexibly to depend on the results of previous stages. (Unfortunately, this is not typically the case in clinical trials.) In the context of asymptotic theory of hypothesis tests, Lorden (1983) suggests that three stages are sufficient to equal the performance of fully sequential tests (to first order), but for practical "small sample" problems, it's not clear how to prescribe three-stage procedures whose error probabilities and expected sample sizes can be readily approximated and whose performance is highly efficient.

Department of Mathematics, California Institute of Technology, Pasadena, CA 91125, U.S.A.
E-mail: gary_lorden@caltech.edu

COMMENTS

Adam T. Martinsek

University of Illinois

I would like to congratulate Professor Lai on his excellent presentation of the history and future prospects of sequential analysis. He skillfully interweaves the evolution of key methods with practical considerations in the fields of medicine, engineering and economics that influenced the evolution. Both the work described and the description are therefore interdisciplinary in the truest sense of the word.

I agree completely with Professor Lai's statement "For sequential analysis to remain a vibrant statistical subject, it has to grow not only inwards in the form of further methodological advances and breakthroughs but also outwards through active involvement in the biomedical, socio-economic and engineering sciences". In my comments I will focus on several recent developments in the area of sequential estimation that follow this model, either because they arose from actual applied problems or because they are clearly applicable to other disciplines. These developments fall under three general headings: sequential estimation in logistic regression and generalized linear models, sequential density estimation, and sequential estimation of the maximum and the mean in models for bounded, dependent data. Research under the first two headings is relevant to interdisciplinary work in medicine and engineering. The work described under the third heading has its origin in an engineering problem and is also potentially useful in economics.

In medical applications it is often important to model the probability of getting a disease as a function of a vector of covariates $\mathbf{X}$. A model that has proved enormously useful over the years is the logistic regression model, for which

$$P[Y_i = 1|\mathbf{X_i}] = \frac{\exp(\mathbf{X_i}^T\beta)}{1 + \exp(\mathbf{X_i}^T\beta)},$$

where $Y_i = 1$ if the ith patient has the disease and 0 otherwise, $\mathbf{X_i}$ is a p-dimensional vector of covariates thought to influence the incidence of the disease, and β is an unknown p-dimensional vector of parameters.

The model is especially relevant in observational studies, such as cohort studies, and in many situations it is important to estimate the unknown parameter vector β very accurately. One approach is to construct a confidence ellipsoid for β that is sufficiently small, e.g., for which the longest axis has length no greater than a prescribed upper bound. This problem was first considered by Grambsch

(1989). She proposed a stopping rule to determine the sample size, very much in the spirit of the pioneering work of Chow and Robbins (1965) and subsequent work on linear regression models by Gleser (1965), Albert (1969), and Srivastava (1967, 1971). She showed that the resulting procedure is asymptotically efficient in terms of the almost sure behavior of the sample size and that it achieves the desired accuracy and confidence level.

One drawback of Grambsch's procedure is that it relies on knowledge of the distribution of the covariate vectors $\mathbf{X_i}$. Chang and Martinsek (1992) proposed a different procedure that is nonparametric (or least semiparametric) in that it does not require knowledge of the distribution of the $\mathbf{X_i}$. They showed that the highly desirable behavior obtained by Grambsch (1989), namely almost sure sample size efficiency along with specified accuracy and confidence level, carries over to the nonparametric version of the stopping rule. They also showed that the expected sample size is asymptotically optimal. Chang (1995, 1996) extended these results substantially beyond the logistic regression case, to general link functions F for which $P[Y_i = 1|\mathbf{X_i}] = F(\mathbf{X_i}^T\beta)$.

Sequential nonparametric estimation of a probability density function at a point has been considered by Yamato (1971), Carroll (1976), Isogai (1981, 1987, 1988) and Stute (1983). In applications to engineering it is of interest to estimate the entire density function, and Martinsek (1992) addressed the problem of estimating the density with sufficient global precision, i.e., with mean integrated squared error (MISE) that is smaller than a specified bound. The Parzen-Rosenblatt kernel estimate was considered a stopping rule and was formulated based on an asymptotically optimal sequence of bandwidths and the well-known expansion

$$\int_{-\infty}^{\infty} (\hat{f}_n(x) - f(x))^2 dx$$
$$= (nh_n)^{-1} \int_{-\infty}^{\infty} K^2(t)dt + (k_2^2 h_n^4/4) \int_{-\infty}^{\infty} (f''(x))^2 dx + o((nh_n)^{-1} + h_n^4)$$

for the MISE, where K is the kernel, h_n is the bandwidth, and $k_2 = \int_{-\infty}^{\infty} t^2 K(t)dt$. The resulting sequential procedure was shown to achieve the desired bound on the MISE and to do so using a sample size that is first order optimal both almost surely and in mean. Analogous results for L_1 error, as opposed to mean integrated squared error, have been obtained by Kundu and Martinsek (1997).

A related global density estimation problem is to construct a fixed width confidence band for an unknown density f on a finite interval, as opposed to merely providing a fixed width confidence interval for the value of the density at a single point. Simultaneous accuracy is especially important when the density estimate will be used for classification, e.g., classification of patients as having

heart disease or not based on heart rate data (Izenman (1991)). Misclassification can have serious consequences, and accurate classification is difficult to achieve without an accurate estimate of the density.

Xu and Martinsek (1995), using results of Bickel and Rosenblatt (1973) on maximal deviations of kernel density estimates, proposed a stopping rule to bound the maximum width of the confidence band by a specified positive number ϵ, while achieving a desired confidence coefficient $1-\alpha$ for the entire band. The user is then guaranteed that all values of the unknown density have been determined to within $\pm\epsilon$, with simultaneous confidence $1-\alpha$. Xu and Martinsek showed that the resulting procedure achieves the desired accuracy and confidence level asymptotically as the upper bound ϵ on the maximum width of the confidence band approaches zero, i.e., when one requires a high degree of accuracy in estimating the density. Martinsek and Xu (1996) obtained similar results for censored data.

Sequential estimation of the maximum and the mean for bounded, dependent data arises naturally in problems involving monitoring of underground pipelines for corrosion. A typical approach in such problems is to send a remote-controlled sensor into the pipeline. As the sensor moves, it uses sound waves to measure the thickness of the wall at a series of locations. The original wall thickness at the time of manufacture is known. Let X_i denote the pit depth (original wall thickness minus current thickness) recorded at the *ith* location. The X_i have support $[a,b]$, where a and b are unknown. It is important to estimate b with a high degree of accuracy, as the estimate will be used to decide whether or not to replace the pipeline, and the consequences of an incorrect decision are serious. Imprecise estimates may result in needless and expensive replacement of a pipeline that is still in good shape, or else leaving in place a pipeline that presents a significant threat of near-term leakage. Similarly, it is essential to obtain an accurate estimate of the mean pit depth in the pipeline, as it gives an overall measure of the extent of corrosion. Sampling efficiency, i.e., producing an accurate estimate with as few observations as possible, is also important when estimating these parameters: running the sensor involves significant expense and one would like to minimize this expense, subject to the accuracy requirement.

Because the X_i are bounded, it is natural to model them by

$$X_i = (b-a)Y_i + a, \tag{1}$$

where the Y_i have a beta density

$$f_{\alpha,\beta}(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \tag{2}$$

for $0<y<1$, with α and β positive but unknown. This is a fairly flexible family for modeling the marginal distribution of the pit depths: as α and β vary, a

wide variety of density shapes and endpoint behavior can be achieved. Because the observations X_i are clearly positively correlated rather than independent, we assume further that the Y_i satisfy the following random coefficient autoregressive model, introduced by McKenzie (1985):

$$Y_n = 1 - U_n(1 - W_n Y_{n-1}), \tag{3}$$

where Y_0 has density (2), U_n and W_n are independent sequences of independent, identically distributed (i.i.d.) random variables, U_n has a beta distribution with parameters β and $\alpha - p$, W_n has a beta distribution with parameters p and $\alpha - p$, and p is positive and strictly less than α. Then (3) defines a strictly stationary sequence with marginal density given by (2) and autocorrelation function $\rho(k) = \rho^k$ for $k = 0, 1, \ldots$, where

$$\rho = E(U_n)E(W_n) = p\beta/\alpha(\alpha + \beta - p). \tag{4}$$

For fixed α and β, (4) is increasing in p, and by varying p one can achieve any positive correlation ρ. McKenzie (1985) refers to this model as a PBAR model, where P denotes positive correlation. Note that the correlation structure of the observable X_i's will be the same as that of the unobservable Y_i's.

Using results from strong mixing and extreme value theory, Martinsek (2000a) shows that under the model given by (1)-(3), the maximum observation $M_n = max(X_1, \ldots, X_n)$ has the following limit distribution as $n \to \infty$:

$$P\Big[\frac{n^{1/\beta}(M_n - b)}{(b-a)[\beta/C(\alpha,\beta)]^{1/\beta}} \le y\Big] \to e^{-(-y)^\beta} \tag{5}$$

for $y < 0$, where $C(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$, i.e., M_n has a Type III limit distribution. We would like to construct a confidence interval for b whose width is at most $2d$ for a prespecified $d > 0$, and whose confidence level is approximately $1 - \gamma$, where $\gamma \in (0, 1)$ is also prespecified. Based on (5), Martinsek (2000a) defines a stopping rule T_d by

$$T_d = \text{ first } n \ge 2 \text{ such that}$$
$$(M_n - m_n)[\frac{\hat\beta_n}{C(\hat\alpha_n, \hat\beta_n)}]^{1/\hat\beta_n}[(-\log(\gamma_1))^{1/\hat\beta_n} - (-\log(\gamma_2))^{1/\hat\beta_n}] + n^{-1} \le 2dn^{1/\hat\beta_n},$$

where $0 < \gamma_1 < \gamma_2 \le 1$ satisfy $\gamma_2 - \gamma_1 = 1 - \gamma$, $m_n = min(X_1, \ldots, X_n)$, and $\hat\alpha_n$ and $\hat\beta_n$ are suitable estimates of α and β, respectively. Then the confidence interval

$$I_{T_d} = [M_{T_d} + T_d{}^{-1/\hat\beta_{T_d}}(M_{T_d} - m_{T_d})[\hat\beta_{T_d}/C(\hat\alpha_{T_d}, \hat\beta_{T_d})]^{1/\hat\beta_{T_d}}(-\log(\gamma_2))^{1/\hat\beta_{T_d}},$$
$$M_{T_d} + T_d{}^{-1/\hat\beta_{T_d}}(M_{T_d} - m_{T_d})[\hat\beta_{T_d}/C(\hat\alpha_{T_d}, \hat\beta_{T_d})]^{1/\hat\beta_{T_d}}(-\log(\gamma_1))^{1/\hat\beta_{T_d}}]$$

has width at most $2d$. As $d \to 0$, $P[b \in I_{T_d}] \to 1-\gamma$, i.e., the confidence level of the procedure is close to the target value when one requires a high degree of accuracy. Moreover, for

$$n_d^* = (2d)^{-\beta}(b-a)^{\beta}[\beta/C(\alpha,\beta)][(-\log(\gamma_1))^{1/\beta} - (-\log(\gamma_2))^{1/\beta}]^{\beta},$$

we have $T_d/n_d^* \to 1$ a.s. Rounding n_d^* to the nearest integer yields the smallest nonrandom sample size that, if it were known, would provide the desired accuracy and confidence level, asymptotically. Thus the stopping rule T_d is asymptotically efficient almost surely in the sense that it is equivalent to the ideal but unavailable n_d^*.

Because the parameter a is of interest in other applications of the model given by (1)-(3), e.g., to economic data such as the market share of a product or the unemployment rate, Martinsek (2000a) also considers sequential estimation of the minimum, as well as joint estimation of the maximum and minimum. Work addressing accurate estimation of the mean of the model appears in Martinsek (2000b). Shibata (1996) provides a nice survey of a variety of previous statistical and probabilistic approaches to corrosion.

Department of Statistics, University of Illinois, Champaign, IL 61820, U.S.A.
E-mail: martins@stat.uiuc.edu

COMMENTS

Moshe Pollak

Hebrew University

Professor Lai should be thanked for a very interesting and informative paper. It provides the reader with a wealth of knowledge regarding the state of the art of sequential analysis, as well as an ordered (sequential!) description of the evolution of its various successes.

The saying goes: "Necessity is the mother of invention". As is clear from Professor Lai's presentation, sequential analysis was born because it was necessary. However, although the subject has been alive and kicking for six decades, a nonnegligible part of its development seems to have been born out of the minds of theoreticians: the problems are mathematically interesting, but their application in real life seems to be meager (e.g., risk-bounded estimation, multi-hypotheses testing). One of the reasons mentioned in Professor Lai's article for this dearth

of application is that sometimes (as in medicine) the maximal potential number of observations is so small that one does not want to try to diminish it (for reasons of power). However, there is another side to the phenomenon: necessity is not only the mother of invention, but also the mother of application. In many areas, observations are abundant and cheap. Sequential analysis requires formidable mathematics, and the investment required to apply it is heavy, as it cannot be applied by a layman (the usual case is that one must hire a statistician to explain it and to apply it), and even run-of-the-mill statisticians balk at learning the subject. Consequently, most people will try to do without sequential analysis unless absolutely necessary. (Proof? Even today, Shewhart is the most popular form of control chart, and one can even find papers studying its properties in contexts where it clearly shouldn't be applied.) Also, the history of application of sequential analysis does not go hand-in-hand with its theoretical evolution. Quite a few of the methods proposed over the years were originally ad hoc procedures; some of them were later proven to possess optimality properties (e.g., Wald's SPRT or Page's or cusum), others remain popular even though sub-optimal (e.g., Shewhart and EWMA control charts).

In parallel to the story of the problems, I would like to see a delineation of the techniques characteristic of constructions and proofs in sequential analysis. For example, finding a martingale within a problem's structure can do wonders (especially in sequential analysis): the fact that a sequence of likelihood ratios is a martingale (even when the observations are dependent) plays a major role in the evaluation of operating characteristics of tests of hypotheses (c.f. Robbins and Siegmund (1970, 1973)); other martingales play similar roles in other contexts (e.g., Pollak (1987), Novikov (1990)). Change-of-measure considerations are a powerful tool: Wald (1947) used a transformation from H_0-measure to H_1-measure to evaluate the power and expected sample size of sequential tests of hypotheses; Yakir (1995) developed a special change-of-measure technique which is powerful for assessing probabilities (and other functionals) associated with random fields; even simulations can benefit from changes of measure (c.f. Siegmund (1975)).

Regarding the various methods and methodologies described in Professor Lai's paper, I would alter some of its emphasis. For example, I would put more weight on Wald's "weight functions" (referred to in the paper as mixture rules) and other methods of estimation (such as martingale-preserving methods of estimation, c.f. Robbins and Siegmund (1973), Dragalin (1997)). Such methods can be shown to have very strong ("second order") optimality properties (which the GLR methods emphasized in Professor Lai's paper may also possess, but proof as of now is lacking).

Finally, to the list of challenging problems in sequential analysis I would add those which deal with robustness of the methods. Some work on this has

been done: a few sequential rank methods (even for non-contiguous alternatives; e.g., Savage and Sethuraman (1966), Sethuraman (1970), McDonald (1990), Gordon and Pollak (1995)) have been developed, and some work in the line of Huber's approach to robustness has been done (e.g., Quang (1985)). But all in all, this is one of the underdeveloped areas of sequential analysis which deserve attention.

Department of Statistics, Hebrew University, Jerusalem 91905, Israel.
E-mail: msmp@mscc.huji.ac.il

COMMENTS

David Siegmund

Stanford University

Professor Lai has given us an authoritative review of both the theoretical advances in sequential analysis since its beginnings in the 1930's and '40's, and its applications in medicine, engineering and economics. My own immersion in sequential analysis began in graduate school and lasted for roughly three decades. For me the appeal of the subject is the combination of challenging theory and interesting applications along with a natural focus on certain issues at the foundation of statistical inference. Following are some idiosyncratic thoughts that occurred as I read this stimulating paper.

Sequential analysis presents difficult technical problems, which can be so fascinating that they take on lives of their own. On occasion solutions to these problems expand our conceptual horizons. An outstanding example is the beautiful proof of the optimality of the sequential probability ratio test, where a formal Bayes optimization problem is introduced as a *deus ex machina* in order to prove an optimality property that is stated entirely in terms of error probabilities and expected sample sizes. The papers by Wald and Wolfowitz (1948) and Arrow, Blackwell and Girshick (1949) spawned the field of dynamic programming/stochastic control theory.

Sequential analysis was closely associated with statistical decision theory, also developed by Wald; and its applications to quality control seem well suited to a decision theoretic framework. Applications to clinical trials have been most successful when approached from the more flexible inferential viewpoint of Armitage's *Sequential Clinical Trials* (Armitage (1975)). Anscombe's (1963) extended review of the first edition of Armitage's book failed in its primary goal

of changing the *Weltanschauung* of clinical trials from an inferential to a decision making viewpoint, although it did provide a class of challenging theoretical problems, reviewed here by Professor Lai. (Anscombe wrote as a Bayesian, but I find this aspect of his paper much less significant than its decision making orientation.)

Sequential analysis has also provided grist for philosophical ruminations about the foundations of statistical inference, for example by illustrating the effect that uncritical adherence to the likelihood principle might have on "sampling to a forgone conclusion." In this regard I find particularly interesting the paper of Cornfield (1966), who wrote as a Bayesian adherent of the likelihood principle, but with an inferential viewpoint that at the end of the day produced a test very much in tune with Armitage's repeated significance tests, which by their focus on Type I and Type II errors are philosophically, if not practically, at odds with the likelihood principle.

Other problems suggested by applications to clinical trials have been solved by systematic progress over a number of years. A not uncommon sentiment of several decades ago was, "You can't use sequential methods in clinical trials because $\cdots$" (fill in the blanks) (a) we cannot afford the unbounded sample size of a sequential probability ratio test, (b) truncated sequential tests are less powerful than comparable fixed sample tests, (c) we don't know how to estimate a treatment effect if we use a sequential test, (d) we don't know how to deal with censored survival data if we use a sequential test. Professor Lai shows us that these problems have by now been successfully addressed.

Continuing challenges in the design and analysis of sequential clinical trials are problems of multiple endpoints (e.g., Lin (1991)) and comparisons of more than two treatments or multiple treatment levels (e.g., Lin and Liu (1992), Siegmund (1993)), where one may want to eliminate treatments during the course of the trial.

Professor Lai also shows us an example in option pricing where methods originally developed to solve problems of sequential analysis can be applied to solve new problems. Change-point detection, which was originally studied in a sequential formulation and applied to quality control, provides another example where methods developed in sequential analysis have proved useful in a broader domain. An example of personal interest of the technological "spin off" of these ideas is found in the statistical theory of genetic mapping, or linkage analysis. The primary goal of genetic mapping is to identify certain regions of the genome as the location of the genes controlling particular traits, e.g., inherited diseases in humans. The standard paradigm of (parametric) linkage analysis was handed down, as if from Mount Sinai, by Morton (1955), who was very much influenced by Wald's ideas of sequential testing. Although the sequential aspect, at least

with regard to sample size determination, was probably never more than a conceptual possibility, Morton's arguments, which were built on Wald's analysis of the error probabilities of the sequential probability ratio test, have dominated the field until relatively recently. Spurred by the recognition that Morton's approach depends on strong modeling assumptions and by the new possibilities opened up by the abundance of molecular genetic markers now available, a different approach has developed, which emphasizes natural connections to (fixed sample) change-point problems. For example, the Type I error probability of a genome scan can be approximated (Feingold, Brown and Siegmund (1993), Lander and Kruglyak (1995)) by adapting methods developed during the 1970's in sequential analysis (e.g., Woodroofe (1976b)). Genetic mapping also poses incompletely understood problems of sequential design, where one must decide how many markers to use in a preliminary genome scan, then how to increase the marker density for a closer examination of genomic regions where there is some evidence of linkage.

Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.
E-mail: dos@stat.stanford.edu

COMMENTS

Gordon Simons

University of North Carolina

I am impressed by the quality and breadth of Lai's presentation but remain less than fully convinced of the paper's thesis. Professor Lai is to be congratulated for his splendid exposition of major themes in sequential analysis – past and present. And perhaps of the future. I shall be teaching a course in sequential analysis this spring and will undoubtedly find his description of recent developments a most helpful reference source.

But to the question: *does sequential analysis have a promising future?* Perhaps this question needs to be raised within a broader context. There are many mathematicians today – we certainly have them here at the University of North Carolina – who wonder aloud, but not too publicly, whether pure mathematics has a promising future. Closer to home, a professionally active, distinguished colleague poses this question to me on an irregular basis: does (mathematical)

statistics have a future? Though, to be honest, his musings seem primarily concerned with future attitudes 50 and 100 years from now, when statistical theory might be of historical interest only. But still, his pessimism on this matter reflects his personal assessment of current trends in the subject.

Perhaps this is a normal concern of senior academics as they are closing their careers and the baton is being passed on to a new generation. Clearly, subject areas do not remain static – even if they continue to develop. In this regard, it is interesting happened in the last couple of decades within the realm of Bayesian statistics. While modern Bayesians, when challenged, are quick to pay lip service to the brilliant foundational work of Leonard Savage – in defense of genuinely *subjective* priors – these same modern Bayesians are busily working away on statistical applications using priors (and hyperpriors) of convenience, with little or no regard for issues of subjectivity. Two weeks ago I attended an interesting talk at Duke concerned with *objective* considerations in choosing a prior – with not a single apology offered to Savage!

In his last section entitled **Future Opportunities and Challenges**, Professor Lai writes: "Sequential analysis has been developing steadily but at a somewhat uneven pace during the past six decades. There is now a rich arsenal of techniques and concepts, methods and theories, that will provide a strong foundation for further advances and breakthroughs. The subject is still vibrant after six decades of continual development, with many important unsolved problems and with new interesting problems brought in from other fields."

Despite my earlier remarks, I readily concede that Lai has made a persuasive case. But is this enough? Academic subjects operate within a *culture*, and it must be said that the present culture is very different from the one that spawned the development of sequential analysis, a culture which was warm and receptive to the seminal ideas of Wald and Wolfowitz and Stein, receptive to the brilliant conceptual innovations of a Herbert Robbins, and to the probabilistic prowess of a Y. S. Chow and a Francis Anscombe, receptive to the mathematical elegance of a Wassily Hoeffding, and receptive to the path-breaking analytical innovations of a Herman Chernoff, a David Siegmund and a Michael Woodroofe, and, of course, of a Tze Leung Lai – and receptive to the contributions of a host of others.

Whether or not sequential analysis will see a successful seventh and eighth decades will depend on its ability to adapt to, without being consumed by, a culture that seriously undervalues mathematical analysis in favor of computational facility.

In conclusion, I remain more than a little bit skeptical concerning the future of sequential analysis.

Department of Statistics, University of North Carolina, Chapel Hill, NC 27599, U.S.A.
E-mail: simons@stat.unc.edu

COMMENTS

Bruce W. Turnbull

Cornell University

Although he claims to be giving only a "brief review", Professor Lai is to be commended for his insightful and comprehensive tour of the state of the art of sequential statistical analysis. Indeed it has been Professor Lai himself who has contributed much to the advancement of this area of statistical research over the past thirty years. As the author states, it is indeed a "vibrant" field as witnessed by the many new articles, books and computer software that continue to appear, and the success of the specialized journal *Sequential Analysis.* This vibrancy is not extraordinary — we live in a world that is inherently dynamic, not static. Situations are constantly changing and our statistical toolbox must reflect this reality. Here I will comment on just a few applications where there may be "new challenges", particularly in the field of clinical biostatistics.

First we consider a major "success" of sequential statistical methods – that is the design and monitoring of clinical trials. Initially, Armitage (1975) had proposed the application of sequential analysis to clinical trials; however, it was not until the mid 1980's, when formal DSMB's became established, that use of group sequential methods became widely recognized. In the U.S., the Food and Drug Administration (FDA) included requirements of interim analyses in published regulations for Phase III trials in 1985 and 1988. Most recently the FDA's 1998 publication E9 (prepared under the auspices of the International Conference on Harmonization) gives detailed recommendations on all aspects of statistical principles and methodology, including trial monitoring, interim analysis, sample size adjustment and DSMBs. My own thinking on the role of stopping rules in Phase III trials has come full circle. Initially, following the pioneering work of Armitage (1975) and others, the emphasis was on formal stopping rules. Later, as Meier (1975) and others pointed out, the decision to stop a trial is a highly complex one involving many factors, both qualitative and quantitative. This led to the idea of stopping *guidelines,* not stopping *rules* (DeMets (1984)), and the search for more flexible methods. These included repeated confidence intervals and stochastic curtailment — the latter proving especially useful in evaluating the futility of continuing a trial. However, having served on some DSMBs, I have recently seen occasions where members of the DSMB — clinicians, ethicists and patient advocates — were particularly unsure of whether to proceed further with a trial and seemed to be looking to the statisticians on the panel for a formal, objective rule to help them out of their dilemma.

One area where flexibility is needed is in the timing of interim analysis. It is natural to schedule the analyses dynamically, using accumulating results. For example, Lan and DeMets (1989) suggested that a DSMB may decide to switch from group sequential to continuous monitoring if it appears that a testing boundary is likely to be crossed soon. Betensky (1998) has addressed some of computational problems involved. However Proschan, Follmann and Waclawiw (1992) have documented how properties of statistical tests can break down if data dependent timing of analyses is employed. This is a "new challenge" area where further research could prove fruitful.

Section 3 of Professor Lai's paper concerns change-point detection. His main application is to the area of quality control and engineering. However another major application is to disease diagnosis. One area of sustained interest has been the optimum timing of screening tests both in individuals and in populations. This is currently the subject of special effort by the U.S. National Cancer Institute's CISNET program for breast, prostate and colon cancer. Optimal intervals between tests should be tailored to the individual's risk factors and accumulating information, but also incorporating knowledge from existing data bases. These screening intervals are of interest not only to guide physicians' recommendations but also to health insurance companies and HMOs who need to make economic decisions.

Serial biomarkers are becoming increasingly used to monitor individuals for onset of a particular disease of interest — PSA for prostate cancer, or CD4 count for onset of AIDS, for example. Even though for each study subject, the series of biomarker readings is typically quite short, irregularly spaced and subject to error, it is possible to use sequential change-point detection techniques, where the change-point is defined as onset of the disease of interest. For example, Slate and Turnbull (2000) have described two methods. The first uses a fully Bayesian hierarchical model for a mixed effects segmented regression model. Posterior estimates of the change-point distribution in a given individual are obtained by Gibbs sampling, these estimates featuring the "borrowing of strength" from other subjects in the study population, usual in such an approach. The second uses a hidden stochastic process model in which the onset time distribution is instead estimated by maximum likelihood via the EM algorithm. It can be viewed as an empirical Bayes version of the first approach. Both methods lead to a dynamic index that represents a strength of evidence that onset has occurred by the current time in a given subject. A question remains on how to evaluate such diagnostic rules. Tests given at a single time are typically evaluated using ROC curves that plot sensitivity and specificity. For dynamic index rules based on accruing longitudinal data, Slate and Turnbull (2000) proposed a three-dimensional generalization of the ROC curve, but it seems that further research needs to be done in this area.

I find the last paragraph of Section 6 overly pessimistic about the applicability of multi-armed bandit type ideas in biomedical trials. When there are nuisance parameters or where variances depend on means, adaptive designs can lead to a reduced total sample size. This class of designs is not only valuable in the hypothesis testing framework but also for point and interval estimation (Hayre and Turnbull (1981a, b)). When there are three or more treatments and the goal is to select the best one, adaptive allocation can again lead to reduced total sample size — Jennison, Johnstone and Turnbull (1982), Coad (1995). In fact there have been a few clinical trials where a response-adaptive design has been used, notably the ECMO trials reported by Ware (1989) and two trials of anti-depression drugs sponsored by Eli Lilly and Co. (Tamura et al. (1994); Andersen (1996)), the latter being an interesting example of a four-armed adaptive trial. An important consideration in the utility of an adaptive design is that it should be able to be incorporated easily into a familiar group sequential design. Jennison and Turnbull (2000, Chapter 17), describe two-armed group sequential designs in which the allocation fraction by which incoming subjects are randomized to one treatment or the other is different in each stage and depends on responses from previous stages. An asymptotically optimal rule, which can also accommodate nuisance parameters, can be simply described. Because the summary statistics and stopping boundaries are the same as those used in a standard group sequential test, the outlook for the application of such rules in clinical trials may not be as bleak as Professor Lai suggests.

The stochastic approximation ideas discussed in Section 5 are key to the dose finding experimental designs in Phase I clinical trials. As such, they play an extremely important role in the drug development process. However, it seems that the classic methods of Robbins and Monro (1951) and subsequent variations are only rarely used in practice. Rather, more *ad hoc* methods such as the "up-and-down" method (Dixon and Mood (1948), Wetherill (1963)) or the "continual reassessment (CRM) method" suggested by O'Quigley, Pepe and Fisher (1990) and modifications (Faries (1994), Goodman, Zahurak and Piantadosi (1995)) are more commonly applied. Indeed the CRM method has been extended to allow incorporation of auxiliary information such as pharmacokinetic measurements (Piantadosi and Liu (1996)). Use of such information has the potential to increase the efficiency of any stochastic approximation procedure. Thus I feel there are still "challenges" to be found in this area — ones that could have important practical impact in biopharmaceutical research and development.

Suppose we consider the stochastic approximation problem of Section 5, but now impose the constraint that the dose levels $x_1 \leq x_2 \leq \cdots$ are nondecreasing. Thus there is more emphasis on caution early on to avoid "overshooting". This problem arises in animal carcinogenicity studies to find the median (say) onset

time for an occult tumor in a group of mice. The presence of such tumors can only be ascertained upon death of the animal via necropsy. Here $x_1 \leq x_2 \leq \cdots$ correspond to sacrifice times which can be chosen sequentially. However, if too many animals are sacrificed too early before any have tumors, there is little information; and similarly if a majority of the animals are sacrificed too late when almost all possess the tumors. This is the same as the problem in quality control of estimating the reliability of an item when failure can only be ascertained by means of a quantal response at a destructive life test. The addition of this simple monotonicity constraint seems to radically change the nature of the problem and make it much more difficult. Various authors have attempted to attack this problem — Bergman and Turnbull (1983), Louis and Orav (1985), Turnbull and Hayter (1985), Morris (1987), Hu and Wei (1989). However, as yet, there is no established practical solution.

Another example of a very important biostatistical problem where practice does not quite meet theory is that of internal pilot studies or sample size reestimaton. Such designs are often needed where there is an unknown nuisance parameter without knowledge of which it is impossible to design a clinical trial with a desired prespecified power. There is a growing literature on the subject — see Jennison and Turnbull (2000, Chap.14) for a survey. Yet, there has been little theoretical advance since the original paper on two-stage designs by Stein (1945), as described by Professor Lai in Section 4.1. However, the Stein design has proved inefficient and impractical so other more pragmatic solutions have been sought. Often numerical methods are employed. For example, Denne and Jennison (2000) have investigated multi-stage Stein-type procedures, which have the advantage of being quite efficient and incorporate a familiar feature of group sequential monitoring. The importance of the area suggests that more research is needed here also.

So what of the future of sequential statistical methodology? Currently much statistical research is turning towards the analysis of enormous databases — "data mining". Vast improvements in computing power, algorithms and information technology are rapidly becoming available to all. Thus it is essential that statisticians address how sequential decision making can be made practical in the face of a bombardment of accruing information. This might be to aid disease prognosis and diagnosis in an individual patient when 200 biomarkers are being serially monitored, to control inline a manufacturing production line when over 400 quality characteristics are recorded on each item produced, or to make investment decisions when thousands of financial time series are being observed. This is clearly a most challenging "challenge"!

School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853, U.S.A.

E-mail: bruce@orie.cornell.edu

COMMENTS

C. Z. Wei

Academia Sinica, Taipei

This is a long-awaited paper. Professor Lai, an important leader in the field of sequential analysis, should be thanked for his willingness to share his view of the development of this vibrant subject. The directions and challenges pointed out in the paper are definitely going to stimulate a lot of future research. The following discussions are meant to serve as an annotation to some of the issues described in Lai's paper.

1. Recursive Estimation and Adaptive Control

Adaptive prediction is a step between recursive estimation and adaptive control, albeit it is of interest in its own right (cf. Goodwin and Sin (1984)). For the stochastic regression model (4.11), one may use a recursive estimator $\tilde{\theta}_{n-1}$ to construct a predictor $\tilde{\theta}_{n-1}x_n$. The certainty equivalence principle leads one to investigate the performance of the predictor through the study of the consistency of the estimator $\tilde{\theta}_{n-1}$. Recent results of Hu (1997) and Chen and Hu (1998) on the Bayes control problem is in this spirit.

However, the Bayes theory developed so far seems difficult to apply to the dynamic system. The approach developed in Wei (1987b) directly handled the accumulated prediction error squares $\sum_{k=1}^{n}(y_k - \tilde{\theta}_{k-1}x_k)^2$ without recourse to consistency. It had been used to help establish the logarithm law of the least squares self- tuning controller (Guo (1994)) under a non-Baysian setting. A parallel theory for the Baysian problem is of interest.

2. Stochastic Approximation and Sequential Optimization

The multivariate Robbins-Monro procedure given in Wei (1987) can be viewed as an asymptotically efficient procedure that achieves specific means of several characteristics simultaneously. In the field of quality technology, Taguchi's philosophy emphasizes the design that obtains the target mean value while minimizing the variance. For the past ten years, a dual response surface approach has been developed to achieve Taguchi's goal. (cf. Vining and Myers (1990); see also Fan (2000) for more recent references).

However, a stochastic approximation procedure that achieves the same goal remains to be developed. It seems that a hybrid of the Robbins-Monro procedure and the Kiefer-Wolfowitz procedure would suffice. But the setting of this problem

needs to be explored. The priority of estimating variance first, as advocated in the Taguchi method, still requires a satisfactory theory.

Furthermore, in contrast with the "target is the best" case as described above, one may consider "the larger (smaller) the better" one. In this situation, the mean response is maximized (minimized) while controlling the variance at a specific level. It is not clear what is the effect of the roles of the mean and variance on the associated procedures.

3. Adaptive Treatment Allocation and the Multi-armed Bandit Problem

As described in Lai's paper, the multi-armed bandit problem has a fundamental conflict between "information" and "control". To learn sufficient information from all populations $\{\pi_j, j = 1, \ldots, k\}$ about their parameter values, the allocation (control) rule is allowed to sample from any population irrespective of the choice of ϕ_t. The "information learning flow" among all populations is therefore completely connected. From this point of view, it seems that the results in the classical bandit problem can be extended to the case where the unknown parameters $\{\theta_j\}$ do not necessarily belong to the same set.

Motivated by serial sacrifice experiment problems, a variant of the multi-armed bandit problem was proposed by Hu and Wei (1989). In this problem, the information learning flow has a strict order which forces a constraint $\phi_t \leq \phi_{t+1}$ on the allocation rules. These rules are irreversible. Due to this nature, these rules are only applicable to the experiments which are specified by distributions with one parameter.

The research on irreversible adaptive allocation rules is still underdeveloped. A recent result due to Fuh and Hu (2000) relaxes the i.i.d. sampling mechanism to being Markov. Their motivating example is computerized adaptive testing. In destructive experiments, as in the study of estimating the due date of a film, multi- parameter problems are often encountered. The multi-phase problems in which certain proportions of the samples have to be saved for the later phases are also interesting.

In practice, one may require rules which are simple (and possibly sub-optimal) but easy to implement. For theory, one may consider the problem with general structure on the information learning flow. These problems seem to be of interest too.

Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan.
E-mail: czw@stat.sinica.edu.tw

COMMENTS

Michael Woodroofe

University of Michigan

I want to thank Professor Lai for summarizing his contributions to sequential analysis along with those of his students and many others, and for setting these contributions in historical perspective. This review will be a valuable resource for current and future workers in the area. In the following comments, I advocate some particular ways of formulating problems.

Group Sequential Tests. I certainly agree that it was not feasible to arrange for continuous examination of the data as they accumulate in double blind clinical trials when group sequential tests were introduced, about 25 years ago. Is this still the case? Certainly data that are recorded at any site can simultaneously be recorded at a central processing site that can be completely separated from the treatment of patients. It should be possible for the central site to monitor the data frequently, for example daily or weekly, computing current likelihood functions and/or test statistics. At the very least, this frequent monitoring could be used to trigger a meeting of an oversight committee. Implementing such a scheme would be difficult, both technically and administratively, and the reduced expected sample size may only be worth the effort in cases where the outcomes are literally life and death. In such cases, however, statisticians should advocate efficient designs.

The Change Problem. Wu, Woodroofe and Mentz (2000) have identified a connection between change point problems and isotonic methods, as described by Robertson, Wright and Dykstra (1988). Suppose that there are observations of the form $X_k = \mu_k + Y_k$, where Y_k is a stationary process that exhibits suitable short range dependence and μ_k is known to be non-decreasing. Consider the null hypothesis $H_0 : \mu_1 = \cdots = \mu_n$, where n is the horizon. The alternative here allows an arbitrary non-decreasing trend and, therefore, differs from that of the change point problem which allows only a single change. For this problem, Wu, Woodroofe and Mentz (2000) suggest the test statistic

$$\Lambda_n = \frac{1}{\hat{\sigma}_n^2} \sum_{k=1}^{n} (\hat{\mu}_{n,k} - \bar{X}_n)^2,$$

where $\hat{\mu}_{n,k}$ are the isotonic estimators of μ_k, penalized as in Woodroofe and Sun (1993), and $\hat{\sigma}_n$ is an estimate of scale. Under assumptions like (3.14) of the paper,

they obtain the limiting null distribution of Λ_n and the distribution under local alternatives.

Very Weak Expansions. Very weak expansions agree with Edgeworth expansions in many cases, especially if the latter hold uniformly on compacts in the parameter and the coefficients of $n^{-\frac{1}{2}r}$ are equicontinuous, where n is either the sample size or a design parameter, like a. There are many problems in which the coefficients are not equicontinuous, and especially problems in which they oscillate wildly as n increases. Brown, Cai and Das Gupta (2000) have recently considered cases like this involving discrete exponential families and provide informative graphs. When Edgeworth expansions contain small oscillations, is it really necessary for the actual coverage probabilities to exceed or equal the nominal value (or the nominal value $+o(1/n)$) for all θ; or is it enough for the nominal value to pass through the middle of the oscillations over small intervals or regions that contain many oscillations? The conventional definition of confidence requires the actual to be at least the nominal. Very weak expansions come down squarely in the middle.

Bayesian Bandit Problems. In the context of Bayesian bandit problems, Equation (6.9) of the paper, results of Woodroofe (1979) and Sarkar (1991) suggest that the nature of the problem may change radically in the presence of a suitable covariate. They show that the myopic procedure, which treats each subject as is currently thought to be best for him or her, is asymptotically optimal to second order (minimizes the asymptotic regret). That is (under Sarkar's conditions), the fundamental conflict between learning and immediate gain disappears; and if subjects represent patients that have to be assigned to a standard or experimental treatment, then the ethical problems disappear. The exact meaning of "suitable" is complicated; it requires that there be enough variability in the covariate and also that regression lines (or curves) not be parallel. To illustrate, consider one of Sarkar's examples in which there are two treatments, labeled 0 and 1, and responses are either success or failure, labeled 1 and 0. Suppose that for each subject, there is a covariate X and potential responses Y^i to treatment i, $i = 0$, 1, that X has a known distribution F over [0,1], and that $P_\theta[Y^0 = 1|X = x] = x$, $P_\theta[Y^1 = 1|X = x] = \theta$, where θ is an unknown parameter. Thus, X is the probability of success with treatment 0, assumed known. Finally suppose that θ has bounded density ξ over [0,1]. The *myopic procedure* assigns the kth subject to treatment 0 or 1 accordingly as $E_\xi[\theta|X_k, \mathcal{F}_{k-1}] <$ or $> X_k$ where $\mathcal{F}_k$ is the sigma-algebra generated by the first k covariates and responses, and E_ξ denotes expectation in the Bayesian model. If F has a positive, bounded density f, then the myopic procedure is asymptotically optimal to second order (minimizes the asymptotic regret).

Acknowledgements

Thanks to Bob Keener for helpful conversations. Research supported by the National Security Agency and U. S. Army.

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.
E-mail: michaelw@umich.edu

COMMENTS

Benjamin Yakir

Hebrew University

Introduction

The paper of T. L. Lai summarizes recent developments in several topics related to sequential analysis. Undoubtedly, the field of sequential statistics looks very different today than it did only a decade ago. The author's impressive contribution to the field has much to do with this fact. Naturally, the choice of topics and results is biased towards those which involve the author's contributions. I will take the same liberty and remark only on the topic of optimality in the context of sequential change-point detection.

The two main approaches for proving optimality rely either on decision theory, or on the construction of a lower bound and the demonstration that a given procedure attains the lower bound (or at least approaches it). The first approach has been successfully applied in the setting of independent observation and simple hypothesis. In more complex situations, only the second approach seems to be working. Indeed, using inequality (3.8) of his paper, Lai was able to demonstrate the asymptotic optimality for detecting changes in the important setting of state-space models.

It should be noted, however, that inequality (3.8) provides only a first-order asymptotic lower bound. Unfortunately, the resolution of first-order asymptotics is very limited. Indeed, in this resolution one cannot distinguish between the performance of most reasonable candidate procedures (like the cusum, Shiryayev-Roberts, EWMA, etc.). Only when a higher-order asymptotic expansion is considered can we hope to be able to rank the different procedures.

In these notes we attempt to extend Lai's results and in the direction of addressing this drawback. Theorem 1 of the next section formulate an alternative lower bound, motivated by the known optimality properties of the Shiryayev-Roberts procedure. This lower bound is demonstrated in Section 3 in the simple setting of a shift of a normal mean. The proof of the theorem is given in an appendix.

A lower bound on the expected delay in detection

We start with some notations: Let $X_1, X_2, \ldots,$ be a random sequence. Denote by P the *in-control* distribution on the sequence. The distribution of the process may change at some unknown point in time k, $0 \le k < t$. Denote by P_k the distribution when a change occurs at time k. (The measure P_k is a change-point measure in the sense that the marginal joint distribution of $X_1, \ldots, X_n$ is identical both under P and under P_k for all $n \le k$. In other words, the first k observations are from the in-control distribution. However, the observations that follow — $X_{k+1}, X_{k+2} \ldots$ — are from the new regime.) We denote by $\ell_k(n)$ the log-likelihood ratio of P_k, relative to the null distribution P. This log-likelihood is based on the first n observations.

Denote,

$$R(n) = \sum_{k=0}^{n} e^{\ell_k(n)}, S = \sum_{n=0}^{t-1} \sum_{k=0}^{n} e^{\ell_k(n)} = \sum_{n=0}^{t-1} R(n), \text{ and } M = \max_{0 \le n < t} R(n).$$

Given $a > 0$, consider the Shiryayev-Roberts stopping time $N_a = \min\{n : R(n) \ge e^a\}$. In particular, note that $\{N_a < t\} = \{M \ge e^a\}$.

Let N be any stopping time N (with respect to the sequence of observations) that satisfies the constraint:

$$\mathrm{P}(N \ge t) \ge \mathrm{P}(N_a \ge t). \tag{1}$$

We are interested in comparing the properties of N as a change-point detection procedure to those of N_a. One can use, for example, the average delay in detection over the interval $[0, t)$, $\frac{1}{t}\sum_{k=0}^{t-1} \mathrm{E}_k(N-k; N \ge k)$, as a measure of the efficiency of N. In Theorem 1 below, the comparison is carried out in terms of a lower bound on the difference of efficiency between the two procedures. The lower bound in Theorem 1 is independent of N. It depends only on the properties of S, M, N_a, and the sequence $\{R(n) : 0 \le n < t\}$.

Theorem 1. *Let N be any stopping time which satisfies the constraint* (1) *with respect to the Shiryayev-Roberts stopping rule N_a. Then*

$$\frac{1}{t}\sum_{k=0}^{t-1} \mathrm{E}_k(N-k; N \ge k)$$

$$\geq \frac{1}{t}\sum_{k=0}^{t-1} \mathrm{E}_k(N_a - k; N_a \geq k)$$
$$-\frac{1}{t}\Big\{\mathrm{E}\,(S; S \geq e^a; M < e^a) - \mathrm{E}\,(S; e^a \leq S < e^{a_s}) + \sum_{i=1}^{3} r_i\Big\},$$

where $r_1 = \sum_{k=0}^{t-1}\sum_{n=t}^{\infty} \mathrm{P}_k(N_a > n) \leq \sum_{k=0}^{t-1}\sum_{n=t}^{\infty} \mathrm{P}_k(\ell_k(n) < a)$, $r_2 = e^{-a}$ $\sum_{n=0}^{t-1} \mathrm{E}\,(R^2(n); R(n) < e^a)$, $r_3 = e^{-a}\sum_{n=0}^{t-1}(t-n)\mathrm{E}\,(R(n); R(n) < e^a)$, *and* a_s *is the solution of the equation*

$$\mathrm{P}(S \geq e^{a_s}) = \mathrm{P}(N_a < t) = \mathrm{P}(M \geq e^a). \tag{2}$$

Quantification of the terms appearing in the lower bound can lead to statements of optimality. An example is given in the next section. The proof of the theorem is given in an appendix.

An optimality result in terms of the average delay in detection, subject to a constraint on the probability of false alarm, can be translated to other forms of optimality. The translation to a Bayesian formulation with the uniform prior is straightforward. The consideration of other priors requires only minor modifications.

Moreover even the traditional formulation, which measures the efficiency in terms of the worst (conditional) delay in detection and puts a constraint on the expected run length to false alarm can be handled. For example, one can reformulate the constraint (1) in terms of not stopping in the interval $(jt, (j+1)t]$, given that no stopping occurred prior to jt — a hazard rate type of condition. The relation between the expectation and the hazard rate can be combined with the (asymptotic) quasi-stationarity of the Shiryayev-Roberts procedure in order to prove the asymptotic optimality of N_a. Indeed, we believe that using this approach may potentially yield refined optimality results in complex models like the state-space models discussed in Lai. We have applied it, for example, in the context of the detection of a change of the slope in a regression model (Yakir (2001)).

An example: detecting a shift in a normal mean

Let us demonstrate Theorem 1 in the simple case of a normal shift of the mean. For this case it is known that for some formulation the cusum, and for other formulations the Shiryayev-Roberts or a variant thereof, are optimal in a strict sense (see the references in Lai's paper, and also Yakir (1997b)). Still, one better try one's new machinery in a familiar environment before trying to expand into a new territory.

We specify here the random sequence to be a sequence of independent normal random variables with variance 1. The mean prior to the change-point is 0, and the mean thereafter is $\mu > 0$. The log-likelihood ratios, $\{\ell_k(n) : 0 \le k \le n < t\}$, are normally distributed. Moreover, it is useful to note that in this case these log-likelihood ratios are also log-likelihood ratios for the complete sequence of observation. The measure for which $\ell_k(n)$ is a likelihood ratio, denoted here by $\mathrm{P}_{k,n}$, assigns mean μ to the observations $X_{k+1}, \ldots, X_n$. The observation up to time k and the observation past the time n are assigned a zero mean.

A measure transformation technique or the more traditional renewal theory can be used in order to bound the tail of the statistics S, M, and $R(n)$. (More details on the measure transformation technique, with application in various fields, can be found in Yakir and Pollak (1998); Siegmund (1999); Siegmund and Yakir (2000, 2000b). This bound, together with some straightforward derivations can be used in order to show that $r_i \to 0$, $i = 1, 2, 3$ (provided $a \to \infty$ and $a \ll t \ll e^{a/2}$).

Next, since S now is a sum of likelihood ratios, one can rewrite the remaining lower bound terms in the form:

$$\mathrm{E}\,(S; S \ge e^a; M < e^a) = \sum_{n=0}^{t-1} \sum_{k=0}^{n} \mathrm{P}_{k,n}(S \ge e^a; M < e^a), \text{ and}$$
$$\mathrm{E}\,(S; e^a \le S < e^{a_s}) = \sum_{n=0}^{t-1} \sum_{k=0}^{n} \mathrm{P}_{k,n}(e^a \le S < e^{a_s}).$$

Define $s_{k,n} = \log(S/e^{\ell_k(n)})$, $m_{k,n} = \log(M/e^{\ell_k(n)})$ and $\mu_{k,n} = \mathrm{E}_{k,n}\ell_k(n) = (n-k)\mu^2/2$. The $\mathrm{P}_{k,n}$-asymptotic independence between $\ell_k(n)$ and $(s_{k,n}, m_{k,n})$ can be used in order to show that

$$\begin{aligned}\mathrm{P}_{k,n}(S \ge e^a; M < e^a) &= \mathrm{P}_{k,n}(a - s_{k,n} \le \ell_k(n) < a - m_{k,n}) \\ &\sim \frac{1}{(2a)^{1/2}} \phi\left(\frac{a - \mu_{k,n}}{(2a)^{1/2}}\right) \mathrm{E}_{k,n}(s_{k,n} - m_{k,n}).\end{aligned}$$

This approximation is valid whenever $\mu_{k,n} \in a \pm \mathrm{o}\,(a)$, the relevant region of indices k and n. Similarly,

$$\begin{aligned}\mathrm{P}_{k,n}(e^a \le S < e^{a_s}) &= \mathrm{P}_{k,n}(a - s_{k,n} \le \ell_k(n) < a_s - s_{k,n}) \\ &\sim \frac{1}{(2a)^{1/2}} \phi\left(\frac{a - \mu_{k,n}}{(2a)^{1/2}}\right) (a_s - a).\end{aligned}$$

The terms $\mathrm{E}_{k,n}(s_{k,n} - m_{k,n})$ converge, as $n - k \to \infty$, to a constant we denote by $-\mathrm{E}_{\mu}[\log(M/S)]$. The limit is independent of k and n and can be associated

with the additive constant in the expansion of the average run length of the power-one SPRT under the alternative. Likewise, the difference $a_s - a$ converges to a term we denote by $-\log(\mathrm{E}_\mu[M/S])$. This limit doesn't depend on both k and n as well, and is associated with the overshoot correction of the significance level of the SPRT.

By summing over k and n, $0 \le k \le n < t$, we can conclude that

$$\frac{1}{t}\sum_{k=0}^{t-1} \mathrm{E}_k(N-k, N \ge k) \ge \frac{1}{t}\sum_{k=0}^{t-1} \mathrm{E}_k(N_a - k, N_a \ge k) - \frac{1}{\mu^2}\{\log(\mathrm{E}_\mu[M/S]) - \mathrm{E}_\mu[\log(M/S)]\} + \mathrm{o}(1).$$

The asymptotic lower bound, as a function of the mean μ, is presented in Figure 1. Note that as the mean increases, the bound becomes tighter. Only when the mean is below 0.4 do we get that the gap between the efficiency of the Siryayev-Roberts procedure and the given bound is less then -1.

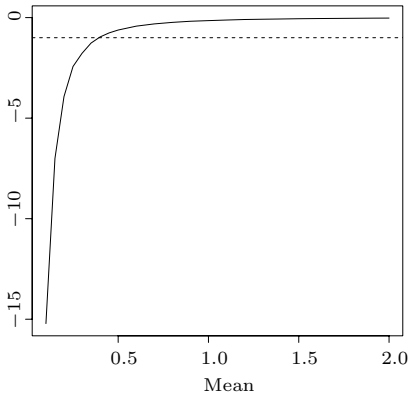

Figure 1. The asymptotic lower bound.

Acknowledgement

The author was supported by the US-Israel Bi-national Science Foundation Grant #9700214.

Appendix. A proof of Theorem 1.

Note that $\mathrm{E}_k(N-k, N \ge k) = \sum_{n=k}^{\infty} \mathrm{P}_k(N > n) \ge \sum_{n=k}^{t-1} \mathrm{P}_k(N > n)$. But, $\mathrm{P}_k(N > n) = \mathrm{E}\,(e^{\ell_k(n)}; N > n)$, by the likelihood-ratio identity, since the event $\{N > n\}$ is determined by $X_1, \ldots, X_n$. The random variables $e^{\ell_k(n)}$ are

non-negative and $\{N \geq t\} \subset \{N > n\}$. Thus,

$$\sum_{k=0}^{t-1} \mathrm{E}_k(N-k; N \geq k) \geq \sum_{k=0}^{t-1} \sum_{n=k}^{t-1} \mathrm{E}\,(e^{\ell_k(n)}; N > n) \geq \mathrm{E}\,(S; N \geq t).$$

The fact that $\mathrm{E}\,(Y; C) \geq \mathrm{E}\,(Y; Y < y)$ for any non-negative random variable Y, event C and real number y, such that $\mathrm{P}(C) \geq \mathrm{P}(Y < y)$, can be used in order to conclude that

$$\frac{1}{t} \sum_{k=0}^{t-1} \mathrm{E}_k(N-k; N \geq k) \geq \mathrm{E}\,(S/t; S < e^{a_s}). \tag{3}$$

Next we consider the average delay in detection of the Shiryayev-Roberts stopping time N_a. As in the previous derivation,

$$\begin{aligned}\sum_{k=0}^{t-1} \mathrm{E}_k(N_a-k; N_a \geq k) &= \sum_{k=0}^{t-1} \sum_{n=k}^{\infty} \mathrm{P}_k(N_a > n) \\ &= \sum_{k=0}^{t-1} \sum_{n=k}^{t-1} \mathrm{E}\,(e^{\ell_k(n)}; N_a > n) + \sum_{k=0}^{t-1} \sum_{n=t}^{\infty} \mathrm{P}_k(N_a > n) \\ &= \sum_{n=0}^{t-1} \mathrm{E}\,(R(n); N_a > n) + r_1.\end{aligned}$$

However, $\mathrm{E}\,(R(n); N_a > n) = \mathrm{E}\,(R(n); N_a \geq t) + \mathrm{E}\,(R(n); n < N_a < t)$. Conditioning on $X_1, \ldots, X_n$, using Doob's inequality and the definition of N_a we get that $\mathrm{E}\,(R(n); n < N_a < t) \leq e^{-a}\{\mathrm{E}\,(R^2(n); R(n) < e^a) + (t-n)\mathrm{E}\,(R(n); R(n) < e^a)\}$. When we sum over all the n's we get

$$\begin{aligned}\sum_{n=0}^{t-1} \mathrm{E}\,(R(n); N_a > n) \leq \mathrm{E}\,(S; N_a \geq t) &+ e^{-a} \sum_{n=0}^{t-1} \mathrm{E}\,(R^2(n); R(n) < e^a) \\ &+ e^{-a} \sum_{n=0}^{t-1} (t-n)\mathrm{E}\,(R(n); R(n) < e^a)].\end{aligned}$$

The above discussion can be summarized by the statement

$$\frac{1}{t} \sum_{k=0}^{t-1} \mathrm{E}_k(N_a-k; N_a \geq k) \leq \mathrm{E}\,(S/t; M < e^a) + \frac{r_1 + r_2 + r_3}{t}. \tag{4}$$

The inequalities (4) and (3), and the fact that $a_s > a$, lead to the theorem.

Department of Statistics, Hebrew University, Jerusalem 91905, Israel.
E-mail: msby@mscc.huji.ac.il

REJOINDER

Tze Leung Lai

I wish to thank all discussants for their inspiring, insightful and informative contributions. In the rejoinder I will address some of the issues they have raised and also use this opportunity to elaborate on several topics that I had left out of the paper to preserve its smooth flow and to keep it within manageable length.

1. Sequential Experimental Design

The paper considered sequential experimentation along the lines pioneered by Robbins (1952), namely stochastic approximation and multi-armed bandits, for which experimentation is targeted toward and constrained by certain control objectives, but does not treat more traditional experimentation goals. Professor Chernoff's discussion has filled this gap for the case where the goal of experimentation is to discriminate among a finite number of hypotheses and the experimenter can choose among a set of experiments. Excellent introductions to the topic and additional references can be found in Chernoff (1972, 1975). Another important case is optimal experimental design in nonlinear regression. Although classical optimal design theory for least squares estimation in linear regression models can in principle be extended to nonlinear regression models of the form $y_n = f(x_n;\theta) + \epsilon_n$ (using the same notation as in (5.1) with $M(x) = f(x;\theta)$), the extension has serious practical difficulties since an optimal design measure typically involves the unknown parameter θ. To circumvent these difficulties, a commonly used approach is to construct designs sequentially, using observations made to date to estimate θ, and choosing the next design point by replacing the unknown θ in the optimal design with its estimate (cf. Federov (1972)). Thus sequential experimentation has played a central role in experimental designs for nonlinear regression models since the seminal paper of Box and Wilson (1951) on response surface methodology.

Unlike classical experimental design theory where the design levels are nonrandom constants, the design levels x_n in a sequential design are sequentially determined random variables, and it remained a long-standing problem concerning whether inference based on the repeated-sampling principle is asymptotically valid for such sequential designs. For linear regression models, a positive answer was provided by Wu (1985) by making use of the asymptotic theory of (linear) least squares estimates in stochastic regression models (Lai and Wei (1982)). Lai (1994), and recently Skouras (2000), extended that theory to nonlinear least squares estimates, thereby providing a positive answer for sequential designs in nonlinear regression models. In the case of linear regression models with normal

errors, Woodroofe (1989) and Woodroofe and Coad (1997) have developed "very weak" asymptotic expansions of the sampling distributions to refine the normal approximations for least squares estimates in sequentially designed experiments.

2. Multivariate Observations, Multiple Endpoints and Multiple Tests

Professors Lorden, Pollak and Simons have given various reasons why sequential analysis has not been as "popular" as we think it should be. I will give a few additional reasons and offer some suggestions. One important reason, in my opinion, is that a large portion of the sequential analysis literature has been narrowly focused on very simple models, in which the observations are i.i.d. univariate random variables having some density function f_θ with unknown parameter θ belonging to some simple set (e.g., finite, or a subset of the real line). Although this narrow focus provides a good starting point, since simplicity enables one to see the central issues clearly without being sidetracked into complicated technical details that are "non-sequential" in essence, it is too restrictive for a "final product" to attract the interest of researchers not working in sequential analysis. We should go far beyond the relatively simple settings in which fundamental innovations have been made, and follow up on the initial breakthroughs with further advances to make the methodology widely applicable. To accomplish this, interactions with other areas of statistics and probability, and with substantive fields of application, are of vital importance.

A case in point is the interface of sequential analysis with multivariate statistics. The need to extend sequential analysis, from the simple univariate settings considered by Wald to multivariate problems with nuisance parameters, became apparent soon after Wald's introduction of the SPRT and, in the fifties and sixties, there was considerable effort to develop sequential tests for multivariate data and general linear hypotheses; see Ghosh (1970). As pointed out in Section 2 of the paper, most of this work involved the use of invariance to reduce composite hypotheses to simple ones so that Wald's SPRT could be applied to the maximal invariants, but this approach only applies to a restrictive class of composite hypotheses. Moreover, except for Wald's approximations (ignoring overshoots) to the type I error probabilities, the operating characteristics and optimality properties, if any, of these invariant SPRTs remained an open problem until the mid-seventies and early eighties; see Lai (1981). The insistence on using likelihood ratios (via invariance or weight functions) also made it difficult to relate to concurrent developments in fixed sample size tests, for which GLR statistics had been found to provide flexible and highly efficient tests for multivariate data and in multiparameter problems with nuisance parameters; see e.g., Neyman (1965) and Bahadur (1967).

One reason why GLR statistics were not used in developing sequential tests during this period is that it was not clear how the stopping boundaries should be set for GLR statistics. Section 2 of the paper reviews several developments to resolve this problem, first for one-parameter and eventually for multiparameter exponential families. There is now a comprehensive theory involving stopping rules of the type (2.4). Another reason why GLR statistics were not used in developing sequential tests during the fifties and sixties is that Wald's method to analyze error probabilities of the SPRT can no longer be applied when the likelihood ratios are replaced by GLRs. To get around this difficulty, Lai and Siegmund (1977) and Woodroofe (1978) made use of nonlinear renewal theory to develop asymptotic approximations to the type I error probabilities of sequential GLR tests. A review of subsequent developments was given by Hu (1988), who also extended the "backward method" of Siegmund (1985) to multiparameter exponential families. Recently Chan and Lai (2000c) developed a general method to derive asymptotic approximations to the error probabilities (*both* type I and type II) that can be applied not only to GLR statistics but also to other functions of the sufficient statistics in a multiparameter exponential family. Moreover, the method is applicable to both sequential and fixed sample size tests. In the fixed sample size case, following the seminal work of Chernoff (1952), Hoeffding (1965) and Bahadur (1967) on asymptotic efficiencies of tests at non-local alternatives, most papers on large deviation approximations to type I and type II error probabilities give only the order of magnitude of the logarithms of the probabilities. Chan and Lai (2000c) have refined these approximations to the form $(C+o(1))e^{-\rho n}$ and have further improved them to higher-order approximations for moderate deviation probabilities (when the alternatives approach the null hypothesis as $n \to \infty$). In the sequential case, they have also shown how the method can be used to develop approximations to the error probabilities of sequential GLR tests when the underlying parametric family is misspecified.

Therefore, it is now ripe for a much better interface of sequential analysis with multivariate statistics. There are many important problems in sequential analysis that are of a multivariate nature, and conversely there are many exciting new directions in multivariate analysis that involve sequential ideas. Concerning the former, Professor Siegmund has pointed out active areas of research in group sequential methods for clinical trials with multiple endpoints and for trials with several arms. Beginning with O'Brien's (1984) seminal paper, the problem of constructing one-sided tests for comparing multivariate treatment effects has received much attention for fixed sample size tests, and subsequently for group sequential trials; see Lin (1991), Jennison and Turnbull (1993, 2000) and the references therein. Bloch, Lai and Tubert-Bitter (2000) recently gave a new formulation of the multiple-endpoint problem, motivated by certain studies in

arthritis, rheumatism and cancer, in which treatment effects are measured by both efficacy and toxicity. In these studies, efficacy is often measured by more than one response variable and so is toxicity. Univariate methods for assessing each response variable individually have been widely used in these studies because of their flexibility and ease of interpretation. On the other hand, usually there is additional need for a single, overall comparison, and combining the univariate comparisons by Bonferroni's inequality ignores the correlations between the response variables and therefore lacks power for alternatives at which the response variables are strongly correlated. In the fixed sample size case, Bloch, Lai and Tubert-Bitter (2000) developed a bootstrap test that incorporates the essential univariate and multivariate features of the treatment effects to be compared. Extending it to the group sequential setting will entail modifying the hybrid resampling approach described in Section 4.2 for this problem.

While the multiple endpoint problem is concerned with testing a single null hypothesis on multivariate treatment effects, the multi-hypothesis (or multiple) testing problem is concerned with testing k hypotheses on multidimensional parameters. Contrary to what Professor Pollak says, the "application in real life" of multi-hypothesis testing is far from being "meager". In classical (nonsequential) multivariate analysis, it arises e.g., in multiple comparisons (developed by Tukey, Scheffé, Dunnett and others) and classification. Using a sequential (stepdown) approach and Bonferroni's inequality, Holm (1979) introduced a multiple test procedure that controls the so-called "family-wise type I error rate" (FWE). The procedure rejects the hypotheses sequentially one at a time until no further rejections can occur, and is therefore called *sequentially rejective*. After Holm's seminal paper, a number of variants of sequentially rejective procedures appeared, including the step-up procedures of Hommel (1988) and Hochberg (1988), and there was a resurgence of interest in multiple tests, partly spurred by applications to psychology and educational testing. Shaffer (1995) gave a survey of these developments and their applications. Recently there have been exciting applications to microarray analysis in genetics which also call for new techniques in multiple testing; see Dudoit, Yang, Callow and Speed (2000). Sequential analysis ideas and related boundary crossing probability approximations may provide important advances in this area. Professor Simons says, "Clearly, subject areas do not remain static – even if they continue to develop." Multiple testing, which is not much younger than sequential analysis, is another example besides the Bayesian methodology he cites.

3. Sequential Analysis and Time Series

The assumption of i.i.d. observations in traditional sequential analysis is too restrictive for many applications in engineering and economics. As pointed out in

the last paragraph of Section 5, stochastic approximation underwent a separate development in the engineering literature where, unlike static regression functions in the statistics literature, dynamic input-output systems are considered. This was also the motivation behind Anantharam, Varaiya and Walrand (1987) in extending the work of Lai and Robbins (1985) on the multi-armed bandit problem from i.i.d. to Markovian arms. Motivated by queuing control and applications to machine learning, Lai and Yakowitz (1995) developed a nonparametric bandit theory that makes no parametric assumptions on the underlying dynamics of the individual arms and yet can still attain a regret of the order $O(\alpha_n \log n)$, where α_n is any nondecreasing sequence of positive numbers such that $\lim_{n\to\infty} \alpha_n = \infty$.

To expand its scope and increase its impact, sequential analysis should have more interactions with time series analysis. These two areas of statistics are actually closely related, particularly if we broadly interpret sequential analysis as statistical methods (including design and decisions) for data that arrive sequentially. In Lai (1981, 1995, 1998, 2000), it is shown how lower bounds on the expected sample size/expected detection delay can be developed for sequential testing/changepoint detection procedures subject to type I error/false detection constraints and how asymptotically efficient procedures can be constructed to attain these lower bounds for general stochastic sequences (including i.i.d. sequences as special cases). The method of Chan and Lai (2000c) described above can be extended to derive asymptotic approximations to error probabilities of sequential tests when the observations are Markovian (instead of i.i.d.). Details of the extension are given in Chan and Lai (2000a,b) for the considerably more difficult problem of sequential change-point detection. Melfi (1992) extended nonlinear renewal theory while Fuh and Lai (1998) extended Wald's equation and the Wiener-Hopf factorization to Markov random walks. Sriram (1987) generalized Robbins' (1959) theory of asymptotically risk-efficient estimation of means from i.i.d. to stationary AR(1) models, while Fakhre-Zakeri and Lee (1992) further generalized that to linear processes. Lai (1996) subsequently provided a general theory of asymptotically risk-efficient estimation of the parameters of stochastic systems satisfying certain conditions. Professor Martinsek's discussion describes recent interesting applications of fixed width confidence intervals in random coefficient autoregressive models to monitoring underground pipelines for corrosion. It appears that powerful tools for dependent data are now in place that allow sequential analysis to interact more closely with time series analysis.

4. Quality Control, Fault Detection and Diagnosis

Professor Pollak's comment that sub-optimal procedures are often used in practice in lieu of better statistical methods applies not only to industrial quality control, but also to data analysis and study design by practitioners in industry

and business. What they seem to lack in statistical sophistication is often counterbalanced by experience and practical insights. There are also many other factors in their choice of methods, such as ease of interpretation and implementation within the company's structure, and whether the methods used can already meet the company's needs. In the case of the Shewhart or EWMA chart versus the CUSUM or Shiryaev-Roberts chart, the Shewhart chart may already be adequate for monitoring departures of mean level or variability from the state of statistical control if sampling inspection is scheduled periodically with a sufficiently large sample size at each inspection. Moreover, industrial engineers and plant workers are familiar with the Shewhart chart. Moving averages are intuitively appealing for monitoring potential changes. Besides quality control, moving average charts are widely used in the so-called "technical analysis" of financial markets; see e.g., Neftci (1991). In fact, Page (1954, page 100) might have first tried to use moving averages to improve the Shewhart charts, but because "the consequences of rules based on moving averages are difficult to evaluate", finally came up with the CUSUM chart for which he could use the theory of Wald's SPRT to evaluate the ARL. Page's comment on the analytic difficulty motivated me to develop tools to analyze the ARL of moving average charts in Lai (1974). Thinking that such moving average rules would be sub-optimal in comparison with the CUSUM rule, for which Lorden (1971) had already established an asymptotic optimality property, I did not pursue the theory of moving average charts further. Nearly twenty years later, in my attempts to address some open problems concerning the Willsky-Jones (1976) rule, I was pleased to find that, by a suitable choice of the window size, the moving average rule can be asymptotically as efficient as the CUSUM rule; see Lai (1995).

Although the Shewhart and EWMA rules are adequate for simple manufacturing systems, they are simplistic and perform poorly for monitoring more complicated engineering systems, for which high-dimensional and serially correlated observations (from many sensors and actuators) arrive sequentially and fault detection has to be carried out in real time. This is the setting considered by Willsky and Jones (1976) when they developed the window-limited GLR rule (3.11) for the state-space model (3.10a,b). I expect that many of the ideas in the fault detection literature will be adopted in the next generation of control chart schemes. In his critique of EWMA charts, Hunter (1990) says, "Current literature on statistical-process-control schemes seems to be captured by the Shewhart paradigm of constant mean and independent errors. The problems faced by Shewhart concerned the quality of product manufactured for the bits-and-pieces industries. In those early days, it was common to produce a batch of items and then randomly to sample the batch ... But that condition occurs much less frequently today. In today's bits-and-pieces industries, one often finds

each sequentially produced item measured and recorded ... Certainly, many of today's industrial environments violate the independence assumption employed by Shewhart." Moreover, because of advances in instrumentation and computer technology, the quality characteristics recorded in modern production plants are typically multidimensional, so there has been growing interest in multivariate control charts during the past decade. It appears that many of these issues have already been considered in the fault detection literature, where a comprehensive methodology is currently available.

Because of the multivariate nature of the measurements and of the system states, an important problem of fault detection is diagnosis of the fault that has occurred (e.g., whether it is a sensor or actuator failure and, in the former case, which of the sensors failed). This problem is called *fault isolation* and is closely related to multiple testing discussed earlier in the rejoinder. Nikiforov (1995), Lai (2000), and the references therein give important advances in the theory and applications of the sequential fault detection-isolation (FDI) in the last fifteen years.

5. Nonlinear Filtering and Estimation of Time-Varying Parameters

Barnard (1959) was led to the GLR rule (3.9) in the case of independent normal X_t by an estimation problem, namely, estimating possibly changing means μ_t when the variance remains constant and known. Chernoff and Zacks (1964) considered the same estimation problem and gave an expression for the Bayes estimate. Specifically, assuming the sequence of change-points of $\{\mu_t\}$ to form a discrete renewal process with geometric interarrival times with parameter p and the jumps of $\{\mu_t\}$ to be i.i.d. normal, their expression for the Bayes estimate $\hat{\mu}_n = E(\mu_n|X_1, \ldots, X_n)$ requires $O(2^n)$ operations to compute. Yao (1984) later found another representation of $\hat{\mu}_n$ that requires only $O(n^2)$ operations. Thus, even in this simple example, the memory required and the number of operations needed to compute the Bayes estimate grow to ∞ with n. Although in practice mean shifts typically occur very infrequently (i.e., p is very small), the unknown times of their occurrence leads to the great complexity of the Bayes estimate $\hat{\mu}_n$. By extending the "window limiting" idea in the detection problem, Lai and Liu (2000) have developed an alternative estimate $\tilde{\mu}_n$ which involves no more than a fixed number (depending on p) of parallel recursions such that the Bayes risk $\Sigma_{t=1}^n E(\tilde{\mu}_t - \mu_t)^2$ is asymptotically equivalent to that with $\hat{\mu}_t$ in place of $\tilde{\mu}_t$, as $p \to 0$ and $np \to \infty$. This alternative estimator approximates $\hat{\mu}_t$ by using a bounded-complexity mixture (BCMIX). The BCMIX estimator $\tilde{\mu}_t$ is close in spirit to Yao's method for computing the exact Bayes procedure $\hat{\mu}_t$ but keeps only a fixed number of linear filters at every stage. The BCMIX procedure can also be readily extended to the smoothing problem of estimating μ_t from $X_1, \ldots, X_n$

for $1 \leq t \leq n$. Yao (1984) developed an algorithm involving $O(n^3)$ operations to compute the Bayes estimator $E(\mu_t|X_1, \ldots, X_n)$ by combining $E(\mu_t|X_1, \ldots X_t)$ with the (backward) Bayes estimate $E(\mu_t|X_t, \ldots, X_n)$ that can be evaluated by time reversal. For the BCMIX approximation to $E(\mu_t|X_1, \ldots, X_n)$, m forward linear filters are combined with m backward (time-reversed) linear filters in a way similar to that of Yao (1984). By choosing m suitably, Lai and Liu (2000) showed that the BCMIX estimates μ_t^* of μ_t have a cumulative Bayes risk $\Sigma_{t=1}^n(\mu_t^* - \mu_t)^2$ that is asymptotically equivalent to $\Sigma_{t=1}^n\{E(\mu_t|X_1, \ldots, X_n) - \mu_t\}^2$ as $p \to 0$ and $np \to \infty$. They also extended the BCMIX approach to more general situations in which the unknown parameters may undergo occasional changes, and to empirical Bayes models in which the hyperparameters of the Bayesian model (p and the variance of a jump in μ_t) are not specified *a priori* and have to be estimated from the data.

When one analyzes data that arrive sequentially over time, it is important to detect secular changes in the underlying model which can then be adjusted accordingly. Estimation of time-varying parameters in stochastic systems is, therefore, of fundamental interest in sequential analysis. Furthermore, it arises in many engineering, econometric and biomedical applications and has an extensive literature widely scattered in these fields. Lai and Liu (2000) have given a review of some recent literature and applied the BCMIX procedure to estimate the compositional variations of a genome sequence, yielding a procedure with much lower computational complexity than those in the literature but with similar statistical properties. Clark (1991) has considered the problem of estimating occasionally changing means in the framework of continuous-time Wiener process and casts it in the framework of nonlinear filtering theory. Since optimal nonlinear filters are typically infinite-dimensional, a major thrust in the literature is to develop implementable finite-dimensional approximations. Understanding how this can be done efficiently in the special case of estimating time-varying parameters that are treated as states (undergoing Bayesian dynamics) will provide new insights and advances in nonlinear filtering.

6. Computational Issues

Branching out from simple i.i.d. models in traditional sequential analysis to multivariate time series models whose parameters may undergo secular changes raises many new computational issues. The real-time computational requirement for on-line implementation of sequential procedures in engineering applications poses additional challenges. As Professors Lorden and Turnbull point out, the vast improvements in computing power, algorithms and information technology have well equipped sequential analysis to take on these challenges.

Professor Ghosh comments that "practically all useful results in sequential analysis since 1970 are asymptotic by nature" and that today's computing power can be used to check their adequacy and perhaps also to alter the reliance on asymptotic approximations. Such accuracy checks have in fact been carried out for many of these approximations although they have not been extensively documented because of journal space limitations. To those not working in the field, a puzzling question may be the following: since numerical or simulation procedures are purportedly available to provide accurate answers so that they can be used as benchmarks to check the asymptotic approximations, why should one worry about approximations in the first place? This question can be answered from three viewpoints.

First, the numerical or simulation procedure may be very time-consuming and difficult to program for the general user without access to the developer's software, whereas fast and simple approximations whose accuracy has been ascertained are indeed very useful in practice. A case in point is American option pricing described in Section 7 of the paper. The Cox-Ross-Rubinstein (1979) binomial tree method with $10,000$ (or more) time steps has been used in the finance literature to price an American option accurately. However, implementation of dynamic hedging strategies depends on fast computation of a large number of option prices and hedge parameters (instead of the price of a single option), and therefore fast and accurate approximations to option values are needed for practical management of option books. Anyway there is no exact solution to the American option valuation problem. The binomial tree or other finite difference methods with a large number of time steps are also numerical approximations themselves.

Second, while the computer can be used to generate a numerical result for each special case, it does not provide a general picture or pattern. Asymptotic analysis is particularly useful in leading us to a general theory. A classic example is Fisher's theory of maximum likelihood, and sequential analysis needs asymptotic analysis even more because of its greater complexity. Asymptotic analysis and numerical computations are complementary to each other, and both are needed to develop a successful statistical procedure.

Third, asymptotic analysis often provides valuable clues in the development of efficient computational procedures. For example, the change-of-measure argument that Wald (1945) used in his derivation of approximations (ignoring overshoots) for the error probabilities of the SPRT has played a key role in the development of importance sampling methods to evaluate boundary crossing probabilities by Monte Carlo simulation, as pointed out by Professor Pollak. To develop importance sampling algorithms for Monte Carlo computation of boundary crossing probabilities in sequential change-point detection, Lai (1995b) and

Lai and Shan (1999) independently arrived at the same tilting measure used by Yakir and Pollak (1998) to derive their new characterization of the renewal-theoretic constant that appears in large deviation approximations to first passage probabilities. In multiparameter problems and Markov-dependent situations, asymptotic approximations to boundary crossing probabilities are often difficult to compute directly, since they involve multivariate integrals over complicated regions and Markovian fluctuation theory, but their derivation illuminates how the tilting measure for importance sampling in these complicated situations can be constructed; see Chan and Lai (2000c, d). Making use of similar asymptotic analysis, Chen (2001) has recently developed a systematic theory for *sequential importance sampling with resampling*, introduced by Liu and Chen (1995, 1998) to implement nonlinear filtering in state space models via Monte Carlo, and by Kong, Liu and Wong (1994) to perform sequential imputations in missing data problems.

Professor Gu has described a very interesting "hybrid" of stochastic approximation and Markov Chain Monte Carlo (MCMC) and its applications to spatial statistics; see also Gu and Kong (1998) for other applications to incomplete data problems. Stochastic approximation has also provided powerful tools for the convergence analysis of recursive simulation in *neuro-dynamic programming*, which attempts to overcome the "curse of dimensionality" in dynamic programming by combining function approximation, via neural networks and other basis functions, with temporal-difference learning algorithms to simulate the cost-to-go function of a policy in a controlled Markov chain; see Bertsekas and Tsitsiklis (1996).

7. Theoretical Issues

Professor Pollak points out that, whereas mixture likelihood ratios or the adaptive likelihood ratio martingales introduced by Robbins and Siegmund (1973) have second-order optimality properties for testing sequentially a simple null hypothesis against a composite alternative in a one-parameter exponential family, it is not known if the GLR also possesses such property. His comment prompted Lai (2001) to prove the second-order optimality of sequential GLRs with suitably chosen time-varying stopping boundaries in multiparameter exponential families, where the null hypothesis is a q-dimensional submanifold of the p-dimensional parameter space, $q < p$. The special case $q = 0$ reduces to the simple null hypothesis considered by Pollak (1978). This second-order optimality theory of sequential GLRs is, therefore, applicable to a much wider class of null hypotheses (including the presence of nuisance parameters) than those considered for mixture likelihood ratios or adaptive likelihood ratio martingales. Besides the obvious connections to the well developed theory of fixed sample size tests explained earlier, the reason why I prefer GLR statistics is that they are

"automatic" and there is no apparent loss of information. In contrast, mixture likelihood ratio tests require suitable specification of the mixing distribution, which depends on what kind of alternatives one wants to focus on. It is possible to improve finite-sample performance by using an elaborate mixing distribution that puts equal mass at close alternatives and at distant ones, as in Kao and Lai (1980). However, this approach is quite limited in practice since one may not be able to integrate the mixture likelihood ratio in closed form for general mixing distributions or for more general parametric families. The adaptive likelihood ratio martingales that replace the unknown θ in $f_\theta(X_i)/f_{\theta_0}(X_i)$ by $f_{\hat{\theta}_{i-1}}(X_i)/f_{\theta_0}(X_i)$, where $\hat{\theta}_{i-1}$ is an estimate of θ based on $X_1, \ldots, X_{i-1}$, are easier to use, but they suffer from the loss of information due to ignoring $X_i, \ldots, X_n$ that are also available at stage n. This loss of information is particularly serious in the multivariate case, where simulation studies have shown that the method can perform poorly in comparison with sequential GLR tests.

Professor Yakir's discussion is also about second-order optimality but in the context of sequential change-point detection where the pre- and post-change distributions of independent observations are completely specified. A subtle point here is the criteria which one uses to assess detection rules. If one uses the traditional false alarm ARL constraint and the worst-case expected delay criterion in (3.2), then the CUSUM rule is (exactly) optimal, as shown by Moustakides (1986). Professor Yakir introduces a different constraint and considers a sharper criterion for assessing detection delay to show that the Shiryaev-Roberts rule is second-order optimal. Although his result is interesting and elegant, it does not provide definitive arguments for choosing the Shiryaev-Roberts rule over the CUSUM or moving average rules, contrary to what he asserts in his third paragraph. His choice of performance criteria to demonstrate the second-order superiority of the Shiryaev-Roberts rule may not be agreeable to those who have used an alternative set of performance criteria to establish the optimality of the CUSUM rule. Asymptotic theory and simulation studies have shown that CUSUM, Shiryaev-Roberts and moving average rules (with suitably chosen window sizes) have similar performance according to various criteria in the simple normal model of known means before and after a single change-point, and second-order optimality according to one of several equally plausible performance criteria does not seem to be a compelling reason for favoring the Shiryaev-Roberts rule over its competitors.

Concerning multi-armed bandits, Professor Woodroofe comments that the covariate bandit problem with non-parallel regression lines is qualitatively very different from the traditional bandit problems reviewed in Section 6. The dilemma between information and control disappears in these covariate bandits and, under some regularity conditions, the myopic rule is second-order optimal.

A unified theory incorporating both covariate bandits and traditional bandits is provided by the asymptotic optimality theory of adaptive choice from a finite set of stationary control laws in controlled Markov chains. In this more general context, what is at stake is a certain set $B(\theta)$ of parameter values, called the "bad set", following the notation of Graves and Lai (1997). For covariate bandits, if no two regression lines are parallel, then $B(\theta)$ is empty and the myopic rule is asymptotically optimal. On the other hand, if some regression lines are parallel, then $B(\theta)$ is non-empty and some "uncertainty adjustments" have to be introduced into the myopic rule, as in traditional bandits.

8. Optimization Over Time and Its Role in Sequential Statistical Procedures

Professor Lorden says that "optimality theory has not been very influential in the development of sequential analysis in the last forty years." One reason is that, except for some simple cases such as the Wald-Wolfowitz solution of the optimal stopping problem that proves the optimality of the SPRT, optimal stopping and other sequential optimization problems in sequential testing, detection or estimation do not have closed-form solutions. Numerical solution of the dynamic programming equation often does not provide much insight into the general pattern since it requires precise specification of the loss function and prior distribution, and the "curse of dimensionality" also makes it hard to carry out for multivariate problems.

My experience, however, is that the ability to solve some related optimal stopping/dynamic programming problem that is both more tractable and prototypical can provide important insights. In Lai (1987, 1988, 1997) and Lai, Levin, Robbins and Siegmund (1980), solution of a tractable optimal stopping problem in some special case has shed light on easily interpretable approximations in a general setting, and the performance of these approximations can then be studied by asymptotic analysis and Monte Carlo simulations. Thus, one does not have to solve an optimal stopping problem in each special (and potentially difficult) case, but can make guesses on what nearly optimal procedures should look like based on what one learns from a particular generic example together with one's statistical insights and intuition. Sometimes, as Professors Chernoff and Yakir point out, one can bypass backward induction/dynamic programming calculations altogether by asymptotic analysis and/or lower bound arguments.

Irrespective of its impact on the development of efficient sequential statistical procedures, optimal stopping and other sequential optimization techniques (see e.g., Whittle (1982, 1983)) are important topics that should be included in courses on sequential analysis. Optimization over time has become an active area of research in several other fields, such as nonlinear partial differential

equations, control engineering, operations research, economics and finance, and a well designed course on sequential analysis that incorporates both sequential optimization techniques and sequential statistical methods, and which strikes a good balance between theory and applications, should appeal to graduate students in statistics and other disciplines.

9. Reinvigoration and Adaptation to Changing Times and Trends

Even though Professor Simons doubts that sequential analysis can thrive in the contemporary "culture that seriously undervalues mathematical analysis in favor of computational facility," I remain optimistic. Perhaps at this point in the development of sequential analysis, putting more emphasis on "computation" (and implementation) than on "mathematical analysis" (targeted towards conceptual innovations of the kind mentioned in Professor Simons' discussion) may actually promote (rather than inhibit) its growth. As pointed out in Sections 2, 3, 5 and 6 of this rejoinder, what is needed for sequential analysis to be more widely applicable and to broaden its appeal is implementation (and extension to more complex models) of its conceptual innovations by making use of current computer power and other advanced technology. Now is the time for sequential analysis to reinvigorate itself by taking advantage of (instead of being depressed by) the strengths (instead of the weaknesses) of the current environment. With many powerful analytic tools and fundamental concepts already in place, sequential analysis can now take a bold step forward and change its traditional outlook so that it can relate more effectively to new trends in the world of science, technology and socio-economic activities. In a broad sense, sequential analysis is about adaptation – how statistical modeling and decisions can be efficiently adapted to new information and to changes in the underlying system. It should, therefore, practice what it preaches and be able to "adapt to, without being consumed by" the changing culture that Professor Simons alludes to.

10. New Challenges and Emerging Opportunities

Professors Siegmund and Turnbull have described a number of exciting applications of sequential analysis and related boundary crossing problems to genetic analysis and clinical biostatistics. They have also indicated that what was regarded as impractical in biomedical applications of sequential methods is now gaining wide acceptance. Even the adoption of adaptive treatment allocation rules in clinical trials, whose difficulty I briefly summarized in the last paragraph of Section 6 of the paper, has been gaining ground. Incidentally, I only cited Armitage (1985) and the discussants of his paper who painted what Professor Turnbull calls a "bleak" outlook for the application of outcome-dependent allocation rules in clinical trials, but did not "purport" such outlook myself. However there

are still many practical difficulties to be resolved, as pointed out by Armitage and other discussants in their discussion of the ECMO trials in Ware (1989). Of course, some of these difficulties will disappear when Professor Woodroofe's vision of a "centralized" multi-center trial, supported by data communication networks, materializes.

There are also emerging opportunities for sequential analysis to make important contributions in other fields. An example is the adaptive control of engineering systems in Professor Wei's discussion. Making use of the convergence properties of recursive least squares estimates in stochastic regression models in Lai and Wei (1982), Lai (1986), Wei (1987b), together with an ingenious analysis of the dynamics of the associated certainty equivalence control rule, Guo (1995) established a logarithmic rate for the regret (similar to (5.12)) of the least squares certainty equivalence rule in his IFAC (International Federation of Automatic Control) award winning paper. Another fast growing field that offers new opportunities and challenges to sequential analysis is cognitive science, encompassing a broad spectrum of areas like neuropsychology, learning and memory, artificial intelligence and machine learning. An application area associated with learning and memory is educational testing. Computerized adaptive testing, which is an active area of research in psychometrics and education, is clearly sequential in nature. The irreversible correlated multi-armed bandit problem with Markovian rewards described by Professors Fuh and Hu was motivated by applications to computerized adaptive testing. Lai and Yakowitz (1995) and Kaebling, Littman and Moore (1996) have reviewed several applications of multi-armed bandits to machine learning. Finally, the last paragraph of Professor Turnbull's discussion describes many other emerging opportunities and new challenges for sequential analysis in this "information age" and "new economy".

Additional References

Albert, A. E. (1961). The sequential design of experiments for infinitely many states of nature. *Ann. Math. Statist.* **32,** 774-799.

Albert, A. (1966). Fixed size confidence ellipsoids for linear regression parameters. *Ann. Math. Statist.* **37**, 1602-1630.

Andersen, J. S. (1996). Clinical trial designs—made to order. *J. Biopharm. Statist.* **6**, 515-522.

Appleby, R. H. and Freund, R. J. (1962). An empirical evaluation of multivariate sequential procedure for testing means. *Ann. Math. Statist.* **33**, 1413-1420.

Armitage, P. (1975). *Sequential Medical Trials.* 2nd edition. Blackwell, Oxford.

Bahadur, R. R. (1967). An optimal property of the likelihood ratio statistic. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1**, 13-26. University California Press.

Bergman, S. W. and Turnbull, B. W. (1983). Efficient sequential designs for destructive life testing with application to animal serial sacrifice experiments. *Biometrika* **70**, 305-314.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming.* Athena Scientific, Belmont (MA).

Bessler, S. (1960a). *Theory and Applications of the Sequential Design of Experiments, k-actions and Infinitely Many Experiments, Part I - Theory.* Technical Report 55, Department of Statistics, Stanford University, California.

Bessler, S. (1960b). *Theory and Applications of the Sequential Design of Experiments, k-actions and Infinitely Many Experiments, Part II - Applications.* Technical Reportt 56, Department of Statisitics, Stanford University, California.

Betensky, R. A. (1998). Construction of a continuous stopping boundary from an alpha spending function. *Biometrics* **54**, 1061-1071.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of the density function estimates. *Ann. Statist.* **1**, 1071-1095.

Bloch, D., Lai, T. L. and Tubert-Bitter, P. (2000). One-sided tests in clinical trials with multiple endpoints. Technical Report, Department of Statistics, Stanford University.

Box, G. E. P. and Wilson, K. P. (1951). On the experimental attainment of optimal conditions (with discussion). *J. Roy. Statist. Soc. Ser. B* **13**, 1-45.

Brown, L. D., Cai, T. and Das Gupta, A. (2000). Interval estimation in discrete exponential family. Manuscript, Purdue University.

Carroll, R. J. (1976). On sequential density estimation. *Z. Wahrsch. Verw. Gebiete* **36**, 137-151.

Chan, H. P. and Lai, T. L. (2000c). Asymptotic approximations for error probabilities of sequential or fixed sample size tests in exponential families. *Ann. Statist.* **28**, in press.

Chan, H. P. and Lai, T. L. (2000d). Importance sampling for Monte Carlo evaluation of boundary crossing probabilities in Markov chains, hypothesis testing and changepoint detection. Technical Report, Department of Statistics, Stanford University.

Chang, Y.-C. I. (1995). Estimation in some binary regression models with prescribed accuracy. *J. Statist. Plann. Inference* **44**, 313-325.

Chang, Y.-C. I. (1996). Sequential fixed size confidence regions for regression parameters in generalized linear models. *Statist. Sinica* **6**, 899-916.

Chang, Y.-C. I. and Martinsek, A. T. (1992). Fixed size confidence regions for parameters of a logistic regression model. *Ann. Statist.* **20**, 1953-1969.

Chen, K. N. and Hu, I. (1998). On consistency of Bayes estimates in a certainty equivalence adaptive system. *IEEE T. Automat. Contr.* **43**, 943-947.

Chen, Y. (2001). Sequential importance sampling with resampling: theory and applications. Ph.D. dissertation (in preparation), Department of Statistics, Stanford University.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of hypotheses based on the sum of observations. *Ann. Math. Statist.* **23**, 493-507.

Chernoff, H. (1959). Sequential design of experiments. *Ann. Math. Statist.* **30,** 755-770.

Chernoff, H. (1965). Sequential tests for the mean of a normal distribution IV (discrete case). *Ann. Math. Statist.* **26**, 55-68.

Chernoff, H. (1972). *Sequential Analysis and Optimal Design.* Soc. Industrial Appl. Math., Philadelphia.

Chernoff, H. (1975). Approaches in sequential design of experiments. In *A Survey of Statistical Design and Linear Models* (Edited by J. N. Srivastava), 67-90. North Holland, Amsterdam.

Chernoff, H. and Petkau, A. J. (1985). *Sequential Medical Trials with Ethical Cost.* Proceedings of the Berkeley Conference in honor of J. Neyman and J. Kiefer, Vol. II (Edited by LeCam and Olshen), 521-537. Wadsworth, Inc., Califiornia.

Chernoff, H. and Zacks (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Statist.* **35**, 999-1018.

Clark, J. M. C. (1991). Anatomy of a low-noise jump filter: part I. In *Stochastic Analysis* (Edited by E. Mayer-Wolf et al.), 111-124. Academic Press, New York.

Coad, D. S. (1995). Sequential allocation involving several treatments. In *Adaptive Designs* (Edited by N. Flournoy and W. F. Rosenberger), 95-109. Inst. Math. Statist., Hayward, CA.

Cornfield, J. (1966). A Bayesian test of some classical hypotheses—with applications to sequential clincal trials. *J. Amer. Statist. Assoc.* **61**, 577-594.

DeMets, D. L. (1984). Stopping guidelines vs stopping rules: a practitioner's point of view. *Commun. Statist. A.*, **13**, 2395-2418.

Denne, J. S. and Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika* **87**, 125-134.

Dixon, W. J. and Mood, A. M. (1948). A method for obtaining and analyzing sensitivity data. *J. Amer. Statist. Assoc.* **43**, 109-126.

Dragalin, V. P. and Novikov, A. A. (1987). Asymptotic solution of the Kiefer-Weiss problem for processes with independent increments. (Russian) Teor. Veroyatnost. i Primenen. 32, no. 4, 679-690.

Dragalin, V. P. (1997). The sequential change point problem. *Economic Quality Control* **12**, 95-122.

Dudoit, S., Yang, Y. H., Callow, M. J. and Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report, Department of Statistics, University of California at Berkeley.

Fan, S.-K. S. (2000). A generalized global optimization algorithm for dual response systems. *J. Qual. Tech.* **32**, 444-456.

Faries, D. (1994). Practical modifications of the continual reassment method for phase I cancer clinical trials. *J. Biopharm. Statist.* **4**, 147-164.

Fakhre-Zakeri, I. and Lee, S. (1992). Sequential estimation of the mean of a linear process. *Sequential Anal.* **11**, 181-197.

Federov, V. V. (1972). *Theory of Optimal Experiments.* Academic Press, New York.

Feingold, E., Brown, P. O. and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent. *Amer. J. Hum. Genetics* **53**, 234-251.

Feldman, D. (1962). Contributions to the "Two-Armed Bandit" problem. *Ann. Math. Statist.* **33,** 847-856.

Fuh, C. D. (1998). Efficient likelihood estimation for hidden Markov models. Tentatively accepted by *Ann. Statist.*

Fuh, C. D. (2000). SPRT and Cusum in hidden Markov models. Technical Reports.

Fuh, C. D. and Hu, I. (2000). Asymptotically efficient strategies for a stochastic scheduling problem with order constraints. *Ann. Statist.* **28**, in press.

Fuh, C. D. and Lai, T. L. (1998). Wald's equations, first passage times and moments of ladder variables in Markov random walks. *J. Appl. Probab.* **35**, 566-580.

Gleser, L. J. (1965). On the asymptotic theory of fixed-size sequential confidence bounds for linear regression parameters. *Ann. Math. Statist.* **36**, 463-467.

Goodwin, G. C. and Sin, K. S. (1984). *Adaptive Filtering, Prediction and Control.* Prentice Hall, Englewood Cliffs N. J.

Goodman, S. N., Zahurak, M. L. and Piantadosi, S. (1995). Some practical improvements to the continual reassessment method for phase I studies. *Statist. Med.* **14**, 1149-1161.

Gordon, L. and Pollak, M. (1995). A robust surveillance scheme for stochastically ordered alternatives. *Ann. Statist.* **23**, 1350-1375.

Grambsch, P. (1989). Sequential maximum likelihood estimation with applications to logistic regression in case-control studies. *J. Statist. Plann. Inference* **22** 355-369.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 7270-7274.

Gu, M. G. and Zhu, H.-T. (2000). Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation. *J. Roy. Statist. Soc. Ser. B.* To appear.

Guo, L. (1994). Further results on least squares based adaptive minimum variance control. *SIAM J. Contr. Optimiz.* **32**, 187-212.

Guo, L. (1995). Convergence of logarithmic laws of self-tuning regulators. *Automatica* **31**, 435-450.

Hayre, L. S. and Turnbull, B. W. (1981a). Sequential estimation in two-armed exponential clinical trials. *Biometrika* **68**, 411-416.

Hayre, L. S. and Turnbull, B. W. (1981b). Estimation of the odds ratio in the two-armed bandit problem. *Biometrika* **68**, 661-668.

Hochberg, Y. (1998). A sharper Bonferonni procedure for multiple tests of significance. *Biometrika* **75**, 800-802.

Hoeffding, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36**, 369-405.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* **6**, 65-70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferonni test. *Biometrika* **75**, 383-386.

Hu, I. (1988). Repeated significance tests for exponential families. *Ann. Statist.* **16**, 1643-1666.

Hu, I. (1997). Strong consistency in stochastic regression models via posterior covariance matrices. *Biometrika* **84**, 744-749.

Hu, I. (1998). On sequential designs in nonlinear problems. *Biometrika* **85**, 496-503.

Hu, I. and Wei, C. Z. (1989). Irreversible adaptive allocation rules. *Ann. Statist.* **17**, 801-823.

Huffman, M. D. (1983). An efficient approximate solution to the Kiefer-Weiss problem. *Ann. Statist.* **11**, 306-316.

Isogai, E. (1981). Stopping rules for sequential density estimation. *Bull. Math. Statist.* **19**, 53-67.

Isogai, E. (1987). The convergence rate of fixed-width sequential confidence intervals for a probability density function. *Sequential Anal.* **6**, 55-69.

Isogai, E. (1988). A note on sequential density estimation. *Sequential Anal.* **7**, 11-21.

Izenman, A. (1991). Recent developments in nonparametric density estimation. *J. Amer. Statist. Assoc.* **86**, 205-224.

Jennison, C., Johnstone, I. M. and Turnbull, B. W. (1982). Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. In *Statistical Decision Theory and Related Topics III,* Vol. 2 (Edited by S. S. Gupta and J. O. Berger), 55-86. Academic Press, New York.

Jennison, C. and Turnbull, B. W. (1993). Group sequential tests for bivariate response: interim analyses of clinical trials with both efficacy and safety endpoints. *Biometrics* **49**, 741-752.

Kaebling, L. P., Littman M. C. and Moore, A. W. (1996). Reinforcement learning: a survey. *J. Artificial Intelligence Res.* **4**, 237-285.

Kao, S. and Lai, T. L. (1980). Sequential selection procedures based on confidence sequences for normal populations. *Comm. Statist. Ser. A* **9**, 1657-1676.

Kiefer, J. and Sacks, J. (1963). Asymptotically optimum sequential inference and design. *Ann. Math. Statist.* **34**, 705-750.

Kong, A., Liu J. S. and Wong, W. H. (1994). Sequential imputations and Bayesian missing-data problems. *J. Amer. Statist. Assoc.* **89**, 113-119.

Kundu, S. and Martinsek, A. T. (1997). Bounding the L_1 distance in nonparametric density estimation. *Ann. Inst. Statist. Math.* **49**, 57-78.

Lai, T. L. (1986). Asymptotically efficient adaptive control in stochastic regression models. *Adv. Appl. Math.* **7**, 23-45.

Lai, T. L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann. Statist.* **22**, 1917-1930.

Lai, T. L. (1995b). Boundary crossing problems in sequential analysis and time series. *Bull. Internat. Statist. Inst.* **56**, 499-515.

Lai, T. L. (1996). On uniform integrability and asymptotically risk-efficient sequential estimation. *Sequential Anal.* **15**, 237-251.

Lai, T. L. (2001). Second-order optimality of sequential generalized likelihood ratio tests in multiparameter exponential families. In preparation.

Lai, T. L. and Liu, T. (2000). A bounded complexity mixture method for estimating time-varying parameters. Technical Report, Department of Statistics, Stanford University.

Lai, T. L. and Robbins, H. (1981). Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. *Z. Wahr. verw. Gebiete* **56**, 329-360.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10**, 154-166.

Lai, T. L. and Yakowitz, S. (1995). Machine learning and nonparametric bandit theory. *IEEE Trans. Automat. Contr.* **40**, 1199-1209.

Lan, K. K. G. and DeMets, D. L. (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics* **45**, 1017-1020.

Lander, E. S. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**, 241-247.

Lin, D. Y. (1991). Nonparametric sequential testing in clinical trials with incomplete multivariate observations. *Biometrika* **78**, 123-131.

Lin, D. Y. and Liu, P. Y. (1992). Nonparametric sequential tests against ordered alternatives in multiple-armed clinical trials. *Biometrika* **79**, 420-425.

Liu, J. and Chen, R. (1995). Blind deconvolution via sequential imputation. *J. Amer. Statist. Assoc.* **90**, 567-576.

Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* **93**, 1032-1044.

Lorden, G. (1983). Asymptotic efficiency of three-stage hypothesis tests. *Ann. Statist.* **11**, 129-140.

Louis, T. A. and Orav, E. J. (1985). Adaptive sacrifice plans for the carcinogen assay. In *Proc. of Symp. on Long-Term Animal Carcinogenicity Studies: A Statistical Perspective*, 36-41. Amer. Statist. Assoc., Washington DC.

Martinsek, A. T. (1992). Using stopping rules to bound the mean integrated squared error in density estimation. *Ann. Statist.* **20**, 797-806.

Martinsek, A. T. and Xu, Y. (1996). Fixed width confidence bands for densities under censoring. *Statist. Probab. Lett.* **30**, 257-264.

Martinsek, A. T. (2000a). Estimation of the maximum and minimum in a model for bounded, dependent data. To be submitted.

Martinsek, A. T. (2000b). Sequential estimation of the mean in a random coefficient autoregressive model with beta marginals. To appear in *Statist. Probab. Lett.*

McDonald, D. (1990). A cusum procedure based on sequential ranks. *Naval Res. Logist.* **37**, 627-646.

McKenzie, E. (1985). An autoregressive process for beta random variables. *Management Sci.* **31**, 988-997.

Meier, P. (1975). Statistics and medical experimentation. *Biometrics* **31**, 511-529.
Melfi, V. F. (1992). Nonlinear Markov renewal theory with statistical applications. *Ann. Probab.* **20**, 753-771.
Morton, N. E. (1955). Sequential tests for the detection of linkage. *Amer. J. Hum. Genet.* **7**, 277-318.
Morris, M. D. (1987). A sequential experimental design for estimating a scale parameter from quantal life testing data. *Technometrics* **29**, 173-181.
Neftci, S. N. (1991). Naive trading rules in financial markets and Wiener-Kolmogorov prediction theory: a study of "technical analysis". *J. Business* **64**, 549-571.
Neyman, J. (1965). Discussion of Hoeffding's paper. *Ann. Math. Statist.* **36**, 401-405.
Novikov, A. A. (1990). On the first passage time of an autoregressive process over a level and an application to a "disorder" problem. *Theory Probab. Appl.* **35**, 269-279.
O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079-1087.
O'Quigley, J., Pepe, M. S. and Fisher, L. (1990). Continual reassessment method: a practical design for phase I trials in cancer. *Biometrics* **46**, 33-48.
Piantadosi, S. and Liu, G. (1996). Improved designs for phase I studies using pharmacokinetic measurements. *Statist. Med.* **15**, 1605-1618.
Pollak, M. (1987). Average run lengths of an optimal method of detecting a change in distribution. *Ann. Statist.* **15**, 749-779.
Proschan, M. A., Follman, D. A. and Waclawiw, M. A. (1992). Effect of assumption violations on Type I error rate in group sequential monitoring. *Biometrics* **48**, 1131-1143.
Quang, P. X. (1985). Robust sequential testing. *Ann. Statist.* **13**, 638-649.
Robbins, H. and Siegmund, D. (1970). Boundary crossing probabilities for the Wiener process and sample sums. *Ann. Math. Statist.* **41**, 1410-1429.
Robbins, H. and Siegmund, D. (1973). A class of stopping rules for testing parametric hypotheses. Proc. 6th Berkely Symp. Math. Statist. Probab. 4, University of California Press.
Robertson, T., Wright, F. and Dykstra, R. (1988). *Order Restricted Inference*. Wiley, New York.
Sarkar, J. (1991). One armed bandit problems with covariates. *Ann. Statist.* **19**, 1978-2002.
Savage, I. R. and Sethuraman, J. (1966). Stopping time of rank-order sequential probability ratio tests based on Lehmann alternatives. *Ann. Math. Statist.* **37**, 1154-1160.
Sethuraman, J. (1970). Stopping time of rank-order sequential probability ratio tests based on Lehmann alternatives II. *Ann. Math. Statist.* **41**, 1322-1333.
Shaffer, J. P. (1995). Multiple hypothesis testing. *Ann. Rev. Psychology* **46**, 561-584.
Shibata, T. (1996). 1996 Whitney Award Lecture: statistical and stochastic approaches to localized corrosion. *Corrosion Sci.* **52**, 813-830.
Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory Probab. Appl.* **8**, 22-46.
Shiryaev, A. N. (1978). *Optimal Stopping Rules*. Springer-Verlag, New York.
Siegmund, D. (1975). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4**, 673-684.
Siegmund, D. (1978). Estimation following sequential tests. *Biometrika* **65**, 341-349.
Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.
Siegmund, D. (1993). A sequential clinical trial for comparing three treatments. *Ann. Statist.* **21**, 464-483.

Siegmund, D. (1999). Note on a stochastic recursion. (To appear in *Festschrift for Willem Van Zwet*, Institute of Mathematical Statistics, Harvard, California.

Siegmund, D. and Yakir, B. (2000b). Approximate p-values for local sequence alignments. *Ann. Statist.* **28**, 637-680.

Skouras, K. (2000). Strong consistency in nonlinear stochastic regression models. *Ann. Statist.* **28**, 871-879.

Slate, E. H. and Turnbull, B. W. (2000). Statistical models for longitudinal biomarkers of disease onset. *Statist. Med.* **19**, 617-637.

Sriram, T. N. (1987). Sequential estimation of the mean of a first order stationary autoregressive process. *Ann. Statist.* **15**, 1079-1090.

Srivastava, M. S. (1967). On fixed-width confidence bounds for regression parameters and mean vector. *J. Roy. Statist. Soc. Ser. B* **29**, 132-140.

Srivastava, M. S. (1971). On fixed-width confidence bounds for regression parameters. *Ann. Math. Statist.* **42**, 1403-1411.

Stute, W. (1983). Sequential fixed-width confidence intervals for a nonparametric density function. *Z. Wahrsch. Verw. Gebiete* **62**, 113-123.

Tamura, R. N., Faries, D. E., Andersen, J. S. and Heiligenstein, J. H. (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *J. Amer. Statist. Assoc.* **89**, 768-776.

Turnbull, B. W. and Hayter, A. J. (1985). A forward stochastic approximation procedure for scheduling sacrifices in tumorigenicity studies. *Proceedings Biopharmaceutical Section, Annual Meeting of American Statistical Association.* Las Vegas, Nevada, August 1985, pp.131-136.

Vining, G. G. and Myers, R. H. (1990). Combining Taguchi and response surface philosophies: a dual response approach. *J. Qual. Tech.* **22**, 38-45.

Wald, A. (1947). *Sequential Analysis.* John Wiley and Sons, New York.

Ware, J. H. (1989). Investigating therapies of potentially great benefit: ECMO (with discussion). *Statist. Sci.* **4**, 298-340.

Wei, C. Z. (1987a). Multivariate adaptive stochastic approximation. *Ann. Statist.* **15**, 1115-1130.

Wei, C. Z. (1987b). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. *Ann. Statist.* **15**, 1667-1682.

Whittle, P. (1982 & 1983). *Optimization over Time*, vol. 1 & 2. Wiley, New York.

Woodroofe, M. (1976b). Frequentist properties of Bayesian sequential tests. *Biometrika* **63**, 101-110.

Woodroofe, M. (1978). Large deviations of likelihood ratio statistics with applications to sequential testing. *Ann. Statist.* **6**, 72-84.

Woodroofe, M. (1979). A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.* **74**, 799-806.

Woodroofe, M. (1989). Very weak expansions for sequentially designed experiments: linear models. *Ann. Statist.* **17**, 1087-1102.

Woodroofe, M. and Sun, J. (1999). A penalized maximum likelihood estimate of $f(0+)$ when f is non-increasing. *Statist. Sinica* **3**, 501-515.

Wu, C. F. J. (1985). Asymptotic inference from sequential design in a nonlinear situation. *Biometrika* **72**, 553-558.

Wu, W., Woodroofe, M. and Mentz, G. (2000). Isotonic regression: another look at the change point problem. Technical Report 355, Statistics Department, University of Michigan. Submitted to *Biometrika*.

Xu, Y. and Martinsek, A. T. (1995). Sequential confidence bands for densities. *Ann. Statist.* **23**, 2218-2240.

Yakir, B. (1995). A note on the run length to false alarm of a change-point detection policy. *Ann. Statist.* **23**, 272-281.

Yakir, B. (1997b). A note on optimal detection of a change in distribution. *Ann. Statist.* **25**, 2117-2126.

Yakir, B. (2001). Optimal detection of a change of a slope. In preparation.

Yakir, B. and Pollak, M. (1998). A new representation for a renewal-theoretic constant appearing in asymptotic approximations of large deviations. *Ann. Appl. Probab.* **8**, 749-774.

Yao, Y. C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *Ann. Statist.* **12**, 1434-1497.