

# Martingales in Sequential Analysis and Time Series, 1945–1985\*

TZE LEUNG LAI<sup>†</sup>

## Abstract

This paper reviews the history of martingales in sequential analysis, beginning with Wald's ground-breaking paper in 1945 that laid the foundations for the subject, and ending in the decade 1975–1985 when the usefulness of martingale theory was also recognized in time series analysis. The important roles played by martingale inequalities, convergence theory, strong laws, functional central limit theorems, laws of the iterated logarithm, and the optional stopping theorem in developments of the theory of sequential analysis and time series from 1945 to 1985 are also discussed.

---

\*Research supported by the National Science Foundation grant DMS 0805879.

<sup>†</sup>Department of Statistics, Stanford University, Stanford, CA 94305-4065, USA  
lait@stat.stanford.edu

# 1 Introduction

Asymptotic results in statistical theory are closely related to limit theorems in probability. Since martingale theory has played a central role in these limit theorems, one would expect a rich history of martingales in theoretical statistics. However, because the development of statistical theory mostly focused on the setting of independent or exchangeable observations, martingale theory did not play a significant role in statistics until more complicated data-generating mechanisms had to be considered for statistical inference and modeling of the data being analyzed. The history of martingales in statistics, therefore, is closely linked to the developments in sequential analysis, time series and survival analysis, for which the martingale structure inherent in the data and the powerful tools from martingale theory have led to major advances in the statistical methodologies. This paper considers the history of martingales in sequential analysis and time series during the 40-year period 1945–1985 when the role of martingale theory in these areas grew from being peripheral to central. The paper by Aalen, Andersen, Borgan, Gill and Keiding in this special issue describes the history of martingales in survival analysis.

Following the style of Rudin’s (1997) book in the *History of Mathematics* series published jointly by the American Mathematical Society and the London Mathematical Society, I try to describe the history not only from archival works but also from my personal experience, first as a graduate student at Columbia University (from 1968 to 1971) where sequential analysis and martingale theory were advanced courses in the curriculum as well as major research areas of the faculty, and then as a researcher in these fields and in time series. Rudin’s exemplary book inspires me “to tell — as well as I could — where the problems came from, what some of their solutions led to, and who (were) involved.” In view of my personal experience, I find it helpful to divide the period into two parts, namely 1945–1975 and 1976–1985, which are presented in Sections 2 and 3, respectively. The first part, which will be labeled “Sequential Analysis from 1945 to 1975”, covers the beginnings and subsequent developments of this field and how they interacted with martingale theory. It also reflects what I learned as a graduate student writing a Ph.D. thesis entitled *Confidence Sequences and Martingales* that belongs to the intersection of both fields. The second part, labeled “From Sequential Analysis to Time Series”, describes the developments in the last ten years of the period, and in particular how I and others saw and exploited

the power of martingale theory to resolve some open problems in time series analysis, sequential experimentation and adaptive control during that time. The paper concludes with an epilogue in Section 4.

## 2 Sequential Analysis from 1945 to 1975

Sequential analysis refers to the analysis of data generated from sequential experiments, for example, when the sample size is not fixed in advance but may depend on the observations collected so far. Wallis (1980, Sect. 6) gives a detailed historical account of the origins of sequential analysis at the Statistical Research Group (SRG) that was set up at Columbia University during the Second World War to advise the United States Department of Defense. Wald’s (1945) ground-breaking paper on the sequential probability ratio test (SPRT), a preliminary version of which already appeared in a “restricted report” of SRG to the National Defense Research Committee in 1943, marks the beginning of the theory of sequential analysis. Although Wald developed this theory from scratch, it was soon recognized that the theory of martingales could simplify and generalize some of Wald’s methods and results. Moreover, sequential analysis then grew in different directions which also had close interactions with martingale theory. The following subsections describe a number of important developments in the interplay between martingale theory and sequential analysis in the period 1945–1975 that also witnessed the blossoming of both fields.

### 2.1 Martingale theory and sequential likelihood ratio statistics

Wald’s SPRT is concerned with testing a simple null hypothesis  $H_0 : f = f_0$  versus a simple alternative hypothesis  $H_1 : f = f_1$  based on independent and identically distributed (i.i.d.) observations  $X_1, X_2, \dots$ , having a common density function  $f$  (with respect to some measure  $m$ ). Let  $L_n = \prod_{i=1}^n (f_1(X_i)/f_0(X_i))$  be the likelihood ratio statistic based on  $X_1, \dots, X_n$ . The SPRT stops sampling at stage

$$N = \inf \{n \geq 1 : L_n \notin (A, B)\}, \quad (2.1)$$

with  $A < 1 < B$ , and rejects  $H_0$  if  $L_N \geq B$ . To analyze the error probabilities of the SPRT, Wald (1945) introduced the likelihood ratio identities

$$P_0(L_N \geq B) = E_1(L_N^{-1} 1_{\{L_N \geq B\}}), \quad P_1(L_N \leq A) = E_0(L_N 1_{\{L_N \leq A\}}). \quad (2.2)$$

It follows from (2.2) that  $P_0(L_N \geq B) \leq B^{-1}P_1(L_N \geq B)$  and  $P_1(L_N \leq A) \leq AP_0(L_N \leq A)$ , in which  $\leq$  can be replaced by  $=$  if  $L_N$  has to fall on either boundary exactly (i.e., if there is no overshoot). Ignoring overshoots, Wald made use of (2.2) to obtain approximations for the error probabilities  $P_0(L_N \geq B)$  and  $P_1(L_N \leq A)$ .

To analyze the operating characteristics of the SPRT, Wald made use of another identity, which was not “restricted” under the Department of Defense rules and which he published a year earlier. Wald (1944) proved this identity for more general i.i.d. random variables than  $\log(f_1(X_i)/f_0(X_i))$  that are the summands of  $\log L_n$ . Although he derived it without martingale theory by using the particular structure (2.1) of the stopping rule, Blackwell and Girshick (1946) used the martingale structure to generalize the identity to more general stopping times. Doob (1953, Sect. VII.10) extended the result further to the following form, which he called “the fundamental theorem of sequential analysis.” Let  $Y_1, Y_2, \dots$  be i.i.d. random variables and let  $z$  be a complex number such that  $|\psi(z)| \geq 1$ , where  $\psi(z) = E(e^{zY_1})$ . Let  $S_n = Y_1 + \dots + Y_n$ . Then  $\{e^{zS_n}/(\psi(z))^n, n \geq 1\}$  is a martingale with mean 1. Moreover, if  $N$  is a stopping time such that  $\max_{n \leq N} |\mathcal{R}(e^{zS_n})|$  is a bounded random variable, where  $\mathcal{R}(\cdot)$  denotes the real part, then

$$E \left\{ e^{zS_N} / (\psi(z))^N \right\} = 1 \quad (2.3)$$

by the optional stopping theorem for martingales. In the case of real  $z \neq 0$  so that  $\psi(z)$  is the moment generating function of  $Y_1$ , Bahadur (1958) subsequently showed that the left-hand side of (2.3) is equal to  $Q(N < \infty)$  if  $\psi(z) < \infty$ , where  $Q$  is the probability measure under which  $Y_1, Y_2, \dots$  are i.i.d. with common density function  $e^{zy}/\psi(z)$  with respect to the original probability measure.

Another tool Wald (1945) developed to analyze the SPRT was Wald’s equation

$$E \left( \sum_{i=1}^N Y_i \right) = \mu E(N) \quad (2.4)$$

for any stopping time  $N$  and i.i.d. random variables  $Y_i$  with mean  $\mu$ . Doob (1953) derived this result in Section VII.10 by applying the optional stopping

theorem to the martingale  $\{S_n - n\mu, n \geq 1\}$ . Chow, Robbins and Teicher (1965) subsequently made use of martingale theory to analyze the higher moments  $E(\sum_{i=1}^N Y_i)^r$  for  $r = 2, 3, 4$ . Noting that the likelihood ratio statistics  $L_n, n \geq 1$ , form a martingale with mean 1 under  $P_0$ , Doob (1953, Sect. VII.9) used the martingale convergence theorem to show that  $L_n$  converges a.s.  $[P_0]$  (almost surely, or with probability 1, under  $P_0$ ). This martingale property, and therefore also the martingale convergence theorem, are in fact applicable to dependent  $X_i$ , with joint density function  $f_n$  for  $X_1, \dots, X_n$ , so that the likelihood ratio now takes the form

$$L_n = q_n(X_1, \dots, X_n)/p_n(X_1, \dots, X_n), \quad (2.5)$$

where  $f_n = q_n$  under  $H_1$  and  $f_n = p_n$  under  $H_0$ . Doob showed that the a.s. limit of  $L_n$  (under  $H_0$ ) is 0 when the  $X_i$  are i.i.d. except for the case  $P_0\{f_1(X_1) = f_0(X_1)\} = 1$ , or equivalently,  $L_n = 1$  a.s.  $[P_0]$ .

Although Wald (1945) developed from scratch tools to analyze the SPRT, his approach was essentially of “martingale-type”. Alternative approaches that were used subsequently include analytic methods based on the strong Markov property and the fluctuation theory of random walks; see Ghosh (1970), Siegmund (1985) and Woodroffe (1982). Doob noticed the martingale structure in Wald’s work, and Chapter VII of his 1953 classic laid the foundations for the martingale approach to the analysis of randomly stopped sums and other statistics. Lai (2004) gives a survey of likelihood ratio identities and related methods in sequential analysis that have been developed on the foundations laid down by Wald and Doob.

## 2.2 Stochastic approximation

The post-war years between Wald’s (1945) fundamental paper and that of Robbins and Monro (1951) were a fast-growing period for Statistics as an academic discipline in the United States. New departments and programs in Statistics were springing up during this period, beginning in 1946 with the University of North Carolina that lured Hotelling from Columbia to start a Department of Statistics at Chapel Hill, and immediately followed by Columbia that set up a new Department of Mathematical Statistics chaired by Wald. Hotelling recruited Herbert Robbins, who claimed that he “knew nothing about statistics” at that time as he had been trained as a topologist at Harvard, to “teach measure theory, probability, analytic methods, etc. to

the department’s graduate students” (Page, 1984, p. 11). At Chapel Hill, Robbins “attended seminars and got to know several very eminent statisticians”, and soon “began to get some idea about what was going on” in statistics, in which he then “became really interested” and started daringly original research projects. This is the background behind his highly innovative work with graduate student Sutton Monro on stochastic approximation, which opened up a new direction for sequential analysis at that time.

The Robbins–Monro paper represents a major departure from the framework of sequential analysis adopted by Wald and his contemporaries, for whom the *sequential* element of the data-generating mechanism (or *experiment*) came from a data-dependent (instead of predetermined) sample size. The sequential experiments in stochastic approximation do not have stopping times; instead they involve choosing the design levels  $x_i$  in a regression model sequentially, on the basis of past observations, so that the  $x_i$  eventually converge to some desired level. The regression model considered is of the general form

$$y_i = M(x_i) + \varepsilon_i \quad (i = 1, 2, \dots), \quad (2.6)$$

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

$$x_{n+1} = x_n - y_n/M'(x_n). \quad (2.7)$$

When errors  $\varepsilon_i$  are present, using Newton’s method (2.7) entails that

$$x_{n+1} = x_n - M(x_n)/M'(x_n) - \varepsilon_n/M'(x_n). \quad (2.8)$$

Hence, if  $x_n$  should converge to  $\theta$  so that  $M(x_n) \rightarrow 0$  and  $M'(x_n) \rightarrow M'(\theta)$ , assuming  $M$  to be smooth and to have a unique root  $\theta$  such that  $M'(\theta) \neq 0$ , then (2.8) implies that  $\varepsilon_n \rightarrow 0$ , which is not possible for many kinds of random errors  $\varepsilon_i$  (e.g., when the  $\varepsilon_i$  are i.i.d. with mean 0 and variance  $\sigma^2 > 0$ ). To dampen the effect of the errors  $\varepsilon_i$ , Robbins and Monro (1951) replaced  $1/M'(x_n)$  in (2.7) by constants that converge to 0. Specifically, assuming that

$$M(\theta) = 0, \quad \inf_{\varepsilon < |x-\theta| < 1/\varepsilon} (x - \theta)M(x) > 0 \text{ for all } 0 < \varepsilon < 1, \quad (2.9)$$

$$|M(x)| \leq c(|x - \theta| + 1) \text{ for some } c > 0 \text{ and all } x, \quad (2.10)$$

the *Robbins–Monro scheme* is defined by the recursion

$$x_{n+1} = x_n - a_n y_n \quad (x_1 = \text{initial guess of } \theta), \quad (2.11)$$

where  $a_n$  are positive constants such that

$$\sum_1^{\infty} a_n^2 < \infty, \quad \sum_1^{\infty} a_n = \infty. \quad (2.12)$$

A year later, Kiefer and Wolfowitz (1952) modified the Robbins–Monro scheme to find the maximum of the regression function  $M$  in (2.6). Here  $M'(\theta) = 0$  and the Kiefer–Wolfowitz scheme is defined by the recursion

$$x_{n+1} = x_n + a_n \Delta(x_n), \quad (2.13)$$

where at the  $n$ th stage observations  $y_n''$  and  $y_n'$  are taken at the design levels  $x_n'' = x_n + c_n$  and  $x_n' = x_n - c_n$ , respectively,  $a_n$  and  $c_n$  are positive constants, and

$$\begin{aligned} \Delta(x_n) &= (y_n'' - y_n') / 2c_n \\ &= \frac{M(x_n + c_n) - M(x_n - c_n)}{2c_n} + \frac{\varepsilon_n'' - \varepsilon_n'}{2c_n} \quad \text{by (2.6)}. \end{aligned} \quad (2.14)$$

To dampen the effect of the errors  $\varepsilon_n'$  and  $\varepsilon_n''$ , Kiefer and Wolfowitz assumed that

$$c_n \rightarrow 0, \quad \sum_1^{\infty} (a_n/c_n)^2 < \infty, \quad \sum_1^{\infty} a_n c_n < \infty \quad \text{and} \quad \sum_1^{\infty} a_n = \infty. \quad (2.15)$$

Martingale theory provides useful tools to analyze convergence properties of stochastic approximation schemes. However, because martingales were still unfamiliar to the statistical community at that time, they were not invoked in the derivation of the convergence properties and statement of the assumptions. By deriving recursions for  $E(x_{n+1} - \theta)^2$  from (2.11) or (2.13) and the assumption  $\sup_i E(\varepsilon_i^2 | x_1, \dots, x_{i-1}) \leq \sigma^2$ , Robbins and Monro (1951) and Kiefer and Wolfowitz (1952) proved that their stochastic approximation schemes converge in  $L_2$ , and therefore also in probability, to  $\theta$  which is the solution of the equation  $M(\theta) = 0$  or  $M'(\theta) = 0$ . Subsequently, Blum (1954) cited a convergence theorem for square-integrable martingales (although he did not use “martingale” terminology) to prove the a.s. convergence of the

Robbins–Monro and Kiefer–Wolfowitz schemes; he was also able to remove the assumption  $\sum_1^\infty a_n c_n < \infty$  in (2.15). Dvoretzky (1956) then proved the a.s. and  $L_2$  convergence of a general class of recursive stochastic algorithms which include the Robbins–Monro and Kiefer–Wolfowitz schemes as special cases. This result is commonly called *Dvoretzky’s approximation theorem*.

Gladyšev (1965) gave a simple proof of the a.s. convergence of the Robbins–Monro scheme by an ingenious application of Doob’s supermartingale convergence theorem, paving the way for a subsequent generalization of supermartingales by Robbins and Siegmund (1971). Let  $\{\varepsilon_i, \mathcal{F}_i, i \geq 1\}$  be a martingale difference sequence such that

$$\sup_i E(\varepsilon_i^2 | \mathcal{F}_{i-1}) < \infty \text{ a.s.} \quad (2.16)$$

Putting (2.6) into the recursion (2.11) yields a corresponding recursion for  $V_n := (x_{n+1} - \theta)^2$ . From (2.10) and the assumption that  $E(\varepsilon_i | \mathcal{F}_{i-1}) = 0$ , it then follows from this recursion for  $V_n$  that

$$E(V_n | \mathcal{F}_{n-1}) \leq (1 + 2c^2 a_n^2) V_{n-1} + a_n^2 \{2c^2 + E(\varepsilon_n^2 | \mathcal{F}_{n-1})\} - 2a_n(x_n - \theta)M(x_n), \quad (2.17)$$

which can be written in the form

$$E(V_n | \mathcal{F}_{n-1}) \leq (1 + \alpha_{n-1}) V_{n-1} + \beta_{n-1} - \gamma_{n-1}, \quad (2.18)$$

in which  $\alpha_i$ ,  $\beta_i$  and  $\gamma_i$  are nonnegative  $\mathcal{F}_i$ -measurable random variables. Robbins and Siegmund (1971) call  $V_n$  that satisfies (2.18) an *almost supermartingale*, noting that  $V_n$  is indeed a supermartingale if  $\alpha_{n-1} = \beta_{n-1} = \gamma_{n-1} = 0$ . They showed that if  $V_n$  is a nonnegative almost supermartingale, then

$$V_n \text{ converges and } \sum_1^\infty \gamma_n < \infty \text{ a.s. on } \left\{ \sum_1^\infty \alpha_i < \infty, \sum_1^\infty \beta_i < \infty \right\}. \quad (2.19)$$

They applied this result to derive the a.s. part of Dvoretzky’s approximation theorem and certain convergence results in two-person games and cluster analysis as corollaries.

Although  $V_n$  satisfying (2.18) is not a supermartingale, it can be transformed into one via

$$U_n = V_n \left/ \prod_{i=1}^{n-1} (1 + \alpha_i) - \sum_{i=1}^{n-1} \left\{ (\beta_i - \gamma_i) \left/ \prod_{j=1}^i (1 + \alpha_j) \right. \right\} \right.,$$

which is a supermartingale by (2.18). Let  $\beta'_i = \beta_i / \prod_{j=1}^i (1 + \alpha_j)$ . Although  $U_n$  need not be nonnegative, it is bounded below on the event  $\{\sum_1^\infty \beta'_i \leq k\}$  for every  $k = 1, 2, \dots$ . Therefore by Doob's supermartingale convergence theorem,  $U_n$  converges a.s. on  $\{\sum_1^\infty \alpha_i < \infty, \sum_1^\infty \beta_i < \infty\}$ . Robbins and Siegmund (1971) made use of this argument to prove (2.19). Earlier, Gladyshev (1965) used a somewhat different argument to transform (2.17) to a nonnegative supermartingale, to which he applied Doob's supermartingale convergence theorem.

Robbins and Siegmund (1971, pp. 246–249) also showed how (2.19) can be applied to prove a.s. convergence of stochastic approximation schemes in a Hilbert space, using  $V_n = \|x_{n+1} - \theta\|^2$  in this case. Stochastic approximation was an active area of research during the two decades after the seminal paper of Robbins and Monro (1951). A relatively complete theory on the convergence and asymptotic normality of multivariate stochastic approximation schemes emerged and was summarized in the monograph by Nevel'son and Has'minskiĭ (1973).

### 2.3 Mixtures of likelihood ratio martingales, power-one tests and confidence sequences

The Robbins–Siegmund paper on almost supermartingales was published during the period 1968–1974 when their research focus was the development of martingale methods for boundary crossing probabilities in sequential tests with power 1. In the case of simple hypotheses  $H_0 : f = f_0$  and  $H_1 : f = f_1$ , a one-sided SPRT with stopping rule  $\tilde{N} = \inf\{n : L_n \geq B\}$  (i.e., letting  $A = 0$  in (2.1) and rejecting  $H_0$  upon stopping) has power 1 and type I error probability  $\alpha$  if  $B$  is so chosen that  $P_0(\tilde{N} = \infty) = \alpha$ . On the other hand, for composite hypotheses of the type  $H_0 : \theta \leq 0$  versus  $H_1 : \theta > 0$  when  $f = f_\theta$ , how can power-one tests such that  $\sup_{\theta \leq 0} P_\theta(\text{Reject } H_0) \leq \alpha$  be constructed? In the case where  $X_1, X_2, \dots$  are i.i.d. random variables from an exponential family of densities  $f_\theta(x) = e^{\theta x - \psi(\theta)}$  with respect to  $P_0$  such that  $E_0 X_1 = 0$ , Darling and Robbins (1967) used the fact that  $Z_n(\theta) := e^{\theta S_n - n\psi(\theta)}$ ,  $n \geq 1$ , is a nonnegative martingale with mean 1 under  $P_0$ , where  $S_n = \sum_{i=1}^n X_i$ , to conclude from Doob's martingale inequality that for  $c_i > 1$ ,

$$\begin{aligned} & P_0 \{Z_n(\theta_i) \geq c_i \text{ for some } m_i \leq n < m_{i+1}\} \\ &= P_0 \{S_n \geq \theta_i^{-1} \log c_i + n\theta_i^{-1}\psi(\theta_i) \text{ for some } m_i \leq n < m_{i+1}\} \quad (2.20) \\ &\leq 1/c_i. \end{aligned}$$

By choosing  $m_i$ ,  $c_i$  and  $\theta_i$  suitably, they derived *iterated logarithm inequalities* of the form

$$P_0 \{S_n \geq b_n(\varepsilon) \text{ for some } n \geq 1\} \leq \varepsilon \quad (2.21)$$

for given  $\varepsilon > 0$ , where

$$b_n(\varepsilon) \sim (E_0 X_1^2)^{1/2} (2n \log \log n)^{1/2} \quad \text{as } n \rightarrow \infty. \quad (2.22)$$

In 1968, in collaboration with David Siegmund who was his former Ph.D. student, Robbins came up with a much simpler construction of  $b_n(\varepsilon)$  that satisfies (2.21) and (2.22). It was the year when he and I both came to Columbia; I arrived as a new graduate student and he returned after spending the previous three years at different universities that tried to lure him away, including University of Michigan where he collaborated with Darling. He had moved in 1953 from Chapel Hill to Columbia to chair the Department of Mathematical Statistics after Wald died in a plane crash in 1950. In 1969, after a year of coursework, I began thinking about research and was attracted to the recent work of Robbins and Siegmund after hearing about it that summer through Robbins' Wald Lectures at the annual meeting of the Institute of Mathematical Statistics in New York City. At around the same time Siegmund was moving from Stanford to Columbia, and I took the earliest opportunity to ask him to be my thesis advisor as I wanted to work on problems related to his exciting project with Robbins.

Instead of letting  $\theta$  and  $c$  vary with  $n$  as in (2.20), Robbins and Siegmund integrated  $Z_n(\theta)$  with respect to a probability measure on  $\theta$ , noting that  $\int_0^\infty Z_n(\theta) dF(\theta)$  is also a nonnegative martingale with mean 1 for any probability distribution  $F$  on  $(0, \infty)$  and that

$$\int_0^\infty Z_n(\theta) dF(\theta) \geq c \iff S_n \geq \beta_F(n, c) \quad (2.23)$$

for  $c > 0$ , where  $x = \beta_F(n, c)$  is the unique positive solution of  $\int_0^\infty e^{\theta x - n\psi(\theta)} dF(\theta) = c$ . Therefore Doob's martingale inequality again yields

$$\begin{aligned} & P_0 \{S_n \geq \beta_F(n, c) \text{ for some } n \geq 1\} \\ &= P_0 \left\{ \int_0^\infty Z_n(\theta) dF(\theta) \geq c \text{ for some } n \geq 1 \right\} \leq c^{-1}. \end{aligned} \quad (2.24)$$

They also showed how  $F$  can be chosen so that the boundary  $\beta_F(n, c)$  has the iterated logarithm growth in (2.22). They applied and refined this idea

in a series of papers during the period 1968–1970; see the references cited in Robbins (1970), which was based on his Wald Lectures.

A natural problem related to (2.24), which only provides an upper bound for the boundary crossing probability of the sample sums  $S_n$ , is whether the bound is sharp enough for statistical applications of the type described in Robbins (1970). Note that  $L_n := \int e^{\theta S_n - n\psi(\theta)} dF(\theta)$  is the likelihood ratio statistic (2.5) for testing  $H_0 : P = P_0$  versus  $H_1 : P = Q$ , where  $Q$  is the probability measure under which  $(X_1, \dots, X_n)$  has joint density  $\int \prod_{i=1}^n f_\theta(x_i) dF(\theta)$ . Hence the likelihood ratio identity and Wald's argument described in Section 2.1 can still be used to show that (2.24) is sharp in the sense that it becomes an equality after ignoring the overshoot  $l_{N(c)} - \log c$ , where  $l_n = \log L_n$  and  $N(c) = \inf\{n : l_n \geq \log c\}$ . To improve these Wald-type approximations, one should correct for the overshoot in

$$E_Q(L_{N(c)} 1_{\{N(c) < \infty\}}) = c^{-1} \int E_\theta(e^{(l_{N(c)} - \log c)} 1_{\{N(c) < \infty\}}) dF(\theta). \quad (2.25)$$

Because  $l_n$  is a nonlinear function of the random walk  $S_n$ , conventional renewal-theoretic methods to analyze overshoots could not be applied. This led to the development of a nonlinear renewal theory by Lai and Siegmund (1977), who used it to derive an asymptotic approximation for (2.25) and for the type I error  $P_0(S_n \geq a\sqrt{n}$  for some  $n_0 \leq n \leq n_1$ ) of a repeated significance test considered by Robbins (1952). Nonlinear renewal theory has since become a standard tool in sequential analysis; see Woodroffe (1982, 1991) and Siegmund (1985).

The overshoot problem disappears if one replaces  $S_n$  by a continuous process, e.g., Brownian motion. More generally, Robbins and Siegmund (1970) proved the following result for continuous martingales. Let  $\varepsilon > 0$  and let  $\{Z_t, \mathcal{F}_t, t \geq a\}$  be a nonnegative martingale with continuous sample paths on  $\{Z_a < \varepsilon\}$  and such that  $Z_t 1_{\{\sup_{s>a} Z_s\}} \xrightarrow{P} 0$  as  $t \rightarrow 0$ . Then

$$P\left\{\sup_{t>a} Z_t \geq \varepsilon \mid \mathcal{F}_a\right\} = Z_a/\varepsilon \text{ a.s.} \quad \text{on } \{Z_a < \varepsilon\}. \quad (2.26)$$

Consequently,  $P\{\sup_{t \geq a} Z_t \geq \varepsilon\} = P\{Z_a \geq \varepsilon\} + \varepsilon^{-1}E(Z_a 1_{\{Z_a < \varepsilon\}})$ . Applying this result to  $Z_t = f(W_t + b, t + h)$ , where  $W_t$  is Brownian motion and

$$f(x, t) = \int_0^\infty e^{\theta x - \theta^2 t/2} dF(\theta), \quad (2.27)$$

they showed that for any  $b \in \mathbb{R}$ ,  $h \geq 0$  and  $a > 0$ ,

$$\begin{aligned} & P \{f(W_t + b, t + h) \geq \varepsilon \text{ for some } t \geq a\} \\ &= P \{f(W_a + b, a + h) \geq \varepsilon\} \\ &+ \frac{1}{\varepsilon} \int_0^\infty \exp\left(b\theta - \frac{h}{2}\theta^2\right) \Phi\left(\frac{\beta_F(a + h, \varepsilon) - b}{\sqrt{a}} - \sqrt{a}\theta\right) dF(\theta), \end{aligned} \quad (2.28)$$

where  $\Phi$  is the standard normal distribution function and  $\beta_F(t, \varepsilon) = \inf\{x : f(x, t) \geq \varepsilon\}$ .

Since (2.26) holds for  $Z_t = f(W_t, t)$  that is a nonnegative continuous martingale, one may wonder if the special form (2.27) of  $f$  used in (2.28) is too restrictive. Robbins and Siegmund (1973) gave a definitive answer to this question and in this connection also provided a probabilistic proof of the integral representations, introduced by Widder (1944), of positive solutions of the heat equation. They showed that the following statements are equivalent for any continuous  $f : \mathbb{R} \times (0, \infty) \rightarrow [0, \infty)$ :

$$\partial f / \partial t + \frac{1}{2} \partial^2 f / \partial x^2 = 0 \text{ on } \mathbb{R} \times (0, \infty), \quad (2.29a)$$

$$f(x, t) = \int_{-\infty}^\infty e^{\theta x - \theta^2 t / 2} dF(\theta) \text{ for all } x \in \mathbb{R}, t > 0 \text{ and some measure } F, \quad (2.29b)$$

$$f(W_t, t), t \geq 0, \text{ is a martingale.} \quad (2.29c)$$

Since Widder (1953) had also established integral representations of positive solutions of the heat equation on the half-line  $x > 0$  (semi-infinite rod), Robbins and Siegmund (1973) considered extensions of their result to Brownian motion with reflecting barrier at 0 and to the radial part of 3-dimensional Brownian motion (Bessel process), noting that Brownian motion is recurrent in dimensions 1 and 2 but transient in higher dimensions. They showed that in this case, (2.29c) has to be replaced by

$$f(r_{t \wedge T_a}, t \wedge T_a), t \geq 0, \text{ is a martingale for every } a > 0, \quad (2.29c')$$

where  $r_t$  is either reflected Brownian motion or the Bessel process and  $T_a = \inf\{t : r_t \leq a\}$ . The integral representation (2.29b) takes a different form here: it is a sum of two integrals with respect to measures  $F_1$  on the time axis  $[0, \infty)$  and  $F_2$  on the space axis  $(0, \infty)$ . In the case of reflected Brownian motion, (2.29a) is the time-reversed heat equation on  $(0, \infty) \times (0, \infty)$ , and takes the form  $\partial f / \partial t + \mathcal{A}f = 0$  for the Bessel process, whose infinitesimal generator  $\mathcal{A}$  is given by  $\mathcal{A}f(x) = \frac{1}{2}f''(x) + x^{-1}f'(x)$ ,  $x > 0$ .

Lai (1973) and Sawyer (1974/75) generalized these results on the Bessel process to a more general continuous Markov process  $X_t$  on an interval  $I$  with endpoints  $r_0$  and  $r_1$ , where  $-\infty \leq r_0 < r_1 \leq \infty$ . Let  $\mathcal{A}$  be the infinitesimal generator of  $X_t$ . Then  $\partial/\partial t + \mathcal{A}$  is the infinitesimal generator of the space-time process  $(t, X_t)$ . Suppose  $f : I \times [0, \infty) \rightarrow \mathbb{R}$  satisfies  $(\partial/\partial t + \mathcal{A})f(x, t) = 0$  for  $r_0 < x < r_1$  and  $t > 0$ , which is an extension of (2.29a). Lai (1973) studied the analog of (2.29c), providing conditions on the boundaries  $r_0$  and  $r_1$  under which  $f(X_t, t)$ ,  $t \geq 0$ , is a martingale. As an analog of (2.29b), Sawyer (1974/75) derived integral representations of nonnegative weak solutions of  $(\partial/\partial t + \mathcal{A})f$ , thereby generalizing the Robbins–Siegmund representation described in the preceding paragraph.

As an alternative to mixture likelihood ratios, Robbins and Siegmund (1972, 1974) introduced adaptive likelihood ratio statistics of the form

$$\tilde{L}_n = \prod_{i=1}^n \left[ f_{\hat{\theta}_{i-1}}(X_i) / f_{\theta_0}(X_i) \right] \quad (2.30)$$

to construct power-one tests of  $H_0 : \theta \leq \theta_0$  versus  $H_1 : \theta > \theta_0$  for the parameter  $\theta$  of an exponential family  $f_\theta(x) = e^{\theta x - \psi(\theta)}$ , where  $\hat{\theta}_{i-1} \geq \theta_0$  is an estimate (e.g., by constrained maximum likelihood) of  $\theta$  based on  $X_1, \dots, X_{i-1}$ . Note that  $\hat{\theta}_{i-1}$  is measurable with respect to the  $\sigma$ -field  $\mathcal{F}_{i-1}$  generated by  $X_1, \dots, X_{i-1}$  while  $X_i$  is independent of  $\mathcal{F}_{i-1}$ . Hence  $\{\tilde{L}_n, n \geq 1\}$  is still a nonnegative martingale under  $P_{\theta_0}$  and therefore Doob's inequality can be applied as in (2.24) to ensure that  $P_\theta\{N_\alpha < \infty\} \leq P_{\theta_0}\{N_\alpha < \infty\} \leq \alpha$  for  $\theta \leq \theta_0$ , where  $N_\alpha = \inf\{n : \tilde{L}_n \geq \alpha^{-1}\}$ . Robbins and Siegmund (1974) showed how  $\hat{\theta}_{i-1}$  can be chosen so that  $E_\theta N_\alpha$  attains Farrell's (1964) asymptotic lower bound, as  $\theta \downarrow \theta_0$ , for  $E_\theta T$  subject to the constraint  $P_{\theta_0}(T < \infty) \leq \alpha$ . Lai (1977) developed a theory of power-one tests of the parameter  $\theta$  of a one-parameter exponential family, based on the sequence of sample sums  $S_n$  which are sufficient statistics for  $\theta$ . Taking  $\theta_0 = 0$  without loss of generality, Lai (1977) made use of Wald's equation and likelihood ratio identities to show that for  $T_b = \inf\{n \geq n_0 : S_n \geq b(n)\}$ ,

$$\lim_{\theta \downarrow 0} E_\theta T_b / g(\mu_\theta) = P_0(T = \infty), \quad (2.31)$$

where  $b(\cdot)$  is a continuous upper-class boundary satisfying certain regularity conditions,  $t = g(\theta)$  is the solution of  $\theta t = b(t)$ , and  $\mu_\theta = E_\theta X_1 = \psi'(\theta)$ . In particular, for an upper-class boundary  $b$  such that  $b(t) \sim (E_0 X_1^2)^{1/2} (2t \log \log t)^{1/2}$  as  $t \rightarrow \infty$ , the stopping rule  $T_b$  attains Farrell's (1964) lower bound.

Robbins' revolutionary idea of terminating a test only when there is enough evidence against the null hypothesis and his theory of power-one tests were described by Neyman (1971) as "a remarkable achievement." Even though practical constraints on time and resources have rendered open-ended tests infeasible in practice, this achievement in statistical theory paved the way for subsequent breakthroughs. In particular, Lorden's (1971) seminal work on the theory of control charts and change-point detection involves the following connection between the stopping time  $N$  of a sequential detection rule and an open-ended test  $\tau$ : Let  $\tau$  be a stopping time based on i.i.d. random variables  $X_1, X_2, \dots$ , such that  $P(\tau < \infty) \leq \alpha$ . For  $k = 1, 2, \dots$ , let  $N_k$  denote the stopping time obtained by applying  $\tau$  to  $X_k, X_{k+1}, \dots$  and let  $N = \min_{k \geq 1} (N_k + k - 1)$ . Then  $N$  is a stopping time and  $EN \geq 1/\alpha$ . This allows one to derive the properties of a sequential detection rule from those of its associated power-one test, as was done by Lorden (1971) in relating the CUSUM rule to the one-sided SPRT.

Let  $X_1, X_2, \dots$  be i.i.d. random variables whose common distribution depends on an unknown parameter  $\theta \in \Theta$ . A sequence of confidence sets  $\Gamma_n = \Gamma_n(X_1, \dots, X_n)$  is called a  $(1 - \alpha)$ -level *confidence sequence* if

$$P_\theta\{\theta \in \Gamma_n \text{ for all } n \geq 1\} \geq 1 - \alpha \text{ for all } \theta \in \Theta. \quad (2.32)$$

Darling and Robbins (1967) introduced this concept and related it to the boundary crossing probabilities developed in that paper by using martingale inequalities. Lai (1976) showed that for an exponential family with parameter  $\theta$ , the Robbins–Siegmund method of mixture likelihood ratio martingales leads to a confidence sequence of intervals which have the desirable property of eventually shrinking to  $\theta$  if the mixing distribution  $F$  is so chosen that  $F(I) > 0$  for every open interval  $I$  contained in the natural parameter space  $\Theta$ . He also used invariance with respect to transformation groups to handle nuisance parameters, thereby constructing invariant confidence sequences. Subsequent applications of confidence sequences to data monitoring in clinical trials were introduced independently by Jennison and Turnbull (1984) and Lai (1984).

## 2.4 Other related developments

We now summarize other important developments in the interplay between martingale theory and sequential analysis during this period. The first is the

theory of optimal stopping and its applications to sequential analysis. The SPRT with stopping rule (2.1) was shown by Wald and Wolfowitz (1948) to be optimal in the sense that it minimizes both  $E_0(T)$  and  $E_1(T)$  among all tests whose sample size  $T$  has a finite expectation under both  $H_0 : f = f_0$  and  $H_1 : f = f_1$ , and whose error probabilities satisfy  $P_0\{\text{Reject } H_0\} \leq \alpha$  and  $P_1\{\text{Reject } H_1\} \leq \beta$ . This had been conjectured by Wald (1945) who developed lower bounds on the expected sample size of  $T$ , which are attained by the SPRT ignoring overshoots. Wald and Wolfowitz showed that the SPRT is a Bayes procedure, assuming a prior probability  $\pi$  in favor of  $H_0$  and a cost  $c$  for each observation. Subsequently, Arrow, Blackwell and Girshick (1949) recognized the optimal stopping aspect of the problem. Snell (1952) made use of martingale theory to establish the existence of optimal stopping rules in general settings and to characterize the optimal values given the  $\sigma$ -field  $\mathcal{F}_n$  generated by observations up to time  $n$ . This work was based on his Ph.D. thesis at the University of Illinois under Doob's supervision, and laid the foundations for rapid development of the subject in the 1960s, culminating in the monographs by Chernoff (1972), Chow, Robbins and Siegmund (1971), Dynkin and Yushkevich (1969) and Shiryaev (1969).

A second area of active development during this period consists of functional central limit theorems for martingales and their applications to sequential analysis. While Billingsley (1961), Brown (1971), Dvoretzky (1972) and McLeish (1974) proved central limit theorems for martingales and also used martingale inequalities to prove tightness for weak convergence, the embedding of martingales in Brownian motion by Dubins and Schwarz (1965) and Strassen (1967) provided a more direct approach to deriving functional central limit theorems (also called *invariance principles*), as shown by Freedman (1971). Approximating sums of weakly dependent variables by martingales led to corresponding central limit theorems for these sums, as in Gordin's (1969) proof of the central limit theorem for stationary sequences. Strassen's embedding result was actually developed in the context of strong approximations (or *almost sure invariance principles*), yielding a negligibly small (in the almost sure sense) error for the problem at hand. Making use of these strong approximations, Heyde (1973) and Heyde and Scott (1973) proved laws of the iterated logarithm for martingales and sums of stationary sequences, Jain, Jogdeo and Stout (1975) obtained integral tests for upper- and lower-class boundaries for martingales and mixing sequences, and Philipp and Stout (1975) extended Gordin's method to derive strong approximations of sums

of weakly dependent random variables by martingales.

Applications of functional central limit theorems and strong invariance principles in sequential analysis during this period were mostly related to nonparametric sequential estimation and testing based on rank statistics,  $U$ -statistics and linear combinations of order statistics. Sen's (1981) monograph gives a review of these developments and their connections to invariance principles and strong approximations. Berk's (1966) proof of the strong consistency of a  $U$ -statistic as an unbiased estimator of  $Eg(X_1, \dots, X_m)$  made use of the martingale convergence theorem applied to reverse martingales; see also Doob (1953, pp. 341–342) and Sen (1981, p. 51).

### 3 From Sequential Analysis to Time Series

An important research direction in sequential analysis and related martingale limit theorems during the decade 1976–1985 was associated with continuous-time processes. The two-volume monograph by Liptser and Shiryaev (1977, 1978) describes developments in nonlinear filtering and other sequential estimation problems for stochastic processes. The functional central limit theorems for stochastic integrals and semimartingales, developed by Jacod, Kłopotowski and Mémin (1982), Liptser and Shiryaev (Liptser and Širjaev, 1980) and Rebolledo (1980), found immediate applications. One such application was to survival analysis and is reviewed in the paper by Aalen et al. in this issue.

Section 2 of Lai (2001), which gives a survey of the theory and applications of sequential tests of composite hypotheses, shows the emergence of a unified complete theory in this period. In particular, the survey points out a historic event occurring outside the research community of sequential analysis and opening up a new direction for the field:

“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 (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. . . . On the other hand, BHAT. . . 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 success story of BHAT led to rapid development of time-sequential methods for censored survival data in the next 10 years; see the review in Lai (2001, pp. 312–315). Besides advances in time-sequential survival analysis, which was closely related to continuous-time martingale theory (see Sellke and Siegmund, 1983 and Section V.6 and Appendix 3 of Siegmund, 1985) this period also witnessed a unified treatment of Bayes tests of composite hypotheses via likelihood ratio martingales and optimal stopping (see Lai, 2001, pp. 306–308). In fact, most of the results in Lai (1988a,b) were already obtained by 1983 and were presented in a number of seminars and conferences around that time, but I did not write them up because my research focus at that time shifted from sequential analysis to time series and stochastic adaptive control. The following subsections will review the developments in these two areas during the period, and the important role that martingale theory had played in these developments.

### 3.1 Adaptive stochastic approximation and the multi-period control problem

Robbins was on sabbatical leave from Columbia in the academic year 1975–76, visiting the University of London. His former Columbia colleague T. W. Anderson, who had moved to Stanford in 1966, was also spending a sabbatical leave in London. Anderson had recently finished a paper with John Taylor, his former Ph.D. student in economics at Stanford, on the following “multi-period control problem” in econometrics: How should inputs  $u_i$  in the regression model  $y_i = \alpha + \beta u_i + \epsilon_i$ , with unknown parameters  $\alpha$  and  $\beta$  and i.i.d. disturbances  $\epsilon_i$  having mean 0 and variance  $\sigma^2$ , be chosen sequentially so that the outputs  $y_i$  are as close as possible in some sense to a given target value  $y^*$ ? Anderson and Taylor (1976) proposed the following certainty-equivalence rule: If  $\alpha$  and  $\beta (\neq 0)$  are both known,  $u_i$  can be optimally set at  $\theta = (y^* - \alpha)/\beta$ . Without assuming  $\alpha$  and  $\beta$  to be known, suppose that bounds  $K_1, K_2$  are given such that  $K_1 < \theta < K_2$ . Assuming

the  $\epsilon_i$  to be normally distributed, the maximum likelihood estimator of  $\theta$  at stage  $t \geq 2$  is

$$\hat{\theta}_t = K_2 \wedge \left\{ \hat{\beta}_t^{-1} (y^* - \hat{\alpha}_t) \vee K_1 \right\}, \quad (3.1)$$

where  $\hat{\beta}_t = \{\sum_1^t (u_i - \bar{u}_t) y_i\} / \{\sum_1^t (u_i - \bar{u}_t)^2\}$ ,  $\hat{\alpha}_t = \bar{y}_t - \hat{\beta}_t \bar{u}_t$  are the least squares estimates of  $\beta$  and  $\alpha$ , and  $\bar{u}_t = t^{-1} \sum_1^t u_i$ . The initial values  $u_1$  and  $u_2$  are distinct but otherwise arbitrary numbers between  $K_1$  and  $K_2$ , and for  $t \geq 2$ , the certainty-equivalence rule sets  $u_{t+1} = \hat{\theta}_t$ . Based on the results of simulation studies, Anderson and Taylor (1976) conjectured that the certainty-equivalence rule converges to  $\theta$  a.s. and that  $\sqrt{t}(\hat{\theta}_t - \theta)$  has a limiting  $N(0, \sigma^2/\beta^2)$  distribution. They also raised the question whether  $\hat{\alpha}_t$  and  $\hat{\beta}_t$  are strongly consistent. Anderson mentioned this problem to Robbins when they met in London. After he returned to Columbia at the end of his sabbatical leave, Robbins posed the problem to me and also suggested trying stochastic approximation as an alternative to the certainty-equivalence rule.

To address the Anderson–Taylor conjecture, we first noted that if the  $x_i$  should cluster around  $\theta$ , then there would not be much information for estimating the slope  $\beta$ . There is, therefore, an apparent dilemma between the control objective of setting the design levels as close as possible to  $\theta$  and the need for an informative design with sufficient dispersion to estimate  $\beta$ . To resolve this dilemma, we began by considering the case of known  $\beta$ . Replacing  $y_i$  by  $y_i - y^*$ , it can be assumed without loss of generality that  $y^* = 0$  so that  $y_i = \beta(x_i - \theta) + \epsilon_i$ . With known  $\beta$ , the least squares certainty-equivalence rule becomes  $x_{n+1} = \bar{x}_n - \bar{y}_n/\beta$ , which turns out to be equivalent to the stochastic approximation scheme (2.11) with  $a_n = (n\beta)^{-1}$ . Since  $\bar{x}_n - \bar{y}_n/\beta = \theta - \bar{\epsilon}_n/\beta$ ,  $E(x_{n+1} - \theta)^2 = \sigma^2/(n\beta^2)$  for  $n \geq 1$  and therefore

$$\begin{aligned} E \left( \sum_{n=1}^N y_n^2 \right) &= \sum_{n=1}^N E \{ \beta^2 (x_n - \theta)^2 + \epsilon_n^2 \} = N\sigma^2 + \beta^2 \sum_{n=0}^{N-1} E(x_{n+1} - \theta)^2 \\ &= \sigma^2 (N + \log N + O(1)). \end{aligned}$$

We applied martingale theory to show that

$$\beta^2 \sum_{n=1}^N (x_n - \theta)^2 \sim \sigma^2 \log N \text{ a.s.} \quad (3.2)$$

Using dynamic programming after putting a prior distribution on  $\theta$ , we found that the optimal control rule when the  $\epsilon_i$  are normal also has expected regret

$\sigma^2 \log N + O(1)$ , where  $\beta^2 \sum_{n=1}^N (x_n - \theta)^2 [= \sum_{n=1}^N (y_n - \epsilon_n)^2]$  is called the *regret* (due to ignorance of  $\theta$ ) of the design; see Lai and Robbins (1982a). Thus, for normally distributed errors, (3.2) shows that the least squares certainty-equivalence rule when  $\beta$  is known yields both asymptotically minimal regret and an efficient final estimate. The next step, therefore, was to try also to achieve this even when  $\beta$  is unknown. We accomplished this in Lai and Robbins (1979, 1981), where we introduced *adaptive* stochastic approximation schemes of the type

$$x_{n+1} = x_n - y_n / (nb_n), \quad (3.3)$$

under the assumptions (2.9) and (2.10) on the regression function  $M$  and also assuming that  $M'(\theta) = \beta > 0$ . The  $b_n$  in (3.3) is assumed to be  $\mathcal{F}_{n-1}$ -measurable and such that  $b_n \rightarrow \beta$  a.s. Not only were these schemes shown to provide asymptotically efficient estimates of  $\theta$  but their regret was also shown to achieve the asymptotically minimal order  $\sigma^2 \log N$ . Martingale theory provided us with important tools to prove these asymptotically optimal properties of adaptive stochastic approximation.

With hindsight, it was lucky that we began this research project by considering adaptive stochastic approximation rather than the certainty-equivalence rule of Anderson and Taylor (1976) whose conjecture had led to this project. After deriving a relatively complete theory for adaptive stochastic approximation, we returned to the Anderson–Taylor paper and found a counter-example to the conjecture. Strictly speaking, (3.1) is undefined when  $\hat{\beta}_n = 0$ , but this is a zero-probability event in the simulation studies of Anderson and Taylor involving continuous (normal)  $\epsilon_n$ . In Lai and Robbins (1982b), we showed that (3.1) does not converge a.s. to  $\theta$  by exhibiting an event which has positive probability and on which  $x_n$  gets stuck at one of the endpoints  $K_1, K_2$  for  $n \geq 2$ . We also established in Lai and Robbins (1982a,b) the asymptotic optimality of *iterated least squares* schemes of the form  $x_{n+1} = \bar{x}_n - \bar{y}_n / b_n$ , in which  $b_n$  is a truncated least squares estimate of  $\beta$  when *a priori* upper and lower bounds  $B_1$  and  $B_2$  of  $\beta$ , having the same sign as  $\beta$ , are known. Unlike adaptive stochastic approximation that has an intrinsic martingale structure, martingale theory is not directly applicable to iterated least squares, which therefore was considerably harder to analyze than adaptive stochastic approximation.

### 3.2 Martingales in stochastic regression and sequential experimental design

A key step in our analysis of adaptive stochastic approximation and iterated least squares procedures was to establish strong consistency of the least squares estimate

$$\hat{\beta}_n = \frac{\sum_{i=1}^n (x_i - \bar{x}_n) y_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \beta + \frac{\sum_{i=1}^n (x_i - \bar{x}_n) \epsilon_i}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

In Lai and Ying (2006), which was written in memory of Ching-Zong Wei who was my Ph.D. student at Columbia from 1977 to 1980, I described how Robbins and I tackled this problem and Wei's subsequent involvement in the project that led to the papers Lai, Robbins and Wei (1978, 1979) and Chen, Lai and Wei (1981) on the strong consistency of least squares estimates in multiple regression models  $y_i = \theta^T x_i + \epsilon_i$ , under the minimal assumption

$$\lambda_{\min} \left( \sum_{i=1}^n \psi_i \psi_i^T \right) \longrightarrow \infty \quad (3.4)$$

on the design vectors  $\psi_i$  when they are nonrandom, as in the traditional Gauss–Markov model; the notation  $\lambda_{\min}(\cdot)$  (or  $\lambda_{\max}(\cdot)$ ) is used to denote the minimum (or maximum) eigenvalue of a nonnegative definite matrix. However, when the  $\psi_i$  are sequentially determined random vectors, Lai and Robbins (1981) had already shown that in the simple linear regression model with  $\psi_i = (1, x_i)^T$ , (3.4) does not ensure strong consistency of the least squares estimate  $\hat{\theta}_n$ . In fact, they gave a counter-example in which (3.4) holds but  $\hat{\theta}_n$  does not converge to  $\theta$  a.s. Using Chow's (1965) strong law for martingales and a recursive representation (of the Kalman filter type) of least squares estimates, Lai and Wei (1982) proved the following fundamental result on strong consistency of least squares estimates in *stochastic regression models* of the form  $y_i = \theta^T \psi_i + \epsilon_i$ , in which  $\{\epsilon_i, \mathcal{F}_i, i \geq 1\}$  is a martingale difference sequence and  $\psi_i$  is  $\mathcal{F}_{i-1}$ -measurable: If  $\sup_i E(|\epsilon_i|^r | \mathcal{F}_{i-1}) < \infty$  a.s. for some  $r > 2$ , then

$$\hat{\theta}_n \rightarrow \theta \text{ a.s. on } \left\{ \lambda_{\min} \left( \sum_{i=1}^n \psi_i \psi_i^T \right) / \log \lambda_{\max} \left( \sum_{i=1}^n \psi_i \psi_i^T \right) \rightarrow \infty \right\}. \quad (3.5)$$

It occurred to me several years afterwards (see Lai, 1989) that the arguments used to prove (3.5) in Lai and Wei (1982) could be linked to Robbins

and Siegmund's almost supermartingales described in Section 2.2. Suppose that (2.18) is modified to

$$V_n \leq (1 + \alpha_{n-1})V_{n-1} + \beta_n - \gamma_n + w_{n-1}\epsilon_n \text{ a.s.}, \quad (3.6)$$

where  $\alpha_n, \beta_n, \gamma_n, V_n$  are nonnegative  $\mathcal{F}_n$ -measurable random variables such that  $\sum \alpha_n < \infty$  a.s. and  $w_n$  is  $\mathcal{F}_n$ -measurable. Then, for every  $\delta > 0$ ,

$$\max \left( V_n, \sum_{i=1}^n \gamma_i \right) = O \left( \sum_{i=1}^n \beta_i + \left( \sum_{i=1}^{n-1} w_i^2 \right)^{\frac{1}{2} + \delta} \right) \text{ a.s.} \quad (3.7)$$

Moreover,  $V_n$  converges and  $\sum E(\gamma_n | \mathcal{F}_{n-1}) < \infty$  a.s. on  $\{\sum E(\beta_n | \mathcal{F}_{n-1}) < \infty\}$ . The convergence of  $V_n$  on  $\{\sum E(\beta_n | \mathcal{F}_{n-1}) < \infty\}$  basically follows by the same argument as that leading to (2.19) upon taking conditional expectation with respect to  $\mathcal{F}_{n-1}$  in (3.6). On the other hand, on  $\{\sum E(\beta_n | \mathcal{F}_{n-1}) = \infty\}$ ,  $V_n$  need not converge and (3.7) provides a bound on the order of magnitude of  $V_n$  and  $\sum_{i=1}^n \gamma_i$ . The function  $V_n = \|x_{n+1} - \theta\|^2$  used by Robbins and Siegmund (1971) is closely related to Liapounov functions in the stability theory of ordinary differential equations (ODEs); see Lai (1989). Analogous to Liapounov functions in ODEs, the *stochastic Liapounov function*  $V_n$  inherits an almost supermartingale structure (2.18) from the dynamics of the original stochastic system. The idea behind Lai's (1989) *extended stochastic Liapounov functions* (3.6) was to achieve greater flexibility by not insisting on the almost supermartingale property that guarantees convergence. In fact, Lai and Wei's (1982) proof of (3.5) basically amounts to these arguments with

$$V_n = (\hat{\theta}_n - \theta)^T \left( \sum_{i=1}^n \psi_i \psi_i^T \right) (\hat{\theta}_n - \theta),$$

$$\gamma_n = \left\{ (\hat{\theta}_{n-1} - \theta)^T \psi_n \right\}^2 \left\{ 1 - \psi_n^T \left( \sum_{i=1}^n \psi_i \psi_i^T \right)^{-1} \psi_n \right\}.$$

The extension to the more general framework (3.6) in Lai (1989) was motivated by a unified treatment of stochastic approximation, least squares and other recursive estimates in the control engineering literature that will be described in the next subsection.

Lai and Wei (1982) also made use of martingale central limit theorems to prove the asymptotic normality of  $\hat{\theta}_n$ . This application of martingale central limit theorems was subsequently used by Ford, Titterton and Wu (1985), Wu (1985a,b), Chaudhuri and Mykland (1993), and Lai (1994) to address some open problems in nonlinear experimental designs and other sequential designs.

### 3.3 Time series analysis, forecasting and control

The title of this subsection is the same (except for minor punctuation differences) as that of the influential textbook by Box and Jenkins (1976). A preliminary edition of that book was published when I was assigned to teach a master's level course on time series at Columbia in 1974, and I used it as a required text for the course. What I liked most about the book was its applications to prediction and control in engineering. Although they were beyond the scope of the one-semester course, they stimulated my interest in the subject and provided me with some subject-matter background when I worked on stochastic adaptive control problems in the 1980s.

I have summarized in Section 2.3 of Lai and Ying (2006) the developments in the engineering literature on adaptive control of linear input/output of the form

$$A(q^{-1})y_n = B(q^{-1})u_{n-1} + C(q^{-1})\epsilon_n \quad (3.8)$$

that attracted me and Wei to work in this area after we obtained the key results in Lai and Wei (1982), around the same time when Goodwin, Ramadge and Caines (1981) provided a major advance in the subject by using stochastic approximation to circumvent the difficulties in the analysis of least squares (or extended least squares) estimates in a feedback control environment. The  $y_n$ ,  $u_n$  and  $\epsilon_n$  in (3.8) denote the output, input and random disturbance at time  $n$ , respectively, and

$$\begin{aligned} A(q^{-1}) &= 1 + a_1q^{-1} + \cdots + a_pq^{-p}, \\ B(q^{-1}) &= b_1 + \cdots + b_kq^{-(k-1)}, \\ C(q^{-1}) &= 1 + \cdots + c_hq^{-h} \end{aligned}$$

are polynomials in the backward shift operator  $q^{-1}$ . The adaptive control problem is to determine the inputs  $u_t$ , based on current and past observations  $y_t, y_{t-1}, u_{t-1}, \dots$ , to keep the outputs  $y_{t+1}$  as close as possible to certain

target values  $y_{t+1}^*$  when the system parameters are unknown and have to be estimated sequentially. If the parameters were known, then one could choose  $\mathcal{F}_t$ -measurable  $u_t$  such that  $E(y_{t+1}|\mathcal{F}_t) = y_{t+1}^*$ , yielding the optimal output  $y_{t+1}^* + \epsilon_{t+1}$ . This suggests defining the regret

$$R_n = \sum_{i=1}^n (y_i - y_i^* - \epsilon_i)^2, \quad (3.9)$$

which is an extension of  $\sum_{i=1}^n (y_i - \epsilon_i)^2$  considered in Section 3.1 for the multi-period control problem in the linear regression model with  $y^* = 0$ . My subsequent work with Wei in Lai and Wei (1986, 1987) showed how to make use of (3.7) to construct modifications of certainty-equivalence rules (also called *self-tuning rules* in the adaptive control literature) by using least squares (or extended least squares) estimates of the parameters in (3.8) to attain a logarithmic order for the regret, i.e.,  $R_n = O(\log n)$  a.s., similar to adaptive stochastic approximation reviewed in Section 3.1; see Lai and Ying (2006, pp. 751–753).

One of the centers of time series research during this period was the statistics group at Australian National University (ANU). The ANU group can be credited as the first to apply systematically martingale theory to time series analysis, dating back to Hannan and Heyde (1972), Heyde and Seneta (1972) and Heyde (1972, 1973). The monograph by Hall and Heyde (1980) gives a review of the martingale applications to estimation theory in time series and branching processes in the 1970s. Subsequent representative work by the group in this direction during the period under review includes Solo (1979), Hannan (1980a,b), Hannan and Rissanen (1982), and An, Chen and Hannan (1982).

Martingale theory also played an important role in addressing some open problems in nonstationary time series during this period, paving the way for major advances in the field in the next five years that will be described in the next section. For the AR( $p$ ) model  $y_n = \beta_1 y_{n-1} + \dots + \beta_p y_{n+p} + \epsilon_n$  with i.i.d. zero-mean random disturbances  $\epsilon_n$  satisfying certain moment conditions, weak consistency of the least squares estimates had been proved by tedious computations of moments of certain linear and quadratic forms involving the observations. These results, under various assumptions on the roots of the characteristic polynomial  $\varphi(z) = z^p - \beta_1 z^{p-1} - \dots - \beta_p$ , were unified by Stigum (1974) who established weak consistency without any assumptions on the roots of  $\varphi(z)$ . By making use of the general results of Lai

and Wei (1982) and a linear transformation of  $Y_n = (y_n, y_{n-1}, \dots, y_{n-p+1})^T$  that corresponds to factorizing  $\varphi$  as a product of a non-explosive polynomial (with zeros on or inside the unit circle) and an explosive one (whose zeros lie outside the unit circle), Lai and Wei (1983) proved strong consistency of the least squares estimate in general AR( $p$ ) models in which the random disturbances form a martingale difference sequence with  $\sup_n E(|\epsilon_n|^r | \mathcal{F}_{n-1}) < \infty$  a.s. for some  $r > 2$ , thereby solving a long-standing problem in the literature.

A special case of nonstationary autoregressive models that has attracted much attention in economics is unit-root nonstationarity ( $\beta = 1$ ) associated with the AR(1) model  $y_n = \beta y_{n-1} + \epsilon_n$ . The test developed by Dickey and Fuller (1979) during this period, for testing the null hypothesis of unit root versus the alternative hypothesis  $|\beta| < 1$ , has become widely used in econometric time series. Other than the Dickey–Fuller test, inference in non-explosive AR(1) models was relatively unexplored because of technical difficulties that Fuller explained in a seminar at Stanford in the early 1980s. Siegmund, who had moved from Columbia to Stanford in 1976, had attended Fuller’s seminar and summarized it for me when we met at a conference in Heidelberg, in which I talked about my recent work with Wei on general AR( $p$ ) models. The main issue is that the least squares estimate  $\hat{\beta}_n$  is asymptotically normal if  $|\beta| < 1$  whereas its limiting distribution after Studentization is highly non-Gaussian if  $|\beta| = 1$ , causing great difficulty in constructing large-sample confidence intervals for  $\beta$  in non-explosive but possibly nonstationary AR(1) models. After some discussion, we decided to circumvent this difficulty by using, instead of a fixed sample size  $n$ , the stopping rule

$$N_c = \inf \left\{ n \geq 1 : \sum_{i=1}^n y_i^2 \geq c \right\}. \quad (3.10)$$

In Lai and Siegmund (1983), we proved the following uniform asymptotic normality property of  $\hat{\beta}_{N_c}$  as  $c \rightarrow \infty$ :

$$\sup_{|\beta| \leq 1, x \in \mathbb{R}} \left| P \left\{ \sqrt{c}(\hat{\beta}_{N_c} - \beta)/\sigma \leq x \right\} - \Phi(x) \right| \longrightarrow 0. \quad (3.11)$$

Since  $\hat{\sigma}_n^2 := n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_n y_{i-1})^2$  is a consistent estimate of  $\sigma^2$ ,  $\hat{\beta}_{N_c} \pm c^{-1/2} \hat{\sigma}_{N_c} \Phi^{-1}(1 - \alpha)$  is an approximate  $(1 - 2\alpha)$ -level confidence interval for

$\beta$  in non-explosive AR(1) models. Moreover, letting

$$N(d) = \inf \left\{ n \geq 1 : (\hat{\sigma}_n \vee n^{-1/2}) \Phi^{-1}(1 - \alpha) \left( \sum_{i=1}^n y_{i-1}^2 \right)^{-1/2} \leq d \right\}, \quad (3.12)$$

$\hat{\beta}_{N(d)} \pm d$  is an approximate  $(1 - 2\alpha)$ -level confidence interval for  $\beta$ , with fixed width  $2d \rightarrow 0$ . Fixed-width confidence intervals based on stopping rules of the type (3.12) were first proposed by Chow and Robbins (1965) for a population mean when the variance  $\sigma^2$  is unknown. Letting  $\bar{Y}_n$  denote the sample mean and  $\hat{\sigma}_n^2$  the sample variance based on a sample of size  $n$ , define  $\tilde{N}(d) = \inf\{n \geq n_0 : (\hat{\sigma}_n \vee n^{-1/2})\Phi^{-1}(1 - \alpha) \leq d\sqrt{n}\}$ . The Chow–Robbins approximate  $(1 - 2\alpha)$ -level confidence interval for the population mean is  $\bar{Y}_{\tilde{N}(d)} \pm d$ . Although the stopping rule (3.12) clearly uses the same idea as that of Chow and Robbins, the analysis of  $\hat{\beta}_{N(d)} \pm d$  is considerably harder than that of Chow and Robbins and relies heavily on the martingale structure of  $\sum_{i=2}^n y_{i-1}\epsilon_i$ . Martingale theory features prominently in the analysis in Lai and Siegmund (1983, pp. 480–482).

## 4 Epilogue

Inspired by Lai and Siegmund (1983), Wei and his Ph.D. student Ngai Hang Chan undertook the study of the behavior of  $\hat{\beta}_n$  in the AR(1) model whose  $\beta$  approaches 1 at a rate depending on  $n$ . Specifically, Chan and Wei (1987) made use of martingale functional central limit theorems to show that for  $\beta = 1 - \gamma/n$ , with  $\gamma$  being a fixed positive constant, the Studentized statistic

$$\left( \sum_{i=1}^n y_{i-1}^2 \right)^{\frac{1}{2}} (\hat{\beta}_n - \beta) = \left( \sum_{i=1}^n y_{i-1}\epsilon_i \right) / \left( \sum_{i=1}^n y_{i-1}^2 \right)^{\frac{1}{2}} \quad (4.1)$$

converges in distribution to a non-normal random variable  $Y(\gamma)$ , with  $Y(\gamma)$  approaching standard normal as  $\gamma \rightarrow \infty$ .

The period 1986–1991 witnessed major advances in multivariate time series with unit-root nonstationarity. Representative works include Chan and Wei (1988), Johansen (1988, 1991) and Phillips (1988, 1991). Martingale theory was extensively used in these works and eventually became standard material in econometric time series, as evidenced by Hamilton’s (1994) popular

textbook on the subject. The same can be said for stochastic adaptive control in linear dynamic systems; see the monographs by Caines (1988), Kumar and Varaiya (1986) and the paper of Guo and Chen (1991) that solved a long-standing problem (Åström and Wittenmark, 1973) on certainty-equivalence-type rules in linear dynamic systems, similar to the Anderson–Taylor problem for the linear regression model.

As a result, the level of exposure to martingale theory in the Ph.D. curriculum in Statistics was much greater in 1987, the year I moved from Columbia to Stanford after Robbins’ retirement (in 1986), than in 1968 when I began my graduate study at Columbia. However, Columbia was somewhat unique in those days because three of the seven faculty members of the department, Chow (who had been Doob’s student at Illinois), Robbins and Siegmund, were working together in martingales and sequential analysis. Therefore I had the good fortune to learn the subject from my teachers as they were still developing it, instead of from systematic courses and textbooks presenting a fully-developed theory. It was even more fortunate that I could join their team and participate in this development. Some of what was produced in those days continued to benefit me in seemingly unrelated problems many years later, and I will conclude by mentioning a relatively recent example.

To begin, note that (4.1) is a *Studentized statistic*, whose name came from Gosset’s (1908) paper on the  $t$ -statistic published under the name ‘Student’. The  $t$ -statistic is prototypical of a large class of *self-normalized* processes of the form  $A_n/B_n$ , with  $A_n = \sum_{i=1}^n X_i$  and  $B_n^2 = \sum_{i=1}^n X_i^2$  in the case of the  $t$ -statistic

$$\frac{\sqrt{n}\bar{X}_n}{\left\{\sum_{i=1}^n (X_i - \bar{X}_n)^2 / (n - 1)\right\}^{1/2}} = \frac{A_n}{B_n} \left\{ \frac{n - 1}{n - (A_n/B_n)^2} \right\}^{\frac{1}{2}}.$$

Active development of the probability theory of self-normalized processes began in the 1990s, first for self-normalized sums of i.i.d. random variables belonging to the domain of attraction of a stable law and then more generally for martingales self-normalized by their quadratic or predictable variation processes; see de la Peña, Lai and Shao (2009). Chapter 11 of that book points out the connection between the method of mixtures that we have described in Section 2.3 and the general theory of self-normalized processes, developed in de la Peña, Klass and Lai (2004) under the canonical assumption

that

$$\left\{ e^{\theta A_t - \theta^2 B_t^2 / 2}, \mathcal{F}_t, t \in T \right\} \text{ is a supermartingale with mean } \leq 1 \quad (4.2)$$

for all  $0 \leq \theta < \theta_0$ , where  $T$  is either  $\{1, 2, \dots\}$  or  $[0, \infty)$ , or under variants thereof. Our key observation of de la Peña et al. (2004) was that

$$\frac{A_t^2}{2B_t^2} = \max_{\theta} \left( \theta A_t - \frac{\theta^2 B_t^2}{2} \right).$$

Although maximizing the supermartingale (4.2) over  $\theta$  would not yield a supermartingale and the maximum may also occur outside the range  $0 \leq \theta < \theta_0$ , integrating the supermartingale with respect to the measure  $f(\lambda)d\lambda$  still preserves the supermartingale property. This is the essence of the Robbins–Siegmund method of mixtures for  $A_t = W_t$  and  $B_t = \sqrt{t}$  that we have reviewed in Section 2.3.

## References

- An, H. Z., Chen, Z. G. and Hannan, E. J. (1982). Autocorrelation, autoregression and autoregressive approximation. *Ann. Statist.* 10: 926–936.
- Anderson, T. W. and Taylor, J. B. (1976). Strong consistency of least squares estimates in normal linear regression. *Ann. Statist.* 4: 788–790.
- Arrow, K. J., Blackwell, D. and Girshick, M. A. (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica* 17: 213–244.
- Åström, K. J. and Wittenmark, B. (1973). On self-tuning regulators. *Automatica—J. IFAC* 9: 195–199.
- Bahadur, R. R. (1958). A note on the fundamental identity of sequential analysis. *Ann. Math. Statist.* 29: 534–543.
- Berk, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist.* 37: 745–746, correction: *ibid*, 51–58.
- Billingsley, P. (1961). The Lindeberg–Lévy theorem for martingales. *Proc. Amer. Math. Soc.* 12: 788–792.

- Blackwell, D. and Girshick, M. A. (1946). On functions of sequences of independent chance vectors with applications to the problem of the “random walk” in  $k$  dimensions. *Ann. Math. Statistics* 17: 310–317.
- Blum, J. R. (1954). Multidimensional stochastic approximation methods. *Ann. Math. Statistics* 25: 737–744.
- Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco, Calif.: Holden-Day, revised ed., holden-Day Series in Time Series Analysis.
- Brown, B. M. (1971). Martingale central limit theorems. *Ann. Math. Statist.* 42: 59–66.
- Caines, P. E. (1988). *Linear Stochastic Systems*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.
- Chan, N. H. and Wei, C. Z. (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Ann. Statist.* 15: 1050–1063.
- Chan, N. H. and Wei, C. Z. (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Ann. Statist.* 16: 367–401.
- Chaudhuri, P. and Mykland, P. A. (1993). Nonlinear experiments: Optimal design and inference based on likelihood. *J. Amer. Statist. Assoc.* 88: 538–546.
- Chen, G. J., Lai, T. L. and Wei, C. Z. (1981). Convergence systems and strong consistency of least squares estimates in regression models. *J. Multivariate Anal.* 11: 319–333.
- Chernoff, H. (1972). *Sequential Analysis and Optimal Design*. Philadelphia, Pa.: Society for Industrial and Applied Mathematics, conference Board of the Mathematical Sciences Regional Conference Series in Applied Mathematics, No. 8.
- Chow, Y. S. (1965). Local convergence of martingales and the law of large numbers. *Ann. Math. Statist.* 36: 552–558.

- 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.
- Chow, Y. S., Robbins, H. and Siegmund, D. (1971). *Great Expectations: The Theory of Optimal Stopping*. Boston, Mass.: Houghton Mifflin Co.
- Chow, Y. S., Robbins, H. and Teicher, H. (1965). Moments of randomly stopped sums. *Ann. Math. Statist.* 36: 789–799.
- Darling, D. A. and Robbins, H. (1967). Iterated logarithm inequalities. *Proc. Nat. Acad. Sci. U.S.A.* 57: 1188–1192.
- de la Peña, V. H., Klass, M. J. and Lai, T. L. (2004). Self-normalized processes: Exponential inequalities, moment bounds and iterated logarithm laws. *Ann. Probab.* 32: 1902–1933.
- de la Peña, V. H., Lai, T. L. and Shao, Q.-M. (2009). *Self-Normalized Processes: Limit Theory and Statistical Applications*. Probability and Its Applications. Berlin Heidelberg: Springer-Verlag.
- Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* 74: 427–431.
- Doob, J. L. (1953). *Stochastic Processes*. New York: John Wiley & Sons Inc.
- Dubins, L. E. and Schwarz, G. (1965). On continuous martingales. *Proc. Nat. Acad. Sci. U.S.A.* 53: 913–916.
- Dvoretzky, A. (1956). On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I*. Berkeley and Los Angeles: University of California Press, 39–55.
- Dvoretzky, A. (1972). Asymptotic normality for sums of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, CA, 1970/1971), Vol. II: Probability Theory*. Berkeley, Calif.: Univ. California Press, 513–535.

- Dynkin, E. B. and Yushkevich, A. A. (1969). *Markov Processes: Theorems and Problems*. Translated from the Russian by James S. Wood. New York: Plenum Press.
- Farrell, R. H. (1964). Asymptotic behavior of expected sample size in certain one sided tests. *Ann. Math. Statist.* 35: 36–72.
- Ford, I., Titterington, D. M. and Wu, C.-F. J. (1985). Inference and sequential design. *Biometrika* 72: 545–551.
- Freedman, D. (1971). *Brownian Motion and Diffusion*. San Francisco, Calif.: Holden-Day.
- Ghosh, B. K. (1970). *Sequential Tests of Statistical Hypotheses*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont.
- Gladyšev, E. G. (1965). On stochastic approximation. *Teor. Veroyatnost. i Primenen.* 10: 297–300.
- Goodwin, G. C., Ramadge, P. J. and Caines, P. E. (1981). Discrete time stochastic adaptive control. *SIAM J. Control Optim.* 19: 829–853.
- Gordin, M. I. (1969). The central limit theorem for stationary processes. *Dokl. Akad. Nauk SSSR* 188: 739–741.
- Gosset, W. S. (1908). On the probable error of a mean. *Biometrika* 6: 1–25.
- Guo, L. and Chen, H. F. (1991). The Åström–Wittenmark self-tuning regulator revisited and ELS-based adaptive trackers. *IEEE Trans. Automat. Control* 36: 802–812.
- Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. New York: Academic Press Inc. [Harcourt Brace Jovanovich Publishers], probability and Mathematical Statistics.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press.
- Hannan, E. J. (1980a). The estimation of the order of an ARMA process. *Ann. Statist.* 8: 1071–1081.

- Hannan, E. J. (1980b). Recursive estimation based on ARMA models. *Ann. Statist.* 8: 762–777.
- Hannan, E. J. and Heyde, C. C. (1972). On limit theorems for quadratic functions of discrete time series. *Ann. Math. Statist.* 43: 2058–2066.
- Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika* 69: 81–94.
- Heyde, C. C. (1972). Martingales: A case for a place in the statistician’s repertoire. *Austral. J. Statist.* 14: 1–9.
- Heyde, C. C. (1973). An iterated logarithm result for martingales and its application in estimation theory for autoregressive processes. *J. Appl. Probability* 10: 146–157.
- Heyde, C. C. and Scott, D. J. (1973). Invariance principles for the law of the iterated logarithm for martingales and processes with stationary increments. *Ann. Probability* 1: 428–436.
- Heyde, C. C. and Seneta, E. (1972). Estimation theory for growth and immigration rates in a multiplicative process. *J. Appl. Probability* 9: 235–256.
- Jacod, J., Kłopotowski, A. and Mémin, J. (1982). Théorème de la limite centrale et convergence fonctionnelle vers un processus à accroissements indépendants: La méthode des martingales. *Ann. Inst. H. Poincaré Sect. B (N.S.)* 18: 1–45.
- Jain, N. C., Jogdeo, K. and Stout, W. F. (1975). Upper and lower functions for martingales and mixing processes. *Ann. Probability* 3: 119–145.
- Jennison, C. and Turnbull, B. W. (1984). Repeated confidence intervals for group sequential clinical trials. *Controlled Clinical Trials* 5: 33–45.
- Johansen, S. (1988). Statistical analysis of cointegration vectors. *J. Econom. Dynam. Control* 12: 231–254, Economic time series with random walk and other nonstationary components.
- Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* 59: 1551–1580.

- Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Statistics* 23: 462–466.
- Kumar, P. R. and Varaiya, P. (1986). *Stochastic Systems: Estimation, Identification and Adaptive Control*. Englewood Cliffs, NJ: Prentice Hall.
- Lai, T. L. (1973). Space-time processes, parabolic functions and one-dimensional diffusions. *Trans. Amer. Math. Soc.* 175: 409–438.
- Lai, T. L. (1976). On confidence sequences. *Ann. Statist.* 4: 265–280.
- Lai, T. L. (1977). Power-one tests based on sample sums. *Ann. Statist.* 5: 866–880.
- Lai, T. L. (1984). Incorporating scientific, ethical and economic considerations into the design of clinical trials in the pharmaceutical industry: A sequential approach. *Comm. Statist. Theory Methods* 13: 2355–2368.
- Lai, T. L. (1988a). Nearly optimal sequential tests of composite hypotheses. *Ann. Statist.* 16: 856–886.
- Lai, T. L. (1988b). On Bayes sequential tests. In *Statistical decision theory and related topics, IV, Vol. 2 (West Lafayette, Ind., 1986)*. New York: Springer, 131–143.
- Lai, T. L. (1989). Extended stochastic Lyapunov functions and recursive algorithms in linear stochastic systems. In *Stochastic differential systems (Bad Honnef, 1988)*. Berlin: Springer, Lecture Notes in Control and Inform. Sci. 126, 206–220.
- Lai, T. L. (1994). Asymptotic properties of nonlinear least squares estimates in stochastic regression models. *Ann. Statist.* 22: 1917–1930.
- Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statist. Sinica* 11: 303–408, with comments and a rejoinder by the author.
- Lai, T. L. (2004). Likelihood ratio identities and their applications to sequential analysis. *Sequential Anal.* 23: 467–497.
- Lai, T. L. and Robbins, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* 7: 1196–1221.

- Lai, T. L. and Robbins, H. (1981). Consistency and asymptotic efficiency of slope estimates in stochastic approximation schemes. *Z. Wahrsch. Verw. Gebiete* 56: 329–360.
- Lai, T. L. and Robbins, H. (1982a). Adaptive design and the multiperiod control problem. In *Statistical decision theory and related topics, III, Vol. 2 (West Lafayette, Ind., 1981)*. New York: Academic Press, 103–120.
- Lai, T. L. and Robbins, H. (1982b). Iterated least squares in multiperiod control. *Adv. in Appl. Math.* 3: 50–73.
- Lai, T. L., Robbins, H. and Wei, C. Z. (1978). Strong consistency of least squares estimates in multiple regression. *Proc. Nat. Acad. Sci. U.S.A.* 75: 3034–3036.
- Lai, T. L., Robbins, H. and Wei, C. Z. (1979). Strong consistency of least squares estimates in multiple regression. II. *J. Multivariate Anal.* 9: 343–361.
- Lai, T. L. and Siegmund, D. (1977). A nonlinear renewal theory with applications to sequential analysis. I. *Ann. Statist.* 5: 946–954.
- Lai, T. L. and Siegmund, D. (1983). Fixed accuracy estimation of an autoregressive parameter. *Ann. Statist.* 11: 478–485.
- 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 Wei, C. Z. (1983). Asymptotic properties of general autoregressive models and strong consistency of least-squares estimates of their parameters. *J. Multivariate Anal.* 13: 1–23.
- 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. Control* 31: 898–906.
- Lai, T. L. and Wei, C. Z. (1987). Asymptotically efficient self-tuning regulators. *SIAM J. Control Optim.* 25: 466–481.

- Lai, T. L. and Ying, Z. (2006). Efficient recursive estimation and adaptive control in stochastic regression and ARMAX models. *Statist. Sinica* 16: 741–772.
- Lipcer, R. Š. and Širjaev, A. N. (1980). A functional central limit theorem for semimartingales. *Teor. Veroyatnost. i Primenen.* 25: 683–703.
- Liptser, R. S. and Shiryaev, A. N. (1977). *Statistics of Random Processes. I.* New York: Springer-Verlag, general theory, Translated by A. B. Aries, Applications of Mathematics, Vol. 5.
- Liptser, R. S. and Shiryaev, A. N. (1978). *Statistics of Random Processes. II.* New York: Springer-Verlag, applications, Translated from the Russian by A. B. Aries, Applications of Mathematics, Vol. 6.
- Lorden, G. (1971). Procedures for reacting to a change in distribution. *Ann. Math. Statist.* 42: 1897–1908.
- McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probability* 2: 620–628.
- Nevel'son, M. B. and Has'minskiĭ, R. Z. (1973). *Stochastic Approximation and Recursive Estimation.* Providence, R. I.: American Mathematical Society, translated from the Russian by the Israel Program for Scientific Translations, Translations of Mathematical Monographs, Vol. 47.
- Neyman, J. (1971). Foundations of behavioristic statistics. In *Foundations of Statistical Inference (Proc. Sympos., Univ. Waterloo, Waterloo, Ontario, 1970)*. Rinehart and Winston of Canada, Toronto, Ont.: Holt, 1–19, with comments by G. A. Barnard, D. J. Bartholomew, I. J. Good and O. Kempthorne and a reply by the author.
- Page, W. (1984). An interview with Herbert Robbins. *College Math. J.* 15: 2–24.
- Philipp, W. and Stout, W. (1975). Almost sure invariance principles for partial sums of weakly dependent random variables. *Mem. Amer. Math. Soc.* 2 issue 2: iv+140.
- Phillips, P. C. B. (1988). Weak convergence of sample covariance matrices to stochastic integrals via martingale approximations. *Econometric Theory* 4: 528–533.

- Phillips, P. C. B. (1991). Optimal inference in cointegrated systems. *Econometrica* 59: 283–306.
- Rebolledo, R. (1980). Central limit theorems for local martingales. *Z. Wahrsch. Verw. Gebiete* 51: 269–286.
- Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* 58: 527–535.
- 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. Statistics* 22: 400–407.
- 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. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics (Proc. Sympos., Ohio State Univ., Columbus, OH, 1971)*. New York: Academic Press, 233–257.
- Robbins, H. and Siegmund, D. (1972). A class of stopping rules for testing parametric hypotheses. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, CA, 1970/1971), Vol. IV: Biology and Health*. Berkeley, Calif.: Univ. California Press, 37–41.
- Robbins, H. and Siegmund, D. (1973). Statistical tests of power one and the integral representation of solutions of certain partial differential equations. *Bull. Inst. Math. Acad. Sinica* 1: 93–120.
- Robbins, H. and Siegmund, D. (1974). The expected sample size of some tests of power one. *Ann. Statist.* 2: 415–436, collection of articles dedicated to Jerzy Neyman on his 80th birthday.
- Rudin, W. (1997). *The Way I Remember It, History of Mathematics 12*. Providence, RI: American Mathematical Society.
- Sawyer, S. (1974/75). A Fatou theorem for the general one-dimensional parabolic equation. *Indiana Univ. Math. J.* 24: 451–498.

- Sellke, T. and Siegmund, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* 70: 315–326.
- Sen, P. K. (1981). *Sequential Nonparametrics*. New York: John Wiley & Sons Inc., invariance principles and statistical inference, Wiley Series in Probability and Mathematical Statistics.
- Shiryaev, A. N. (1969). *Statisticheskii Posledovatelnyi Analiz: Optimalnye Pravila Ostanovki [Statistical Sequential Analysis: Optimal Stopping Rules]*. Izdat. “Nauka”, Moscow.
- Siegmund, D. (1985). *Sequential Analysis*. Springer Series in Statistics. New York: Springer-Verlag.
- Snell, J. L. (1952). Applications of martingale system theorems. *Trans. Amer. Math. Soc.* 73: 293–312.
- Solo, V. (1979). Convergence of AML. *IEEE Trans. Automat. Contr.* 24: 958–962.
- Stigum, B. P. (1974). Asymptotic properties of dynamic stochastic parameter estimates. III. *J. Multivariate Anal.* 4: 351–381.
- Strassen, V. (1967). Almost sure behavior of sums of independent random variables and martingales. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, CA, 1965/66)*. Berkeley, Calif.: Univ. California Press, 315–343, Vol. II: Contributions to Probability Theory, Part 1.
- Wald, A. (1944). On cumulative sums of random variables. *Ann. Math. Statistics* 15: 283–296.
- Wald, A. (1945). Sequential tests of statistical hypotheses. *Ann. Math. Statistics* 16: 117–186.
- Wald, A. and Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Ann. Math. Statistics* 19: 326–339.
- Wallis, W. A. (1980). The Statistical Research Group, 1942–1945. *J. Amer. Statist. Assoc.* 75: 320–335, with comments by F. J. Anscombe and William H. Kruskal and a reply by the author.

- Widder, D. V. (1944). Positive temperatures on an infinite rod. *Trans. Amer. Math. Soc.* 55: 85–95.
- Widder, D. V. (1953). Positive temperatures on a semi-infinite rod. *Trans. Amer. Math. Soc.* 75: 510–525.
- Woodroffe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*, CBMS-NSF Regional Conference Series in Applied Mathematics 39. Philadelphia, Pa.: Society for Industrial and Applied Mathematics (SIAM).
- Woodroffe, M. (1991). The role of renewal theory in sequential analysis. In *Handbook of Sequential Analysis*. New York: Dekker, Statist. Textbooks Monogr. 118, 145–167.
- Wu, C.-F. J. (1985a). Asymptotic inference from sequential design in a non-linear situation. *Biometrika* 72: 553–558.
- Wu, C.-F. J. (1985b). Efficient sequential designs with binary data. *J. Amer. Statist. Assoc.* 80: 974–984.