Statistical Science
2021, Vol. 36, No. 2, 291–302
https://doi.org/10.1214/20-STS784

Stochastic Approximation: From Statistical Origin to Big-Data, Multidisciplinary Applications

Tze Leung Lai and Hongsong Yuan

Abstract. Stochastic approximation was introduced in 1951 to provide a new theoretical framework for root finding and optimization of a regression function in the then-nascent field of statistics. This review shows how it has evolved in response to other developments in statistics, notably time series and sequential analysis, and to applications in artificial intelligence, economics and engineering. Its resurgence in the big data era has led to new advances in both theory and applications of this microcosm of statistics and data science.

Key words and phrases: Control, gradient boosting, optimization, recursive stochastic algorithms, regret, weak greedy variable selection.

1. INTRODUCTION

The year 2021 will mark the seventieth anniversary of the seminal paper of Robbins and Monro (1951) on stochastic approximation. In 1946, Herbert Robbins entered the then-nascent field of statistics somewhat serendipitously. He had enlisted in the Navy during the Second World War and was demobilized as a lieutenant commander in 1945. His interest in probability theory and mathematical statistics began during the war when he overheard conversation between two senior naval officers about the effect of random scatter of bomb impacts. Although he was prevented from pursuing the officers' problem during his service because he lacked the appropriate security clearance, his eventual work on the problem led to his fundamental papers (Robbins, 1944, 1945) in the field of geometric probability. These papers paved the way for his recruitment by Hotelling to teach "measure theory, probability, analytic methods, etc." as associate professor in the new Department of Mathematical Statistics at the University of North Carolina at Chapel Hill, even though he first thought that Hotelling had called the wrong person because he "knew nothing about statistics" (Page, 1984, pp. 8–11). During the 6 years he spent at Chapel Hill before moving to Columbia, he invented compound decision theory and empirical Bayes methods, stochastic approximation and multiarmed bandit theory, and also introduced new approaches to sequential analysis. These accomplishments and their impacts were reviewed by Efron (2003), Zhang (2003), Lai (2003) and Siegmund (2003) in a memorial issue of the *Annals of Statistics* after his death in 2002. They complemented the review, by Lai and Siegmund (1986), of Robbins' work and its impact up to 1984.

Sutton Monro was born in Burlington, Vermont, in 1914, about a year before Robbins. He received B.S. and M.S. degrees from MIT. He enlisted in the Navy during the Second World War and taught mathematics at the University of Maine at Orono from 1946 to 1948 after he was demobilized. Then he became a Ph.D. student in the Department of Mathematical Statistics at the University of North Carolina at Chapel Hill. There was substantial interest during that period in the problem of finding the maximum (or minimum) of a regression function M and choosing design levels around the optimum, beginning the work of Hotelling (1941) on "experimental attainment of optimum conditions" in polynomial regression models; see Box and Wilson (1951). Monro was interested in this problem, but after he arrived at Chapel Hill, Hotelling was no longer working in this area and, therefore, he chose Robbins to be his adviser. In Section 2, we describe Monro's work with Robbins that led to the foundational statistical theory in their 1951 paper and subsequent developments not only in statistics but also in engineering and economics. Our review of these multidisciplinary developments dating back to the 1960s show

Tze Leung Lai is Professor, Department of Statistics, Stanford University, 390 Jane Stanford Way, Stanford, California 94305, USA (e-mail: lait@stanford.edu). Hongsong Yuan is Associate Professor, School of Information Management and Engineering, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China (e-mail: yuan.hongsong@shufe.edu.cn).

how new analytical techniques, algorithms and applications have emerged and enriched stochastic approximation.

The turn of the millennium marks the onset of the big data revolution that has changed the field of statistics in both theory and practice. Section 3 revisits stochastic approximation in the big data era, in which gradient (instead of Hessian) methods play a prominent role in machine/statistical learning of the maximum of a regression function $f(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^p$ with p larger than the size of the training sample. Besides reviewing important advances in this area, Section 3 also describes some new theoretical developments and applications of stochastic approximation. There has been much recent debate on statistics versus data science and on theory versus practice in statistics in this big data era. Our discussion in Section 3.4 shows that in the case of stochastic approximation which belongs to both statistics and data science, the trajectory has been "from theory to practice and back," and the subject has remained vibrant to this day, nearly 70 years after Robbins and Monro's foundational paper.

2. STOCHASTIC APPROXIMATION AND OPTIMIZATION

2.1 The Robbins–Monro and Kiefer–Wolfowitz Recursions

To derive key insights into the underlying issues, Robbins and Monro (1951) considered univariate input x so that finding the optimum of the regression function amounts to solving the equation $M(x) = 0$, where M is the derivative of the regression function. Under the assumption that M is smooth with $M'(\theta) \neq 0$, if one uses Newton's scheme $x_{n+1} = x_n - Y_n/M'(x_n)$, then since $Y_n = M(x_n) + \epsilon_n$, where the ϵ_n represent unobservable random errors,

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

which entails that $\epsilon_n \to 0$ if x_n should converge to θ so that $M(x_n) \to 0$ and $M'(x_n) \to M'(\theta)$. Since this is not possible if ϵ_n are i.i.d. with positive variance, the Robbins–Monro scheme

$$x_{n+1} = x_n - a_n Y_n$$

uses weights $a_n > 0$ such that $\sum_{n=1}^{\infty} a_n^2 < \infty$ and $\sum_{n=1}^{\infty} a_n = \infty$ to average out the errors ϵ_n. In fact, the assumption $\sum_{n=1}^{\infty} a_n^2 < \infty$ ensures that $\sum_{n=1}^{\infty} a_n \epsilon_n$ converges in L_2 and a.s. for many stochastic models of the random errors (including i.i.d. mean-zero ϵ_n with finite variance). Under certain regularity conditions, this in turn implies that $x_n - \theta$ converges in L_2 and a.s., and the assumption $\sum_{n=1}^{\infty} a_n = \infty$ then assures that the limit of $x_n - \theta$ is 0. In their convergence analysis of the recursive scheme, Robbins and Monro (1951) transformed the recursion into a corresponding recursion for $E(x_{n+1} - \theta)^2$ and thereby proved L_2-convergence of the recursive scheme.

Kiefer and Wolfowitz (1952) subsequently used this approach to prove L_2-convergence of the recursive scheme

$$x_{n+1} = x_n - a_n\left(\frac{Y_n'' - Y_n'}{2c_n}\right)$$

to find the minimum θ of the regression function $f(x)$ (or, equivalently, the solution of $M(x) = 0$, where $M = df/dx$). During the nth stage of the Kiefer–Wolfowitz scheme, observations Y_n'' and Y_n' are taken at $x_n + c_n$ and $x_n - c_n$, respectively, where c_n and a_n are positive constants such that $\sum_{n=1}^{\infty} (a_n/c_n)^2 < \infty$ and $\sum_{n=1}^{\infty} a_n = \infty$. Such a scheme was what Monro had been working on before he rejoined the U.S. military as naval officer in the Korean War. Like Monro, Kiefer who was about 10 years younger had been interested in stochastic optimization when he began his doctoral program in mathematical statistics at Columbia. He had already written papers on sequential search schemes when he was a master's student at MIT; these papers were later refined and published as Kiefer, Kiefer (1953, 1957). He changed his thesis topic to the new area of statistical decision theory and worked with Wald, whose death in 1950 resulted in Wolfowitz being his thesis adviser. He moved with Wolfowitz to Cornell in 1951 and his Ph.D. thesis *Contributions to the Theory of Games and Statistical Decision Functions* at Columbia was completed in 1952. After the Korean War, Monro worked at Bell Labs from 1953 to 1959 and then at Lehigh University as professor of industrial engineering until his retirement in 1985. He returned to live in Burlington, Vermont after retirement, and died in 1989, 8 years after Kiefer's death.

2.2 Adaptive Stochastic Approximation

The last sentence of Robbins and Monro (1951) says: "One of us is investigating the properties of this and other sequential designs as a graduate student; the senior author is responsible for the convergence proof." Unlike his graduate student Monro, Robbins was more interested in direct applications of the root-finding Robbins–Monro algorithm than in indirect applications to regression function optimization. Robbins and Monro (1951) gave a concrete application to recursive estimation of the qth quantile θ_q of a distribution function F, for which $M(x) = F(x) - q$. They also discussed a linear regression model $M(x) = \alpha + \beta x$ with unknown regression parameters such that $\beta \neq 0$, for which $M(\theta) = y^*$ has solution $\theta = (y^* - \alpha)/\beta$, saying: "Instead of trying to estimate the (regression) parameters (by least squares), we may try to estimate the value θ such that $M(\theta) = y^*$, without any assumption about the form of $M(x)$," by using a stochastic approximation scheme "when $M(x)$ satisfies the hypothesis of (the convergence) theorem." There was little

progress in this problem until Lai and Robbins (1979) developed a comprehensive theory of adaptive stochastic approximation, which was motivated by a conjecture of Anderson and Taylor (1976) and which we describe below.

Robbins spent the 1975–1976 academic year as a Guggenheim Fellow at Imperial College in London, where he heard a lecture by Anderson on the "multi-period control problem" in econometrics, which is concerned with choosing the inputs $x_1, \dots, x_N$ sequentially in the linear regression model $Y_i = \alpha + \beta x_i + \epsilon_i$ (with unknown parameters $\beta \neq 0$ and α and i.i.d. random errors ϵ_i having mean 0 and variance σ^2) so that the outputs are as close as possible to a target value y^*. Assuming prior knowledge of bounds K_1 and K_2 such that $K_1 < \theta := (y^* - \alpha)/\beta < K_2$, the Anderson–Taylor rule is defined recursively by

$$x_{n+1} = K_1 \vee \{\widehat{\beta}_n^{-1}(y^* - \widehat{\alpha}_n) \wedge K_2\}, \quad n \geq 2,$$

where $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ are the least squares estimates of α and β at stage n. Based on the results of simulation studies, Anderson and Taylor (1976) conjectured that this rule converges to θ a.s. and that $\sqrt{n}(x_n - \theta)$ has a limiting $N(0, \sigma^2/\beta^2)$ distribution. They also raised the question whether $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ are strongly consistent. Clearly, if the x_i should cluster around θ, then there would not be much information for estimating the slope β. There is, therefore, an apparent dilemma between the control objective of setting the design levels as close as possible to θ and the need for an informative design with sufficient dispersion to estimate β.

To resolve this dilemma, Lai and Robbins (1979) began by considering the case of known β. Replacing Y_i by $Y_i - y^*$, it can be assumed without loss of generality that $y^* = 0$ so that $Y_i = \beta(x_i - \theta) + \epsilon_i$. Let $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$, $\bar{Y}_n = n^{-1}\sum_{i=1}^n Y_i$. With known β, 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 recursion $x_{n+1} = x_n - (n\beta)^{-1}Y_n$. 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,

$$E\left(\sum_{n=1}^N Y_n^2\right) = \sum_{n=1}^N E\{\beta^2(x_n - \theta)^2 + \epsilon_n^2\} = \sigma^2\big(N + \log N + O(1)\big).$$

Moreover, $\sqrt{N}(x_N - \theta) \Rightarrow N(0, \sigma^2/\beta^2)$ and $\sum_{n=1}^N (x_n - \theta)^2 \sim (\sigma^2/\beta^2)\log N$ a.s. As shown by Lai and Robbins (1982a) who used dynamic programming after putting a prior distribution on θ, the optimal control rule when the ϵ_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 θ) of the design. Hence, for normally distributed errors, this rule when β is known yields both asymptotically minimal regret and an efficient final estimate. The next step, therefore, is to try also to achieve this even when β is unknown. An obvious way to modify the preceding rule for the case of unknown β is to use an estimate b_n to substitute for β either in the recursion $x_{n+1} = \bar{x}_n - \bar{Y}_n/\beta$ or in the equivalent stochastic approximation scheme $x_{n+1} = x_n - Y_n/(n\beta)$. The equivalence between the two recursive schemes, however, no longer holds when β is replaced by b_n. The second recursion, called *adaptive stochastic approximation*, was treated in Lai and Robbins, Lai and Robbins (1979, 1981) and is described below.

Blum (1954) used martingale theory to prove a.s. convergence of the Robbins–Monro scheme to θ under the following conditions on the regression function M that are weaker than those of Robbins and Monro (1951): (a) $|M(x)| \leq c(|x - \theta| + 1)$ for all x and some $c > 0$, and (b) $\inf_{\epsilon \leq |x-\theta| \leq \epsilon^{-1}} \{M(x)(x - \theta)\} > 0$ for all $0 < \epsilon < 1$. Under these assumptions and also assuming that $M'(\theta) = \beta > 0$, Lai and Robbins (1979) consider adaptive stochastic approximation schemes of the form $x_{n+1} = x_n - Y_n/(nb_n)$, where b_n is $\mathcal{F}_{n-1}$-measurable and $\lim_{n\to\infty} b_n = b > 0$ a.s. By representing x_n as a weighted sum of the i.i.d. random variables ϵ_i, they proved limit theorems on $x_N - \theta$ and $\sum_{n=1}^N (x_n - \theta)^2$. In particular, for $0 < b < 2\beta$, they proved that

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

$$\text{(b)} \quad \sqrt{N}(x_N - \theta) \Rightarrow N\big(0, (\sigma^2/\beta^2)g(b/\beta)\big),$$

where $g(t) = 1/\{t(2 - t)\}$ for $0 < t < 2$ and has a minimum value of 1 at $t = 1$. Lai and Robbins (1981) showed that by choosing b_n to be a truncated version of the least squares estimate in the adaptive stochastic approximation scheme, one indeed has $b_n \to \beta$ a.s. and, therefore, the adaptive scheme has the same asymptotic properties as the "oracle" stochastic approximation that assumes known β. Instead of a step size of order $1/n$, Nemirovski and Yudin, Nemirovsky and Yudin (1978, 1983) and subsequently Polyak (1990) and Ruppert (1991) have proposed to take larger step sizes that are simpler to implement and yet can still attain the asymptotic efficiency result $\sqrt{N}(\bar{x}_N - \theta) \Rightarrow N(0, \sigma^2/\beta^2)$, where $\bar{x}_N = N^{-1}\sum_{n=1}^N x_n$. In other words, averaging "slowly convergent" Robbins–Monro schemes can still yield asymptotically efficient estimates of θ, as in (b) above with $b = \beta$; see Ruppert (1991), p. 515. Note, however, that these slowly convergent schemes x_n have much larger regret than the optimal order $(\sigma^2/\beta^2)\log N$ in (a) above with $b = \beta$.

Ruppert (1985) and Wei (1987) subsequently studied adaptive stochastic approximation in multivariate regression models in which $\mathbf{Y}_n$ and $\mathbf{x}_n$ belong to $\mathbb{R}^p$. An advantage of using adaptive SA (stochastic approximation)

instead of least squares for linear regression models, noted by Lai and Robbins, Ruppert and Wei, is that the procedure can be conveniently extended to nonlinear regression models $\mathbf{Y}_n = \mathbf{f}(\mathbf{x}_n) + \boldsymbol{\epsilon}_n$, where $\mathbf{f}: \mathbb{R}^p \to \mathbb{R}^p$ satisfies certain regularity conditions. Wei uses a generalized Venter estimate of the Jacobian matrix $\partial \mathbf{f}/\partial \mathbf{x}$, which requires taking $2p$ measurements $x_n - c_n, x_{nj} + c_n$ $(j = 1, \ldots, p)$ at stage n of the SA recursion. Ruppert reformulates the problem as finding the minimum of $\|\mathbf{f}\|^2$ and uses $2pm_n$ observations at stage n of the Kiefer–Wolfowitz-type recursion to estimate $\partial \mathbf{f}/\partial \mathbf{x}(\mathbf{x}_n)$ and $\lfloor n^\gamma \rfloor$ observations to estimate $\mathbf{f}(\mathbf{x}_n)$. Instead of using finite-difference approximations for multivariate SA, Spall (1992) introduced simultaneous perturbation (SP) gradient approximations, which he and his coauthors subsequently developed further in a series of papers. Whereas Ruppert's and Wei's approaches require multiple measurements, of the order of p or of higher order, at each stage of the SA recursion, Spall (2000) argues that SPSA requires only 3 gradient measurements at each stage of the recursion or 4 function measurements for gradient-free SA schemes. Noting that as with all stochastic search algorithms the performance of SA schemes depends on the choice of the tuning parameters, he develops an adaptive SPSA algorithm that is "based on the simple idea of using two parallel recursions," one for estimating the Hessian matrix, "while concurrently estimating the (tuning) parameters of interest."

2.3 Recursive Algorithms in Signal Processing and Adaptive Control

From its statistical foundation reviewed in Section 2.1, stochastic approximation flourished under multidisciplinary input and development. Sakrison (1967) and Saridis and Stein (1968) described the use of stochastic approximation for recursive system identification, which Åström and Wittenmark, Åstöm and Wittenmark (1971, 1973) further developed for adaptive control of linear stochastic systems. Ljung (1977) gave a unified convergence analysis of recursive stochastic algorithms for system identification and control, using stability analysis of an associated ordinary differential equation (ODE) which defines the "asymptotic paths" of the recursive algorithm that can only converge to the stable equilibrium points of the ODE. Kushner and Clark (1978) elaborated and refined this ODE approach in connection with earlier work by Kushner and his collaborators on stochastic approximation algorithms for constrained optimization.

An important problem in adaptive control of linear stochastic systems after the seminal paper of Åstöm and Wittenmark (1973) on self-tuning regulators was recursive identification and adaptive control in the ARX and ARMAX models. An ARX model (autoregressive model with exogenous inputs) is a linear time series of the form $A(q^{-1})Y_n = B(q^{-1})u_{n-1} + \epsilon_n$, where $A(q^{-1}) = 1 - a_1 q^{-1} - \cdots - a_p q^{-p}$, $B(q^{-1}) = b_1 + \cdots + b_r q^{-(r-1)}$ and q^{-1} is the unit delay operator (defined by $q^{-1}u_n = u_{n-1}$), and u_t represents the input while Y_t the output and ϵ_t the random disturbance at time t. An ARMAX model (in which MA stands for "moving average") is a more general model of the form $A(q^{-1})Y_n = B(q^{-1})u_{n-1} + C(q^{-1})\epsilon_n$, where $C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_k q^{-k}$ is used to model moving average disturbances. Rewriting the ARX model as a stochastic regression model $y_n = \boldsymbol{\beta}^T \mathbf{x}_n + \epsilon_n$, with $\boldsymbol{\beta} = (-a_1, \ldots, -a_p, b_1, \ldots, b_r)^T$ and $\mathbf{x}_n = (Y_{n-1}, \ldots, Y_{n-p}, u_{n-1}, \ldots, u_{n-r})^T$, the problem is similar to the multiperiod control problem in econometrics that motivates the work of Lai and Robbins (1979) on adaptive stochastic approximation summarized in Section 2.2.

Besides adaptive stochastic approximation, Lai and Robbins (1982b) also used a more direct recursion, originally proposed by Anderson and Taylor (1976) and described in the second paragraph of Section 2.2. Calling this recursion *iterated least squares*, they analyzed the strong consistency of the least squares estimate $\widehat{\boldsymbol{\beta}}_n = \{\sum_{i=1}^n (x_i - \bar{x}_n) y_i\} / \sum_{i=1}^n (x_i - \bar{x}_n)^2$ for sequentially determined x_i, as in iterated least squares which they proved not to converge to β on an event with positive probability. Lai and Wei (1982) proved the following result on strong consistency of the least squares estimate $\widehat{\boldsymbol{\beta}}_n = (\sum_{t=1}^n \mathbf{x}_t \mathbf{x}_t^T)^{-1} \sum_{t=1}^n \mathbf{x}_t y_n$ in the more general stochastic regression model $y_t = \boldsymbol{\beta}\mathbf{x}_t + \epsilon_t$, in which $\mathbf{x}_t \in \mathbb{R}^p$ is $\mathcal{F}_{t-1}$-measurable and ϵ_t is a martingale difference sequence (with respect to a filtration $\{\mathcal{F}_t\}$) such that $\sup_t E(|\epsilon_t|^{2+\delta} | \mathcal{F}_{t-1}) < \infty$ a.s.:

$$\widehat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta} \quad \text{a.s. in the event } \left\{ \frac{\lambda_{\min}(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)}{\log \lambda_{\max}(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)} \to \infty \right\}, \tag{2.1}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues, respectively. A key tool used in their proof is the recursive representation of $\widehat{\boldsymbol{\beta}}_n$:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_n &= \widehat{\boldsymbol{\beta}}_{n-1} + \boldsymbol{\Gamma}_n \mathbf{x}_n (y_n - \boldsymbol{\beta}_{n-1}^T \mathbf{x}_n), \\ \boldsymbol{\Gamma}_n &= \boldsymbol{\Gamma}_{n-1} - \boldsymbol{\Gamma}_{n-1} \mathbf{x}_n \mathbf{x}_n^T \boldsymbol{\Gamma}_{n-1} / (1 + \mathbf{x}_n^T \boldsymbol{\Gamma}_{n-1} \mathbf{x}_n), \end{aligned} \tag{2.2}$$

where $\boldsymbol{\Gamma}_n = (\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T)^{-1}$ has the above explicit recursion using the matrix inversion lemma; see Kumar and Varaiya (1986) who also show (2.2) to be a special case (corresponding to $\mathbf{V} = \mathbf{0}$) of the linear state-space model with unobservable states $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \mathbf{w}_t$, in which $\mathbf{w}_t$ are independent with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}$, and observations $y_t = \boldsymbol{\beta}_t^T \mathbf{x}_t + \epsilon_t$.

Lai (1989) generalized the aforementioned argument used to prove (2.1) into the method of *extended stochastic Liapounov functions*. A precursor of this work was the *nonnegative almost supermartingale* V_n satisfying

$$E(V_n | \mathcal{F}_{n-1}) \le (1 + \alpha_{n-1}) V_{n-1} + \beta_{n-1} - \gamma_{n-1} \tag{2.3}$$

for some nonnegative $\mathcal{F}_i$-measurable random variables V_i, α_i, β_i and γ_i such that $\sum \alpha_i + \sum \beta_i < \infty$ a.s., which was introduced by Robbins and Siegmund (1971) who showed that

$$V_n \text{ converges a.s. and } \sum_{n=1}^{\infty} \gamma_n < \infty \text{ a.s.}, \tag{2.4}$$

generalizing the convergence theorem of nonnegative supermartingales (for which $\alpha_i = \beta_i = \gamma_i = 0$). For the Robbins–Monro scheme $x_{n+1} = x_n - a_n\{M(x_n) + \epsilon_n\}$ with $|M(x)| \le c(|x-\theta| + 1)$ for all x and some $c > 0$, $\inf_{\delta \le |x-\theta| \le \delta^{-1}} \{M(x)(x-\theta)\} > 0$ for all $0 < \delta < 1$, and with martingale difference sequence ϵ_n such that $\sup_n E(\epsilon_n^2 | \mathcal{F}_{n-1}) < \infty$ a.s., (2.3) holds for $V_n := (x_{n+1} - \theta)^2$ with $\gamma_{n-1} = 2a_n M(x_n)(x_n - \theta) \ge 0$, $\alpha_{n-1} = 2c^2 a_n^2$ and $\beta_{n-1} = a_n^2\{2c^2 + E(\epsilon_n^2 | \mathcal{F}_{n-1})\}$. Hence by (2.4), $|x_n - \theta| = \sqrt{V_n}$ converges a.s. and $\sum r_{n-1} = 2\sum a_n M(x_n)(x_n - \theta) < \infty$ a.s., implying that $x_n \to \theta$ a.s. since $\sum a_n = \infty$ in the Robbins–Monro scheme. Noting that $V_n = (x_{n+1} - \theta)^2$ is closely related to the Liapounov functions in the stability theory of ordinary differential equations that feature in the ODE approach to the analysis of recursive stochastic algorithms of Ljung (1977) and Kushner and Clark (1978), Lai (1989) generalized (2.3) to

$$\begin{aligned} V_n \le (1 + \alpha_{n-1})V_{n-1} + \xi_n - \zeta_n \\ + w_{n-1}\epsilon_n \quad \text{a.s.}, \end{aligned} \tag{2.5}$$

where $V_i, \alpha_i, \xi_i, \zeta_i$ are nonnegative $\mathcal{F}_{i-1}$-measurable random variables such that $\sum \alpha_i < \infty$, w_i is $\mathcal{F}_i$-measurable and $\{\epsilon_n, n \ge 1\}$ is a martingale difference sequence satisfying $\sup_n E(\epsilon_n^2 | \mathcal{F}_{n-1}) < \infty$ a.s. Then for every $\delta > 0$,

$$\begin{aligned} &\max\left(V_n, \sum_{i=1}^{n} \zeta_i\right) \\ &\quad = O\left(\sum_{i=1}^{n} \xi_i + \left(\sum_{i=1}^{n-1} w_i^2\right)^{\frac{1}{2}+\delta}\right) \quad \text{a.s.}, \end{aligned} \tag{2.6}$$

$$\begin{aligned} &V_n \text{ converges a.s. and } \sum E(\zeta_i | \mathcal{F}_{i-1}) < \infty \\ &\quad \text{a.s. on } \Big\{\sum E(\xi_i | \mathcal{F}_{i-1}) < \infty\Big\}. \end{aligned} \tag{2.7}$$

Note that taking conditional expectation $E(\cdot | \mathcal{F}_{n-1})$ on both sides of (2.5) yields (2.3) with $\beta_{n-1} = E(\xi_n | \mathcal{F}_{n-1})$ and $\gamma_{n-1} = E(\zeta_n | \mathcal{F}_{n-1})$, hence (2.7) is a "local convergence" version of (2.4). On the other hand, V_n does not converge on $\{\sum E(\xi_i | \mathcal{F}_{i-1}) = \infty\}$ and (2.7) provides a bound on its order of magnitude and that of $\sum_{i=1}^n \zeta_i$. Lai and Wei, Lai and Wei (1982, 1986) basically used the extended stochastic Liapounov function $V_n = (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})^T \boldsymbol{\Gamma}_n^{-1} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})$ or its modification for the ELS (extended least squares) recursive algorithm to prove (2.1) for the least squares estimator in the stochastic regression model or its extension for ELS in ARMAX models that satisfy certain stability and positive real conditions on the moving average operator $C(q^{-1})$.

Widrow and Hoff (1960) introduced the LMS (least mean squares) algorithm as an alternative to the Kalman filter (2.2). It is basically a recursive stochastic gradient algorithm which uses a scalar gain sequence γ_n in lieu of the matrix gain sequence $\boldsymbol{\Gamma}_n$ in the Kalman filter. For the ARMAX model $A(q^{-1})Y_n = B(q^{-1})x_{n-1} + C(q^{-1})\epsilon_n$, Fuchs (1982) used the LMS algorithm

$$\begin{aligned} &\boldsymbol{\theta}_n = \boldsymbol{\theta}_{n-1} + (\alpha/\gamma_n)\boldsymbol{\phi}_n(Y_n - \boldsymbol{\theta}_{n-1}^T \boldsymbol{\phi}_n), \\ &\gamma_n = \gamma_{n-1} + \|\boldsymbol{\phi}_n\|^2 \end{aligned} \tag{2.8}$$

for recursive estimation of the parameter vector $\boldsymbol{\theta} = (-a_1, \ldots, -a_p, b_1, \ldots, b_r, c_1, \ldots, c_k)^T$ under stability and positive real conditions on the ARMAX model, where $\widehat{\epsilon}_t = Y_t - \boldsymbol{\theta}_{t-1}^T \boldsymbol{\phi}_t$ and $\boldsymbol{\phi}_n = (Y_{n-1}, \ldots, Y_{n-p}, x_{n-1}, \ldots, x_{n-r}, \widehat{\epsilon}_{n-1}, \ldots, \widehat{\epsilon}_{n-k})^T$ is a surrogate for $\boldsymbol{\psi}_n = (Y_{n-1}, \ldots, Y_{n-p}, x_{n-1}, \ldots, x_{n-r}, \epsilon_{n-1}, \ldots, \epsilon_{n-k})^T$. He showed that $\sum_{n=1}^{\infty} (\boldsymbol{\theta}_{n-1}^T \boldsymbol{\phi}_n - \boldsymbol{\theta}^T \boldsymbol{\psi}_n)^2 / \gamma_n < \infty$ a.s. and that the adaptive control rule $\boldsymbol{\theta}_{t-1}^T \boldsymbol{\phi}_t = y_t^*$ satisfies the "self-tuning" property

$$\begin{aligned} &n^{-1} \sum_{t=1}^{n} (Y_t - y_t^* - \epsilon_t)^2 \to 0 \quad \text{a.s. and} \\ &\quad n^{-1} \sum_{t=1}^{n} (x_t^2 + Y_t^2) = O(1) \quad \text{a.s.} \end{aligned} \tag{2.9}$$

The second property is often called "bounded average energy" for the inputs and outputs, assuming the target values and random disturbances also have bounded average energy. $\sum_{t=1}^n (Y_t - y_t^* - \epsilon_t)^2$ is called the regret of the control rule, since y_t^* is the target output and ϵ_t is the random disturbance at time t. A similar result was obtained earlier by Goodwin, Ramadge and Caines (1981) after reparameterizing the ARMAX model as $C(q^{-1})E(Y_{n+1} | \mathcal{F}_n) = G(q^{-1})Y_n + B(q^{-1})x_n$, where the polynomial division algorithm was used to give $C(z) = A(z) + zG(z)$. The self-tuning property (2.9) is considerably weaker than the order $\sigma^2 \log n$ for the regret established by Lai and Robbins, Lai and Robbins (1979, 1981) for adaptive stochastic regression. Lai and Ying, Lai and Ying (1991a, 1991b) have shown how parallel recursive algorithms can be used to develop recursive parameter estimates that are asymptotically efficient and adaptive control rules with regret of the order of $\log n$.

In practice the order of (p, r, k) of an ARMAX model is unknown and is often used to approximate on infinite-order model. Guo, Huang and Hannan (1990) considered the ARX(∞) model $Y_n = \sum_{i=1}^{\infty} (a_i Y_{n-i} + b_i x_{n-i}) + \epsilon_n$, in which $Y_t = 0$ and $x_t = 0$ for $t < 0$ and $\sum_{i=1}^{\infty} (|a_i| + |b_i|) < \infty$, as a more realistic data generating mechanism for the observed time series $\{(x_t, Y_t), 1 \le t \le$

$T\}$ than the finite-order ARX(p,r) model in the preceding paragraph, which could be used as an approximation to the model, as Hannan (1987) had considered earlier for transfer function approximations by rational functions (associated with ARMA models). The selected model ARX$(\widehat{p}_T, \widehat{r}_T)$ depends on the sample, and in particular the sample size, hence the notation $\widehat{p}_T$ and $\widehat{r}_T$ used to highlight this point. The assumption $\sum_{i=1}^{\infty}(|a_i|+|b_i|)<\infty$ is similar to the l_1-sparsity constraint on the stochastic regression model $Y_n=\boldsymbol{\beta}^T\mathbf{x}_n+\epsilon_n$, in which $\mathbf{x}_n=(x_{n,1},\ldots,x_{n,p_n})^T$ is $\mathcal{F}_{n-1}$-measurable $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_{p_n})^T$ with $\sup_n\sum_{i=1}^{p_n}|\beta_i|<\infty$, and ϵ_n is $\mathcal{F}_n$-measurable with $E(\epsilon_n|\mathcal{F}_{n-1})=0$ a.s. For $p_n\le n$ and $p_n\le t<n$, let $\mathbf{u}_{t,n}=(Y_t,\ldots,Y_{t-p_n+1},x_t,\ldots,x_{t-p_n+1})^T$ and

$$\widehat{\boldsymbol{\beta}}_{t,n}=\left(\sum_{i=0}^{t-1}\mathbf{u}_{i,n}\mathbf{u}_{i,n}^T+\gamma\mathbf{I}\right)^{-1}\sum_{i=0}^{t-1}\mathbf{u}_{i,n}y_{i+1},$$

where $\gamma>0$ is arbitrarily chosen to ensure that the matrix is invertible. Guo, Huang and Hannan (1990) actually consider multivariate $\mathbf{Y}_t$ and $\mathbf{x}_t$, in which a_i and b_i are replaced by matrices $\mathbf{A}_i$ and $\mathbf{B}_i$ with $\sum_{i=1}^{\infty}(\|\mathbf{A}_i\|+\|\mathbf{B}_i\|)<\infty$ and Y_i and x_i in the definition of $\mathbf{u}_{t,n}$ are replaced by $\mathbf{Y}_i^T$ and $\mathbf{x}_i^T$, where $\|C\|=(\lambda_{\max}(\mathbf{C}\mathbf{C}^T))^{1/2}$ is the maximum singular value of $\mathbf{C}$. Letting $\mathbf{A}(z)=\mathbf{I}-\sum_{i=1}^{\infty}\mathbf{A}_iz^i$ and $\mathbf{B}(z)=\sum_{i=1}^{\infty}\mathbf{B}_iz^i$, they consider the transfer function matrix $\mathbf{G}(z)=(\mathbf{A}(z),\mathbf{B}(z))$ of the ARX(∞) model and its estimate $\widehat{\mathbf{G}}_n(z)=(\widehat{\mathbf{A}}_n(z),\widehat{\mathbf{B}}_n(z))$, in which $\widehat{\mathbf{A}}_n(z)=\mathbf{I}-\sum_{i=1}^{p_n}\widehat{\mathbf{A}}_{i,n}z^i$, $\widehat{\mathbf{B}}_n(z)=\sum_{i=1}^{p_n}\widehat{\mathbf{B}}_{i,n}z^i$, and $\widehat{\mathbf{A}}_{i,n}$ and $\widehat{\mathbf{B}}_{i,n}$ are the component matrices of $\widehat{\boldsymbol{\beta}}_{n,n}=(\widehat{\mathbf{A}}_{1,n},\ldots,\widehat{\mathbf{A}}_{p_n,n},\widehat{\mathbf{B}}_{1,n},\ldots,\widehat{\mathbf{B}}_{p_n,n})$. They use truncation and exponential bounds for double arrays of martingales differences to derive bounds on the Hankel norm $\|\widehat{\mathbf{G}}_n-\mathbf{G}\|_\infty$, where $\|\mathbf{F}\|_\infty=\operatorname{ess\,sup}_{\theta\in[0,2\pi]}\lambda_{\max}^{1/2}(\mathbf{F}(e^{i\theta})\mathbf{F}^*(e^{i\theta}))$ and $\mathbf{F}^*(e^{i\theta})$ is the conjugate transpose of the matrix $\mathbf{F}(e^{i\theta})$ with complex entries.

3. STOCHASTIC APPROXIMATION IN THE BIG DATA ERA

3.1 High-Dimensional Sparse Linear Stochastic Regression Models

Basu and Michailidis (2015) consider a linear stochastic regression model of the form $y_t=\boldsymbol{\beta}^T\mathbf{x}_t+\epsilon_t, t=1,\ldots,n$, where $\mathbf{x}_t\in\mathbb{R}^q$ and $\epsilon_t\in\mathbb{R}$ are "independent, centered, Gaussian stationary processes". Although q can be larger than n, they assume that $\boldsymbol{\beta}$ is k-sparse, that is, $\sum_{i=1}^{q}I_{\{\beta_i\ne0\}}=k$. In addition, like Guo, Huang and Hannan (1990), they assume that the spectral density function $\mathbf{f}$ of the stationary process $\mathbf{x}_t$ exists and consider its maximum singular value $\|\mathbf{f}\|_\infty$; the spectral density function is defined by $(2\pi)^{-1}\sum_{m=-\infty}^{\infty}E\{(\mathbf{x}_0-\boldsymbol{\mu})(\mathbf{x}_m-\boldsymbol{\mu})^T\}e^{-im\theta}$, where $\boldsymbol{\mu}=E\mathbf{x}_0$. Based on observations $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$ from the stochastic regression model, they use the Lasso estimate

$$\widetilde{\boldsymbol{\beta}}_n=\arg\min_{\mathbf{b}\in\mathbb{R}^p}\sum_{t=1}^{n}(y_t-\mathbf{b}^T\mathbf{x}_t)^2+\lambda_n\sum_{j=1}^{p_n}|b_j|$$

to approximate the k-sparse parameter vector $\boldsymbol{\beta}$. Their Section 3 shows that the "restricted eigenvalue condition (commonly assumed for consistent estimation of k-sparse $\boldsymbol{\beta}$ by Lasso with appropriately chosen λ_n) holds with high probability when the sample size is sufficiently large and the process of predictors $\mathbf{x}_t$ is stable, with a full-rank spectral density," and that the concentration condition of $\sum_{t=1}^{n}(y_t-\widetilde{\boldsymbol{\beta}}_n^T\mathbf{x}_t)\mathbf{x}_t$ around $\mathbf{0}$ for consistent estimation also holds for n of larger order of magnitude than $\log p$. Their Section 4 proves analogous results of high-dimensional vector autoregressive (VAR) models. Central to this development are exponential bounds of Ravikumar et al. (2011), Peligrad et al. (2014) and Wu and Wu (2016) for high-dimensional linear models with correlated errors.

Instead of Lasso used by Basu and Michailidis (2015), Lai, Xu and Yuan (2020) use the orthogonal greedy algorithm (OGA) in conjunction with a high-dimensional information criterion (HDIC) to choose regressors in a stochastic regression model $Y_t=\boldsymbol{\beta}^T\mathbf{x}_t+\epsilon_t, 1\le t\le n$, with $\mathcal{F}_{t-1}$-measurable $\mathbf{x}_t\in\mathbb{R}^{p_n}$ and martingale difference sequence $(\epsilon_t,\mathcal{F}_t)_{1\le t\le p_n}$. Assuming that the $\mathbf{x}_t$ and Y_t are centered so that $\bar{\mathbf{x}}=\mathbf{0}$ and $\bar{Y}=0$ and letting $\mathbf{X}_j=(x_{1j},\ldots,x_{nj})^T$ and $\mathbf{X}_J=(\mathbf{X}_j, j\in J)$, OGA is a fast iterative procedure that chooses the set $\hat{J}_k$ of indices of the input variable after k iterations and applies least squares to the residuals $U_t^k=Y_t-\hat{Y}_t^k$ as described below. First initialize with $\mathbf{U}^0=(Y_1,\ldots,Y_n)^T$ and $\hat{J}_0=\varnothing$. For $k=1$ to m, (a) choose $j=\hat{j}_k\notin\hat{J}_{k-1}$ such that $\mathbf{X}_j$ is most correlated with $\mathbf{U}^{k-1}$, (b) update $\hat{J}_k=\hat{J}_{k-1}\cup\{\hat{j}_k\}$, compute the projection $\widehat{\mathbf{X}}_{\hat{j}_k}$ of $\mathbf{X}_{\hat{j}_k}$ into the linear space spanned by $\mathbf{X}_{\hat{j}_1},\mathbf{X}_{\hat{j}_2}^{\perp},\ldots,\mathbf{X}_{\hat{j}_{k-1}}^{\perp}$ and let $\mathbf{X}_{\hat{j}_k}^{\perp}=\mathbf{X}_{\hat{j}_k}-\widehat{\mathbf{X}}_{\hat{j}_k}$, (c) compute $\hat{\beta}_{\hat{j}_k}^k=(\sum_{t=1}^{n}U_t^{k-1}x_{t,\hat{j}_k}^{\perp})/\sum_{t=1}^{n}(x_{t,\hat{j}_k}^{\perp})^2$, and let $\hat{Y}_t^k=\hat{Y}_t^{k-1}+\hat{\beta}_{\hat{j}_k}^k x_{t,\hat{j}_k}^{\perp}$ for $1\le t\le n$, and $\mathbf{U}^k=(Y_1-\hat{Y}_1^k,\ldots,Y_n-\hat{Y}_n^k)^T$. Let $\mathbf{P}_J$ (resp., $\mathbf{P}_J^{\perp}$) denote the matrix associated with orthogonal projection into the space spanned by (resp., orthogonal to) $\mathbf{X}_j, j\in J$. Thus, $\mathbf{P}_J=\mathbf{X}_J(\mathbf{X}_J^T\mathbf{X}_J)^{-1}\mathbf{X}_J^T$, $\mathbf{P}_J^{\perp}=\mathbf{I}-\mathbf{P}_J$. Assuming the $\mathbf{x}_t$ to be i.i.d. and $\log p_n=o(n)$, Ing and Lai (2011) call the regression model *weakly sparse* if $\sup_{n\ge1}\sum_{j=1}^{p_n}|\beta_j\sigma_j|<\infty$, where $\sigma_j^2=\operatorname{Var}(x_{tj})$, and show that under certain finiteness assumptions on moment generating functions,

$$E\big[\{\hat{y}_{\hat{J}}(\mathbf{x})-y_{\hat{J}}(\mathbf{x})\}^2|y_1,\mathbf{x}_1,\ldots,y_n,\mathbf{x}_n\big]=O_p(n^{-1}m\log p_n)$$

if OGA terminates after $m = m_n = O(\sqrt{n/\log p_n})$ iterations, where $\hat{J} = \{\hat{j}_1, \cdots, \hat{j}_m\}$ is the set of variables selected by OGA, $\hat{y}_{\hat{J}}(\mathbf{x}) = \mathbf{x}^T\widehat{\boldsymbol{\beta}}$ in which $\widehat{\boldsymbol{\beta}}$ is the OGA estimate of $\boldsymbol{\beta}$, $(\mathbf{x}, y)$ is independent of $(\mathbf{x}_t, y_t)$ for $t = 1, \ldots, n$ but has the same distribution, and $y_{\hat{J}}(\mathbf{x}) = \sum_{j\in\hat{J}}\beta_j x_j$ is the "semipopulation version" of OGA. They also introduce a high-dimensional information criterion (HDIC) to choose along the OGA path the model that has the smallest HDIC. In this setting of i.i.d. $\mathbf{x}_t$ that are independent of i.i.d. ϵ_t, they assume $E(\exp(s\epsilon)) < \infty$ for $|s| \le s_0$ and $\max_{1\le j\le p_n} E(\exp(sx_j^2)) < \infty$ for $0 < s < s_1$, together with a variant of the restricted eigenvalue condition, and use exponential inequalities to bound $\sum_{t=1}^n \epsilon_t\mathbf{x}_t$ and $\|\widehat{\boldsymbol{\Gamma}}_n^{-1}(J) - \boldsymbol{\Gamma}^{-1}(J)\|$, where $\boldsymbol{\Gamma}(J)$ is the covariance matrix of $(x_j; j \in J)$ and $\widehat{\boldsymbol{\Gamma}}_n(J)$ is the corresponding sample covariance matrix. Extending directly these exponential inequalities to $\mathcal{F}_{t-1}$-measurable regressor $\mathbf{x}_t$ and martingale difference ϵ_t (with respect to $\mathcal{F}_t$) is much more difficult. However, de la Peña, Klass and Lai (2009) have developed exponential and moment inequalities for self-normalized locally square integrable martingales, for which self-normalization consists of multiplying by the inverse of its quadratic or predictable process (or some linear combination thereof) which is a matrix. By making use of these inequalities, Lai, Xu and Yuan (2020) have recently extended Ing and Lai's (2011) theory of OGA+HDIC to stochastic regression models with $\mathcal{F}_{t-1}$-measurable $\mathbf{x}_t \in \mathbb{R}^{p_n}$ and martingale difference sequence $\{\epsilon_t, 1 \le t \le p_n\}$.

3.2 Recursive Gradient Boosting for Nonlinear Stochastic Regression

In Section 2.3, we have reviewed recursive stochastic gradient algorithms beginning with Widrow and Hoff's LMS algorithm in 1960 as an alternative to the Kalman filter, followed by the recursive stochastic gradient algorithms for parameter estimation and associated adaptive control rules in ARMAX models by Goodwin, Ramadge and Caines (1981) and Fuchs (1982). We have noted that although the adaptive control rules have the self-tuning property, they do not have logarithmic regret and the recursive parameter estimates are not as efficient as the considerably more complicated matrix-gain recursions. Here, we review recent developments in stochastic gradient algorithms which can also achieve full asymptotic efficiency via a modification of Friedman's gradient boosting machine, introduced in 2001, called "modified gradient boosting" (MGB). In regression or classification problems with high-dimensional covariate vectors, one faces the problem of minimizing a loss function over a high-dimensional parameter space. When the loss function is convex, regularization methods like Lasso or elastic net are often used to find a sparse solution. However, in a more general framework, the loss function may not be convex in the parameters and difficulties arise for fitting the model. Boosting is a powerful tool to circumvent this difficulty. The essence of boosting is to combine many base learners in a greedy way to produce a powerful predictor. For classification problems, the AdaBoost algorithm introduced by Freund and Schapire (1997) takes votes from many weak classifiers to form a boosted classifier using the majority vote. Friedman, Hastie and Tibshirani's (2000) Real AdaBoost procedure is an extension to prediction problems and provides a statistical framework, via additive modeling and maximum likelihood, to understand why AdaBoost can produce dramatic improvements in performance over the weak classifiers (learners). The "gradient boosting machine" introduced by Friedman (2001) represents a further generalization that connects stagewise additive expansions to steepest-descent minimization, for which "function estimation/approximation is viewed from the perspective of numerical optimization in function space, rather than parameter space." This general paradigm can be "based on any fitting criterion" via the use of loss functions of the form $L(Y_t, f(\mathbf{x}_t))$, in which Y_t and $\mathbf{x}_t$ are the observed response and covariate vector for $t = 1, \ldots, n$ and f is the regression function that has an additive expansion of the form $f(\mathbf{x}) = \alpha + \sum_{k=1}^p \beta_k\phi_k(\mathbf{x}; \mathbf{b}_k)$, in which ϕ_k is a basis function that involves a nonlinear parameter vector $\mathbf{b}_k \in B$ and is linearly associated with a regression coefficient β_k. Friedman assumes $\alpha \equiv 0$ and $\phi_k \equiv \phi$ and his Gradient_Boost algorithm is initialized at $\hat{f}^0(\mathbf{x}) = 0$. It carries out the following steps at the kth iteration:

(a) $\hat{u}_t^{k-1} = -\frac{\partial L}{\partial f}(Y_t, \hat{f}^{k-1}(\mathbf{x}_t)), t = 1, \ldots, n,$

(b) $\hat{\mathbf{b}}_k = \arg\min_{\mathbf{b}\in\Gamma,\beta\in\mathbb{R}} \sum_{t=1}^n [\hat{u}_t^{k-1} - \beta\phi(\mathbf{x}_t; \mathbf{b})]^2,$

(c) $\hat{\beta}_k = \arg\min_\beta \sum_{t=1}^n L(Y_t, \hat{f}^{k-1}(\mathbf{x}_t) + \beta\phi(\mathbf{x}_t; \hat{\mathbf{b}}_k)),$

(d) $\hat{f}^k(\mathbf{x}) = \hat{f}^{k-1}(\mathbf{x}) + \hat{\beta}_k\phi(\mathbf{x}; \hat{\mathbf{b}}_k).$

It terminates after m iterations and outputs $\hat{f}^m(\cdot)$.

For the case $L(y, f) = (y - f)^2/2$, Gradient_Boost reduces to the "pure greedy algorithm (PGA)" of Temlyakov (2000), which is called "matching pursuit" by Mallat and Zhang (1993) for the special case of "time-frequency" dictionaries and also called "L_2-boosting" by Bühlmann (2006), who shows that for the linear regression model with $p_n = \exp(O(n^\xi))$ for some $0 < \xi < 1$, $E\{(f(\mathbf{x}) - \hat{f}^m(\mathbf{x}))^2 | y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n\}$ converges in probability to 0 if $m = m_n \to \infty$ sufficiently slowly, where $\mathbf{x}$ is independent of $(\mathbf{x}_t, y_t)$ and has the same distribution as $\mathbf{x}_t$ which are assumed to be i.i.d., but no results on how slowly m_n should grow have been derived. It is widely recognized that early termination can avoid overfitting, and some variable selection schemes such as AIC have been proposed to choose m_n, but a convergence theory of PGA (or L_2-boosting) is lacking. On the other hand, there is a definitive convergence theory of OGA that is summarized

in the last paragraph of Section 3.3. A major difference between OGA and PGA is that at each iteration OGA selects a new input variable whereas PGA can select the same input variable in multiple iterations. Thus termination of OGA after m_n iterations implies inclusion of m_n regressors in the linear regression model whereas the number of regressors included in PGA after m_n iterations is unclear other than that it cannot be larger than m_n, contributing to the difficulties in analyzing PGA. For OGA, Ing and Lai (2011) have shown that optimal bias-variance tradeoff in high-dimensional sparse linear models entails that m_n should be $O((n/\log p_n)^{1/2})$, suggesting termination of the OGA iterations with $K_n = O((n/\log p_n)^{1/2})$ input variables, assuming that $\log p_n = o(n)$. Letting $y_J(\mathbf{x}) = \sum_{j\in J}\beta_j x_j$, an important tool used by Ing and Lai (2011) is an upper bound, due to Temlyakov (2000), on the conditional mean squared prediction error $E\{(y_{\hat{J}}(\mathbf{x}) - \boldsymbol{\beta}^T\mathbf{x})^2|y_1, \mathbf{x}_1, \ldots, y_n, \mathbf{x}_n\}$ for weak orthogonal greedy algorithms that can be applied to analyze the semipopulation version of OGA. By making use of moderate deviation bounds for the least squares estimates of the unknown regression coefficients in the sample version of the procedure, Ing and Lai (2011) derive the desired convergence rate from that of the semipopulation version. For PGA, there is a corresponding Temlyakov bound for the semipopulation version, which has been used in the aforementioned work of Bülhmann. However, because the same input variable can be used repeatedly in the PGA iterations, there are inherent difficulties in determining the number of iterations and deriving the convergence rate of PGA from the Temlyakov bound for the semipopulation version of PGA. These difficulties become even more intractable for gradient boosting in regression models with nonlinear parameters in the basis functions.

Lai, Xia and Yuan (2020) have recently introduced a *modified gradient boosting* (MGB) algorithm to circumvent these difficulties in the additive expansion and for general loss functions $L(Y_t, f(\mathbf{x}_t))$. Whereas Ing and Lai's OGA uses forward greedy inclusion of regressors until K_n variables have been included, MGB modifies it into a "less greedy" procedure that chooses $(\hat{j}_k, \hat{\mathbf{b}}_k)$ with the smallest number $\#_j^{k-1}$ of iterations up to stage $k-1$, over $1 \le j \le p_n$ and $\mathbf{b} \in B$, that attains at least ϵ times the maximum squared correlation of $(\hat{u}_t^{k-1} - \hat{\alpha}_{k-1})_{1\le t\le n}$ and $\phi_j(\mathbf{x}_t; \mathbf{b})_{1\le t\le n}$, with prescribed $0 < \epsilon < 1$. Following Ing and Lai (2011), MGB stops including new basis functions at stage m_n when K_n distinct $\hat{j}_k$'s are included in the basis expansion. For $k > m_n$, MGB continues the preceding procedure with j restricted to the K_n distinct $\hat{j}_k$'s until stage $\tilde{m}_n$ when loss minimization converges within prescribed tolerance limits. Under certain regularity conditions, Lai, Xia and Yuan (2020) have developed an asymptotic theory of MGB as $n \to \infty$, parallel to that of OGA in linear regression with squared error loss, for stochastic minimization of general loss functions. Gradient descent used in MGB obviates the need for the restricted eigenvalue condition for OGA, and the weak greedy selection of basis functions has much lower computational cost than OGA+HDIC. This shows that a suitably chosen scalar-gain sequence coupled with weak greedy selection of input variables (or basis functions) enables MGB to attain the same asymptotic properties, as $n \to \infty$, as OGA in linear regression models with squared error loss, or more generally, in high-dimensional regression models with nonlinear basis functions and general loss functions.

However, MGB is an off-line algorithm that depends on a given training sample of size n, hence we have used the notation p_n and m_n to denote the number of basis functions to be considered and selected, respectively, for a given sample size n. Lai, Xu and Yuan (2020) have recently developed a recursive MGB algorithm by parallelizing the basis function selection and parameter updating tasks of MGB (represented by steps (b) and (c), respectively, for Gradient_Boost and their modifications described above). Recursive MGB carries out basis function selection only for sample sizes $n_1 < n_2 < \cdots$ so that when the sample size n reaches n_i, $m_{n_{i-1}}$ basis functions are selected on the basis of the first n_{i-1} observations. With the set of basis functions unchanged for $n_i \le n < n_{i+1}$, parameter updating can be carried out by a scalar-gain stochastic approximation algorithm; note that the minimization in step (c) of Gradient_Boost is over $\beta \in \mathbb{R}$.

3.3 Stochastic Approximation in Particle Swarm Optimization and AI

Artificial (or machine) intelligence (AI) is intelligence demonstrated by machines, in contrast with natural intelligence displayed by humans and animals. Many AI models and algorithms emulate those in natural intelligence; neural networks, deep learning and computer vision are well-known examples. Another example that has received much recent attention is particle swarm optimization (PSO) because of its wide range of applications in power systems, mechanical design, polymerization, biological sequence analysis, pharmacodynamics, robotics and industrial engineering; see Yuan and Yin (2015) who have developed a novel stochastic approximation (SA) scheme in connection with the convergence of PSO optimization algorithms. These algorithms involve optimization in network or multiagent systems and in autonomous systems, originating from the example of a swarm of birds searching for food—how each bird adjusts the next search direction in accordance with its current estimate of the best position (of food) and the communicated best position by its neighbors. Kennedy and Eberhart (1995) proposed a seminal PSO algorithm that underwent many subsequent refinements and developments; see Clerc (2006)

and Bonyadi and Michalewicz (2016), in which "birds" are replaced by "particles" for a particle swarm. Let M be the size of the swarm and $f:\mathbb{R}^d\to\mathbb{R}$ be the cost function to be minimized. Letting $\mathbf{x}_i(t)\in\mathbb{R}^d$ be the current position and $\mathbf{v}_i(t)$ be the current velocity of particle i, the dynamics of PSO can be expressed in terms of the recursive algorithms

$$\mathbf{x}_i(t+1)=\mathbf{x}_i(t)+\mathbf{v}_i(t+1), \tag{3.1}$$

$$\begin{aligned}\mathbf{v}_i(t+1)&=(1-\omega)\mathbf{v}_i(t)\\&\quad+c_1\mathbf{U}_{1,i}(t)\otimes\big(\mathbf{x}_i^*(t)-\mathbf{x}_i(t)\big)\\&\quad+c_2\mathbf{U}_{2,i}(t)\otimes\big(\mathbf{x}^*(t)-\mathbf{x}_i(t)\big).\end{aligned} \tag{3.2}$$

In (3.2), $\otimes$ denotes the Kronecker product, and $\mathbf{U}_{1,i}(t)$ and $\mathbf{U}_{2,i}(t)$ are independent random vectors in $\mathbb{R}^d$; the components of $\mathbf{U}_{1,i}(t)$ are i.i.d. with finite second moments and so are those of $\mathbf{U}_{2,i}(t)$. Moreover,

$$\begin{aligned}\mathbf{x}_i^*(t)&=\operatorname*{argmin}_{0\le s\le t} f\big(\mathbf{x}_i(s)\big),\\ \mathbf{x}^*(t)&=\operatorname*{argmin}_{1\le j\le M,0\le s\le t} f\big(\mathbf{x}_j(s)\big)\end{aligned} \tag{3.3}$$

represent the best position found by the ith particle and the "global best" found by the whole swarm up to the tth iteration.

After a review of previous works on convergence analysis of the PSO algorithm, Yuan and Yin (2015) point out that most of them rely on overly restrictive assumptions such as the swarm having "only one particle" and that "there are no rigorous rates of convergence results available to date." To address these issues, Choi et al. (2021) introduce the following enhancements of PSO, which not only can be guaranteed to converge to the global optimum under mild regularity conditions but are also related to stochastic approximation theory that provides the convergence rate results. Smoothed PSO (sPSO) uses martingale differences $\mathbf{Z}_i(t+1)$ and a tuning parameter $\eta\in(0,1)$ to convert the objective function into a regression function and modify the PSO iterations (3.1) and (3.2) as

$$\mathbf{x}_i(t+1)=\mathbf{x}_i(t)+\eta\mathbf{v}_i(t+1), \tag{3.4}$$

$$\begin{aligned}\mathbf{v}_i(t+1)&=(1-\eta\omega)\mathbf{v}_i(t)\\&\quad+\eta c_1\mathbf{U}_{1,i}(t)\otimes\big(\mathbf{x}_i^*(t)-\mathbf{x}_i(t)\big)\\&\quad+\eta c_2\mathbf{U}_{2,i}(t)\otimes\big(\mathbf{x}^*(t)-\mathbf{x}_i(t)\big)\\&\quad+\eta\mathbf{Z}_i(t+1).\end{aligned} \tag{3.5}$$

Following Yuan and Yin (2015), p. 1761, we assume $d=1$ for the convergence proof of sPSO which only entails convergence for each coordinate in the d-dimensional case. Let $\mathbf{x}_t=(x_1(t),\dots,x_M(t))^T$, $\mathbf{v}_t=(v_1(t),\dots,v_M(t))^T$, $\mathbf{x}_t^*=(x_1^*(t),\dots,x_M^*(t))^T$, $\mathbf{Z}_t=(Z_1(t),\dots,Z_M(t))^T$, $\mathbf{e}=(1,\dots,1)^T$, $\boldsymbol{\theta}_t=(\mathbf{x}_t^T,\mathbf{v}_t^T)^T$, and define the $M\times M$ diagonal matrices $\mathbf{D}_{1,t}=\operatorname{diag}(U_{1,1}(t),\dots,U_{1,M}(t))$, $\mathbf{D}_{2,t}=\operatorname{diag}(U_{2,1}(t),\dots,U_{2,M}(t))$. Then we can combine (3.4) and (3.5) for the case $d=1$ into

$$\begin{aligned}\boldsymbol{\theta}_{t+1}&=\begin{pmatrix}\mathbf{I}&\eta(1-\eta\omega)\mathbf{I}\\0&(1-\eta\omega)\mathbf{I}\end{pmatrix}\boldsymbol{\theta}_t\\&\quad+\begin{pmatrix}\eta^2c_1\mathbf{D}_{1,t}&\eta^2c_2\mathbf{D}_{2,t}\\\eta c_1\mathbf{D}_{1,t}&\eta c_2\mathbf{D}_{2,t}\end{pmatrix}\\&\quad\times\begin{pmatrix}\mathbf{x}_t^*-\mathbf{x}_t\\x^*(t)\mathbf{e}-\mathbf{x}_t\end{pmatrix}+\begin{pmatrix}\eta^2\mathbf{Z}_{t+1}\\\eta\mathbf{Z}_{t+1}\end{pmatrix}.\end{aligned} \tag{3.6}$$

In the PSO literature, c_1 and c_2 are called "acceleration constants" and $U_{1,i}(t)$, $U_{2,i}(t)$ are often chosen to be uniform random variables, while $0<\eta<1$ is called the "step-size." Unlike (3.6), Yuan and Yin (2015) do not explicitly describe how $\mathbf{Z}_{t+1}$ enter into their PSO algorithm. Instead, they point out that in practice the form of f in (3.3) is "not known precisely, or too complicated to compute," for which stochastic approximation methods are "well suited" to convergence proofs and derivation of convergence rates. However, since PSO is usually associated with known f in the literature, the sPSO enhancement (3.6) includes additional noise $(\eta^2,\eta)^T\mathbf{Z}_{t+1}$ in the optimization algorithm. The ODE approach used by Yuan and Yin (2015), pp. 1762–1766, can be used to establish the convergence of sPSO (as $\eta\to0$) to the global minimum of f under certain regularity assumptions. This approach embeds the discrete-time iterates $\boldsymbol{\theta}_t$ into a continuous-time process $\boldsymbol{\theta}^\eta(s)=\boldsymbol{\theta}_t$ for $\eta t\le s<(\eta+1)t$ and then shows that with probability $1+o(1)$ as $\eta\to0$, the process $\boldsymbol{\theta}^\eta$, consisting of subvectors $\mathbf{x}^\eta$ and $\mathbf{v}^\eta$, converges weakly to the solution $\boldsymbol{\theta}(\cdot)$ of $(d/dt)\boldsymbol{\theta}(t)=\mathbf{C}\boldsymbol{\theta}(t)+\mathbf{b}(\boldsymbol{\theta}_1(t))$, in which $\boldsymbol{\theta}_1(t)$ denotes the first M components of $\boldsymbol{\theta}(t)$ and $\mathbf{C}$ is a $2M\times2M$ matrix whose first M rows form the submatrix $(\mathbf{0},\mathbf{I})$ and the remaining submatrix is $(\mathbf{0},\omega\mathbf{I})$, under the regularity conditions of Yuan and Yin (2015), pp. 1762–1763. Moreover, its equilibrium (as $t\to\infty$) is given by $\mathbf{x}^*=\operatorname{argmin}_{\mathbf{x}}f(\mathbf{x})$ and $\mathbf{v}^*=\mathbf{0}$ under these assumptions, showing the convergence of sPSO to the global minimum of f. Under additional assumptions, Yuan and Yin (2015), pp. 1766–1768, have shown that $(\mathbf{x}^\eta(t)-\mathbf{x}^*)/\sqrt{\eta}$ converges weakly, as $\eta\to0$, to the solution $\mathbf{z}$ of the stochastic differential equation

$$d\mathbf{z}=(\mathbf{A}+\mathbf{B})\mathbf{z}\,dt+\boldsymbol{\Sigma}^{1/2}\,d\mathbf{w}, \tag{3.7}$$

where $\mathbf{B}$ and $\boldsymbol{\Sigma}$ are related to the tuning parameters ω,c_1,c_2 and the covariance matrices of $\mathbf{U}_{1,i}(t)$, $\mathbf{U}_{2,i}(t)$ and $\mathbf{Z}_i(t+1)$ in (3.5).

Choi et al. (2021) also introduce adaptive PSO (aPSO) that uses an adaptive choice of the parameters of sPSO. It is widely recognized that the finite-sample performance of PSO depends heavily on the choice of tuning parameters, hence it is desirable to estimate the optimal tuning parameters for sPSO sequentially so that this adaptive choice of

tuning parameters can mimic the oracle procedure that has *a priori* knowledge of the optimal parameters. They note that important insights into this problem were provided by adaptive stochastic approximation that we have reviewed in Section 2.2. They also point out that PSO and its enhancements also use two parallel algorithms as Spall's adaptive SPSA procedure in the last paragraph of Section 2.2, one for locating the optimum and the other for determining the velocity, for each of M particles. An important difference between sPSO and SPSA is that SPSA uses gain sequences a_t and c_t similar to Kiefer–Wolfowitz scheme, whereas sPSO uses a step-size η. Convergence analysis and convergence rate results involve the selection of the gain sequences a_t and c_t for SPSA and the step-size η for sPSO. Although the asymptotic theory requires the choices to converge to 0, it does not provide "practical guidelines" for such choices, as noted by Spall (2000) who also gave some guidelines which he subsequently modified to develop adaptive SPSA that uses the data collected so far to estimate the tuning parameters sequentially; see Spall (2003), Sections 7.5, 7.7, 7.8. In particular, he suggests choosing $a_t = a/(A+t)^\alpha$ and $c_t = c/t^\gamma$ with α and γ smaller than their asymptotically optimal values $\alpha = 1$ and $\gamma = 1/6$ (specifically, $\alpha = 0.6$ and $\gamma = 0.1$). Since sPSO uses step-size η instead of gain sequences with prescribed functional forms, adaptation can be done more simply by adaptive selection of the tuning parameters in aPSO that mimics the "oracle PSO," which assumes knowledge of the distribution of the martingale difference sequence $\{Z_i(t), t \geq 0\}$ for determining the step-size η. Although the convergence theory in the preceding paragraph requires sufficiently small η, the oracle sPSO initializes at a larger η_1 that yields the smallest mean of $\tau = \inf\{s \leq \widetilde{T} : f(\mathbf{x}^*(0)) - f(\mathbf{x}^*(s)) \leq \delta(f(\mathbf{x}^*(0)) - f(\mathbf{x}^*(\widetilde{T})))\}$, in which δ represents the descent rate, $\widetilde{T}$ is the time horizon for using the larger step-size η_1 and the convention $\inf \varnothing = \widetilde{T}$ is used. After this fast descent of the cost function with an optimally chosen η_1, the oracle sPSO then chooses a sufficiently small step-size η to minimize $E\|\mathbf{x}^*(T) - \mathbf{x}^*\|^2/\eta$, where T is the maximum number of recursions in the practical implementation of sPSO. Recalling that $(\mathbf{x}^\eta(T) - \mathbf{x}^*)/\sqrt{\eta}$ is approximately $N(\mathbf{0}, \mathbf{\Sigma})$ for sufficiently small η and large T by the asymptotic theory of sPSO, this amounts to finding the step-size that minimizes $\mathbf{\Sigma}$. Since particle swarm optimization involves a swarm of birds (particles), it cannot use multistart local search strategies for the optimal tuning parameters ω and $c(= c_1 = c_2)$. Instead, Choi et al. (2021) use a group sequential ϵ-greedy randomization procedure from reinforcement learning to choose (ω, c) during the flight path of the swarm, and show by theoretical analysis and simulation studies that aPSO has certain oracle properties.

3.4 Discussion—from Theory to Practice and Back

"Theory versus Practice" is a topic that has attracted much recent discussion in statistics in response to the challenges and opportunities in the big data era, and was the title of a panel discussion in the 2018 Joint Statistical Meetings to address the following questions: "What is the current balance between theory and applications in our field, in terms of how research is received/perceived/valued/rewarded? Has this balance changed in the last 20 years? 10 years? 5 years? Have we struck an optimal balance or are we moving in the wrong direction? What is the value of Theoretical work/Applied work? What are some of the main challenges in today's Theory world/Applied world? What is the role of computation and how does it fit in to all of this?" We have demonstrated in the case of stochastic approximation that theory has guided practice, which has in turn led to more powerful and versatile theories, giving persistent vibrancy to this multidisciplinary subject in which computation has also played an increasingly important role.

In their article on "neuroscience inspired AI," Hassabis et al. (2017) discuss the usefulness of neuroscience, which studies the "inner workings on the human brain" and "the behaviors that it generates, and the mechanisms by which it does so," to "accelerate and guide" AI research. They consider in particular deep learning and reinforcement learning. Section 3.2 describes modified gradient boosting that is closely related to deep learning, and its on-line version (recursive MGB) which can be used in sequential or time series settings and which is related to stochastic approximation via stochastic gradient algorithms pioneered by Widrow and Hoff's LMS algorithm. In their development of efficient aPSO, Choi et al. (2021) use adaptive stochastic approximation and reinforcement learning, as noted in Section 3.3 that discusses the role of stochastic approximation in recent developments in AI. An important area of reinforcement learning that also originated in statistical theory and has blossomed into big-data, multidisciplinary applications is multiarmed bandits and bandits with side information (or contextual bandits); see Robbins (1952), Lai and Robbins (1985), Lai (1987), Tewari and Murphy (2017), Luckett et al. (2020), Tomkins et al. (2019), and Lai, Choi and Tsang (2019).

ACKNOWLEDGMENTS

Lai's research is supported by the National Science Foundation grant DMS 1811818. Yuan's research is sponsored by Shanghai Pujiang Program (No. 18PJC047). The corresponding author is Yuan.

REFERENCES

ANDERSON, T. W. and TAYLOR, J. B. (1976). Some experimental results on the statistical properties of least squares estimates in control problems. *Econometrica* **44** 1289–1302.

ÅSTÖM, K. J. and WITTENMARK, B. (1973). On self-tuning regulators. *Automatica* **9** 189–199.

ÅSTRÖM, K. J. and WITTENMARK, B. (1971). Problems of identification and control. *J. Math. Anal. Appl.* **34** 90–113. MR0278799 https://doi.org/10.1016/0022-247X(71)90161-2

BASU, S. and MICHAILIDIS, G. (2015). Regularized estimation in sparse high-dimensional time series models. *Ann. Statist.* **43** 1535–1567. MR3357870 https://doi.org/10.1214/15-AOS1315

BLUM, J. R. (1954). Approximation methods which converge with probability one. *Ann. Math. Stat.* **25** 382–386. MR0062399 https://doi.org/10.1214/aoms/1177728794

BONYADI, M. and MICHALEWICZ, Z. (2016). Analysis of stability, local convergence, and transformation sensitivity of a variant of particle swarm optimization algorithm. *IEEE Trans. Evol. Comput.* **20** 370–385.

BOX, G. E. P. and WILSON, K. B. (1951). On the experimental attainment of optimum conditions. *J. Roy. Statist. Soc. Ser. B* **13** 1–38; discussion: 38–45. MR0046009

BÜHLMANN, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559–583. MR2281878 https://doi.org/10.1214/009053606000000092

CHOI, K. P., LAI, T. L., TONG, X. T. and WONG, W. K. (2021). A statistical approach to adaptive parameter tuning in nature-inspired optimization and optimal sequential design of dose-finding trials. *Statist. Sinica* **31**. To appear.

CLERC, M. (2006). *Particle Swarm Optimization*. ISTE, London. MR2269598 https://doi.org/10.1002/9780470612163

DE LA PEÑA, V. H., KLASS, M. J. and LAI, T. L. (2009). Theory and applications of multivariate self-normalized processes. *Stochastic Process. Appl.* **119** 4210–4227. MR2565565 https://doi.org/10.1016/j.spa.2009.10.003

EFRON, B. (2003). Robbins, empirical Bayes and microarrays. Dedicated to the memory of Herbert E. Robbins. *Ann. Statist.* **31** 366–378. MR1983533 https://doi.org/10.1214/aos/1051027871

FREUND, Y. and SCHAPIRE, R. (1997). A decision-theoretic generalization of online learning and an application to boosting. *J. Comput. System Sci.* **55** 119–139. MR1473055

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232. MR1873328 https://doi.org/10.1214/aos/1013203451

FRIEDMAN, J., HASTIE, T. and TIBSHIRANI, R. (2000). Additive logistic regression: A statistical view of boosting. *Ann. Statist.* **28** 337–407. With discussion and a rejoinder by the authors. MR1790002 https://doi.org/10.1214/aos/1016218223

FUCHS, J.-J. J. (1982). Indirect stochastic adaptive control: The general delay-white noise case. *IEEE Trans. Automat. Control* **27** 219–223. MR0673095 https://doi.org/10.1109/TAC.1982.1102827

GOODWIN, G. C., RAMADGE, P. J. and CAINES, P. E. (1981). Discrete time stochastic adaptive control. *SIAM J. Control Optim.* **19** 829–853. MR0634955 https://doi.org/10.1137/0319052

GUO, L., HUANG, D. W. and HANNAN, E. J. (1990). On ARX(∞) approximation. *J. Multivariate Anal.* **32** 17–47. MR1035605 https://doi.org/10.1016/0047-259X(90)90069-T

HANNAN, E. J. (1987). Rational transfer function approximation. *Statist. Sci.* **2** 135–161. With comments and a reply by the author. MR0904031

HASSABIS, D., KUMARAN, D., SUMMERFIELD, C. and BOTVINICK, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron* **95** 245–258. https://doi.org/10.1016/j.neuron.2017.06.011

HOTELLING, H. (1941). Experimental determination of the maximum of a function. *Ann. Math. Stat.* **12** 20–45. MR0003521 https://doi.org/10.1214/aoms/1177731784

ING, C.-K. and LAI, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica* **21** 1473–1513. MR2895106 https://doi.org/10.5705/ss.2010.081

KENNEDY, J. and EBERHART, R. (1995). Particle swarm optimization. In *Proc. IEEE Int. Conf. Neural Networks* 4 1942–1048, Perth, Australia.

KIEFER, J. (1953). Sequential minimax search for a maximum. *Proc. Amer. Math. Soc.* **4** 502–506. MR0055639 https://doi.org/10.2307/2032161

KIEFER, J. (1957). Optimum sequential search and approximation methods under minimum regularity assumptions. *J. Soc. Indust. Appl. Math.* **5** 105–136. MR0092326

KIEFER, J. and WOLFOWITZ, J. (1952). Stochastic estimation of the maximum of a regression function. *Ann. Math. Stat.* **23** 462–466. MR0050243 https://doi.org/10.1214/aoms/1177729392

KUMAR, P. R. and VARAIYA, P. (1986). *Stochastic Systems*, 1st ed. ed. Prentice-Hall, Englewood Cliffs, NJ. MR3471643

KUSHNER, H. J. and CLARK, D. S. (1978). *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. *Applied Mathematical Sciences* **26**. Springer, New York-Berlin. MR0499560

LAI, T. L. (1987). Adaptive treatment allocation and the multi-armed bandit problem. *Ann. Statist.* **15** 1091–1114. MR0902248 https://doi.org/10.1214/aos/1176350495

LAI, T. L. (1989). Extended stochastic Lyapunov functions and recursive algorithms in linear stochastic systems. In *Stochastic Differential Systems* (*Bad Honnef*, 1988) (N. Christopeit, K. Helmes and N. Kohlmann, eds.). *Lect. Notes Control Inf. Sci.* **126** 206–220. Springer, Berlin. MR1236069 https://doi.org/10.1007/BFb0043786

LAI, T. L. (2003). Stochastic approximation. Dedicated to the memory of Herbert E. Robbins. *Ann. Statist.* **31** 391–406. MR1983535 https://doi.org/10.1214/aos/1051027873

LAI, T. L., CHOI, A. and TSANG, K. W. (2019). Statistical science in information technology and precision medicine. *Ann. Math. Sci. Appl.* **4** 413–438. MR4020370 https://doi.org/10.4310/AMSA.2019.v4.n2.a6

LAI, T. L. and ROBBINS, H. (1979). Adaptive design and stochastic approximation. *Ann. Statist.* **7** 1196–1221. MR0550144

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. MR0621117 https://doi.org/10.1007/BF00536178

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) 103–120. Academic Press, New York. MR0705310

LAI, T. L. and ROBBINS, H. (1982b). Iterated least squares in multiperiod control. *Adv. in Appl. Math.* **3** 50–73. MR0646499 https://doi.org/10.1016/S0196-8858(82)80005-5

LAI, T. L. and ROBBINS, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22. MR0776826 https://doi.org/10.1016/0196-8858(85)90002-8

LAI, T. L. and SIEGMUND, D. (1986). The contributions of Herbert Robbins to mathematical statistics. *Statist. Sci.* **1** 276–284. MR0846004

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. MR0642726

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. MR0855543 https://doi.org/10.1109/TAC.1986.1104138

LAI, T. L., XIA, T. and YUAN, H. (2020). Modified gradient boosting in high-dimensional nonlinear regression. Technical report, Dept. Statistics, Stanford Univ.

LAI, T. L., XU, H. and YUAN, H. (2020). Self-normalized martingales, recursive orthogonal matching pursuit and modified gradient boosting in high-dimensional stochastic regression models. Technical report, Dept. Statistics, Stanford Univ.

LAI, T. L. and YING, Z. (1991a). Recursive identification and adaptive prediction in linear stochastic systems. *SIAM J. Control Optim.* **29** 1061–1090. MR1110087 https://doi.org/10.1137/0329058

LAI, T. L. and YING, Z. (1991b). Parallel recursive algorithms in asymptotically efficient adaptive control of linear stochastic systems. *SIAM J. Control Optim.* **29** 1091–1127. MR1110088 https://doi.org/10.1137/0329059

LJUNG, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. Automat. Control* **AC-22** 551–575. MR0465458 https://doi.org/10.1109/tac.1977.1101561

LUCKETT, D. J., LABER, E. B., KAHKOSKA, A. R., MAAHS, D. M., MAYER-DAVIS, E. and KOSOROK, M. R. (2020). Estimating dynamic treatment regimes in mobile health using V-learning. *J. Amer. Statist. Assoc.* **115** 692–706. MR4107673 https://doi.org/10.1080/01621459.2018.1537919

MALLAT, S. and ZHANG, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.* **41** 3397–3415.

NEMIROVSKI, A. and YUDIN, D. (1978). On Cezari's convergence of the steepest descent method for approximating saddle points of convex-concave functions. *Sov. Math., Dokl.* **19**. MR0482494

NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization. A Wiley-Interscience Publication*. Wiley, New York. MR0702836

PAGE, W. (1984). An interview with Herbert Robbins. *College Math. J.* **15** 2–24. MR0759951 https://doi.org/10.2307/3027425

PELIGRAD, M., SANG, H., ZHONG, Y. and WU, W. B. (2014). Exact moderate and large deviations for linear processes. *Statist. Sinica* **24** 957–969. MR3235407

POLYAK, B. T. (1990). A new method of stochastic approximation type. *Avtomat. i Telemekh.* **7** 98–107. MR1071220

RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G. and YU, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* **5** 935–980. MR2836766 https://doi.org/10.1214/11-EJS631

ROBBINS, H. E. (1944). Two properties of the function $\cos x$. *Bull. Amer. Math. Soc.* **50** 750–752. MR0011342 https://doi.org/10.1090/S0002-9904-1944-08232-3

ROBBINS, H. E. (1945). On the measure of a random set. II. *Ann. Math. Stat.* **16** 342–347. MR0019239 https://doi.org/10.1214/aoms/1177731060

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.* **58** 527–535. MR0050246 https://doi.org/10.1090/S0002-9904-1952-09620-8

ROBBINS, H. and MONRO, S. (1951). A stochastic approximation method. *Ann. Math. Stat.* **22** 400–407. MR0042668 https://doi.org/10.1214/aoms/1177729586

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, Ohio*, 1971) 233–257. MR0343355

RUPPERT, D. (1985). A Newton–Raphson version of the multivariate Robbins–Monro procedure. *Ann. Statist.* **13** 236–245. MR0773164 https://doi.org/10.1214/aos/1176346589

RUPPERT, D. (1991). Stochastic approximation. In *Handbook of Sequential Analysis* (B. K. Ghosh and P. K. Sen, eds.). *Statist. Textbooks Monogr.* **118** 503–529. Dekker, New York. MR1174318

SAKRISON, D. J. (1967). The use of stochastic approximation to solve the system identification problem. *IEEE Trans. Automat. Control* **12** 563–567.

SARIDIS, G. and STEIN, G. (1968). A new algorithm for linear system identification. *IEEE Trans. Automat. Control* **13** 592–594.

SIEGMUND, D. (2003). Herbert Robbins and sequential analysis. Dedicated to the memory of Herbert E. Robbins. *Ann. Statist.* **31** 349–365. MR1983532 https://doi.org/10.1214/aos/1051027870

SPALL, J. C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Automat. Control* **37** 332–341. MR1148715 https://doi.org/10.1109/9.119632

SPALL, J. C. (2000). Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. Automat. Control* **45** 1839–1853. MR1795352 https://doi.org/10.1109/TAC.2000.880982

SPALL, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley-Interscience Series in Discrete Mathematics and Optimization*. Wiley, Hoboken, NJ. MR1968388 https://doi.org/10.1002/0471722138

TEMLYAKOV, V. N. (2000). Weak greedy algorithms. *Adv. Comput. Math.* **12** 213–227. MR1745113 https://doi.org/10.1023/A:1018917218956

TEWARI, A. and MURPHY, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health* (J. Re, S. A. Murphy and S. Kumar, eds.) 495–517. Springer, Berlin.

TOMKINS, S., LIAO, P., YEUNG, S., KLASNJA, P. and MURPHY, S. A. (2019). Intelligent pooling in Thompson sampling for rapid personalization in mobile health. ICML 2019 Workshop RL4RealLife, 2019.

WEI, C. Z. (1987). Multivariate adaptive stochastic approximation. *Ann. Statist.* **15** 1115–1130. MR0902249 https://doi.org/10.1214/aos/1176350496

WIDROW, B. and HOFF, M. E. (1960). Adaptive switching circuits. In *Proc. IRE WESCON Convention Record, Part* 4 96–104.

WU, W.-B. and WU, Y. N. (2016). Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electron. J. Stat.* **10** 352–379. MR3466186 https://doi.org/10.1214/16-EJS1108

YUAN, Q. and YIN, G. (2015). Analyzing convergence and rates of convergence of particle swarm optimization algorithms using stochastic approximation methods. *IEEE Trans. Automat. Control* **60** 1760–1773. MR3365066 https://doi.org/10.1109/TAC.2014.2381454

ZHANG, C.-H. (2003). Compound decision theory and empirical Bayes methods. Dedicated to the memory of Herbert E. Robbins. *Ann. Statist.* **31** 379–390. MR1983534