1
Introduction

Consider an initially empty urn to which balls, either red or black, are added one at a time. Let y_n denote the *number* of red balls at time n and $x_n \stackrel{\text{def}}{=} y_n/n$ the *fraction* of red balls at time n. We shall suppose that the conditional probability that the next, i.e., the $(n+1)$st ball is red given the past up to time n is a function of x_n alone. Specifically, suppose that it is given by $p(x_n)$ for a prescribed $p : [0,1] \to [0,1]$. It is easy to describe $\{x_n, n \geq 1\}$ recursively as follows. For $\{y_n\}$, we have the simple recursion

$$y_{n+1} = y_n + \xi_{n+1},$$

where

$$\begin{aligned} \xi_{n+1} &= 1 \text{ if the } (n+1)\text{st ball is red,} \\ &= 0 \text{ if the } (n+1)\text{st ball is black.} \end{aligned}$$

Some simple algebra then leads to the following recursion for $\{x_n\}$:

$$x_{n+1} = x_n + \frac{1}{n+1}(\xi_{n+1} - x_n),$$

with $x_0 = 0$. This can be rewritten as

$$x_{n+1} = x_n + \frac{1}{n+1}(p(x_n) - x_n) + \frac{1}{n+1}(\xi_{n+1} - p(x_n)).$$

Note that $M_n \stackrel{\text{def}}{=} \xi_n - p(x_{n-1}), n \geq 1$ (with $p(x_0) \stackrel{\text{def}}{=}$ the probability of the first ball being red) is a sequence of zero mean random variables satisfying $E[M_{n+1}|\xi_m, m \leq n] = 0$ for $n \geq 0$. This means that $\{M_n\}$ is a *martingale difference sequence* (see Appendix C), i.e., uncorrelated with the 'past', and thus can be thought of as 'noise'. The above equation then can be thought of

as a noisy discretization (or *Euler scheme* in numerical analysis parlance) for the ordinary differential equation (o.d.e. for short)

$$\dot{x}(t) = p(x(t)) - x(t),$$

for $t \geq 0$, with nonuniform stepsizes $a(n) \stackrel{\text{def}}{=} 1/(n+1)$ and 'noise' $\{M_n\}$. (Compare with the standard Euler scheme $x_{n+1} = x_n + a(p(x_n) - x_n)$ for a small $a > 0$.) If we assume $p(\cdot)$ to be Lipschitz continuous, o.d.e. theory guarantees that this o.d.e. is well-posed, i.e., it has a unique solution for any initial condition $x(0)$ that in turn depends continuously on $x(0)$ (see Appendix B). Note also that the right-hand side of the o.d.e. is nonnegative at $x(t) = 0$ and nonpositive at $x(t) = 1$, implying that any trajectory starting in $[0, 1]$ will remain in $[0, 1]$ forever. As this is a scalar o.d.e., any bounded trajectory must converge. To see this, note that it cannot move in any particular direction ('right' or 'left') forever without converging, because it is bounded. At the same time, it cannot change direction from 'right' to 'left' or vice versa without passing through an equilibrium point: This would require that the right-hand side of the o.d.e. changes sign and hence by continuity must pass through a point where it vanishes, i.e., an equilibrium point. The trajectory must then converge to this equilibrium, a contradiction. (For that matter, the o.d.e. couldn't have been going both right and left at any given x because this direction is uniquely prescribed by the sign of $p(x) - x$.) Thus we have proved that $x(\cdot)$ must converge to an equilibrium. The set of equilibria of the o.d.e. is given by the points where the right-hand side vanishes, i.e., the set $H = \{x : p(x) = x\}$. This is precisely the set of fixed points of $p(\cdot)$. Once again, as the right-hand side is continuous, is ≤ 0 at 1, and is ≥ 0 at 0, it must pass through 0 by the mean value theorem and hence H is nonempty. (One could also invoke the Brouwer fixed point theorem (Appendix A) to say this, as $p : [0, 1] \to [0, 1]$ is a continuous map from a convex compact set to itself.)

Our interest, however, is in $\{x_n\}$. The theory we develop later in this book will tell us that the $\{x_n\}$ 'track' the o.d.e. with probability one in a certain sense to be made precise later, implying in particular that they converge a.s. to H. The key factors that ensure this are the fact that the stepsize $a(n)$ tends to zero as $n \to \infty$, and the fact that the series $\sum_n a(n)M_{n+1}$ converges a.s., a consequence of the martingale convergence theorem. The first observation means in particular that the 'pure' discretization error becomes asymptotically negligible. The second observation implies that the 'tail' of the above convergent series given by $\sum_{m=n}^{\infty} a(n)M_{n+1}$, which is the 'total noise added to the system from time n on', goes to zero a.s. This in turn ensures that the error due to noise is also asymptotically negligible. We note here that the fact $\sum_n a(n)^2 = \sum_n (1/(n+1)^2) < \infty$ plays a crucial role in facilitating the application of the martingale convergence theorem in the analysis of the urn

scheme above. This is because it ensures the following sufficient condition for martingale convergence (see Appendix C):

$$\sum_n E[(a(n)M_{n+1})^2|\xi_m, m \leq n] \leq \sum_n a(n)^2 < \infty, \text{ a.s.}$$

One also needs the fact that $\sum_n a(n) = \infty$, because in view of our interpretation of $a(n)$ as a time step, this ensures that the discretization does cover the entire time axis. As we are interested in tracking the asymptotic behaviour of the o.d.e., this is clearly necessary.

Let's consider now the simple case when H is a finite set. Then one can say more, viz., that the $\{x_n\}$ converge a.s. to some point in H. The exact point to which they converge will be random, though we shall later narrow down the choice somewhat (e.g., the 'unstable' equilibria will be avoided with probability one under suitable conditions). For the time being, we shall stop with this conclusion and discuss the *raison d'etre* for looking at such '*nonlinear urns*' .

This simple set-up was proposed by W. Brian Arthur (1994) to model the phenomenon of increasing returns in economics. The reader will have heard of the 'law of diminishing returns' from classical economics, which can be described as follows. Any production enterprise such as a farm or a factory requires both fixed and variable resources. When one increases the amount of variable resources, each additional unit thereof will get a correspondingly smaller fraction of fixed resources to draw upon, and therefore the additional returns due to it will correspondingly diminish.

While quite accurate in describing the traditional agricultural or manufacturing sectors, this law seems to be contradicted in some other sectors, particularly in case of the modern 'information goods'. One finds that larger investments in a brand actually fetch larger returns because of standardization and compatibility of goods, brand loyalty of customers, and so on. This is the so-called 'increasing returns' phenomenon modelled by the urn above, where each new red ball is an additional unit of investment in a particular product. If the predominance of one colour tends to fetch more balls of the same, then after some initial randomness the process will get 'locked into' one colour which will dominate overwhelmingly. (This corresponds to $p(x) > x$ for $x \in (x_0, 1)$ for some $x_0 \in (0, 1)$, and $< x$ for $x \in (0, x_0)$. Then the stable equilibria are 0 and 1, with x_0 being an unstable equilibrium. Recall that in this set-up the equilibrium x is stable if $p'(x) < 1$, unstable if $p'(x) > 1$.) When we are modelling a pair of competing technologies or conventions, this means that one of them, not necessarily the better one, will come to dominate overwhelmingly. Arthur (1994) gives several interesting examples of this phenomenon. To mention a few, he describes how the VHS technology came to dominate over Sony Betamax for video recording, why the present arrangement of letters and symbols

on typewriters and keyboards (QWERTY) could not be displaced by a superior arrangement called DVORAK, why 'clockwise' clocks eventually displaced 'counterclockwise' clocks, and so on.

Keeping economics aside, our interest here will be in the recursion for $\{x_n\}$ and its analysis sketched above using an o.d.e. The former constitutes a special (and a rather simple one at that) case of a much broader class of stochastic recursions called 'stochastic approximation' which form the main theme of this book. What's more, the analysis based on a limiting o.d.e. is an instance of the 'o.d.e. approach' to stochastic approximation which is our main focus here. Before spelling out further details of these, here's another example, this time from statistics.

Consider a repeated experiment which gives a string of input-output pairs (X_n, Y_n), $n \geq 1$, with $X_n \in \mathcal{R}^m, Y_n \in \mathcal{R}^k$ resp. We assume that $\{(X_n, Y_n)\}$ are i.i.d. Our objective will be to find the 'best fit' $Y_n = f_w(X_n) + \epsilon_n, n \geq 1$, from a given parametrized family of functions $\{f_w : \mathcal{R}^m \to \mathcal{R}^k : w \in \mathcal{R}^d\}$, ϵ_n being the 'error'. What constitutes the 'best fit', however, depends on the choice of our error criterion and we shall choose this to be the popular 'mean square error' given by $g(w) \stackrel{\text{def}}{=} \frac{1}{2}E[||\epsilon_n||^2] = \frac{1}{2}E[||Y_n - f_w(X_n)||^2]$. That is, we aim to find a w^* that minimizes this over all $w \in \mathcal{R}^d$. This is the standard problem of nonlinear regression. Typical parametrized families of functions are polynomials, splines, linear combinations of sines and cosines, or more recently, wavelets and neural networks. The catch here is that the above expectation cannot be evaluated because the underlying probability law is not known. Also, we do not suppose that the entire string $\{(X_n, Y_n)\}$ is available as in classical regression, but that it is being delivered one at a time in 'real time'. The aim then is to come up with a recursive scheme which tries to 'learn' w^* in real time by adaptively updating a running guess as new observations come in.

To arrive at such a scheme, let's pretend to begin with that we do know the underlying law. Assume also that f_w is continuously differentiable in w and let $\nabla^w f_w(\cdot)$ denote its gradient w.r.t. w. The obvious thing to try then is to differentiate the mean square error w.r.t. w and set the derivative equal to zero. Assuming that the interchange of expectation and differentiation is justified, we then have

$$\nabla^w g(w) = -E[\langle Y_n - f_w(X_n), \nabla^w f_w(X_n)\rangle] = 0$$

at the minimum point. We may then seek to minimize the mean square error by gradient descent, given by:

$$\begin{aligned} w_{n+1} &= w_n - \nabla^w g(w_n) \\ &= w_n + E[\langle Y_n - f_{w_n}(X_n), \nabla^w f_{w_n}(X_n)\rangle | w_n]. \end{aligned}$$

This, of course, is not feasible for reasons already mentioned, viz., that the

expectation above cannot be evaluated. As a first approximation, we may then consider replacing the expectation by the 'empirical gradient', i.e., the argument of the expectation evaluated at the current guess w_n for w^*,

$$w_{n+1} = w_n + \langle Y_n - f_{w_n}(X_n), \nabla^w f_{w_n}(X_n)\rangle.$$

This, however, will lead to a different kind of problem. The term added to w_n on the right is the nth in a sequence of 'i.i.d. functions' of w, evaluated at w_n. Thus we expect the above scheme to be (and it is) a correlated random walk, zigzagging its way to glory. We may therefore want to smooth it by making only a small, incremental move in the direction suggested by the right-hand side instead of making the full move. This can be achieved by replacing the right-hand side by a convex combination of it and the previous guess w_n, with only a small weight $1 > a(n) > 0$ for the former. That is, we replace the above by

$$w_{n+1} = (1 - a(n))w_n + a(n)(w_n + \langle Y_n - f_{w_n}(X_n), \nabla^w f_{w_n}(X_n)\rangle).$$

Equivalently,

$$w_{n+1} = w_n + a(n)\langle Y_n - f_{w_n}(X_n), \nabla^w f_{w_n}(X_n)\rangle.$$

Once again, if we do not want the scheme to zigzag drastically, we should make $\{a(n)\}$ small, the smaller the better. At the same time, a small $a(n)$ leads to a very small correction to w_n at each iterate, so the scheme will work very slowly, if at all. This suggests starting the iteration with relatively high $\{a(n)\}$ and letting $a(n) \to 0$. (In fact, $a(n) < 1$ as above is not needed, as that can be taken care of by scaling the empirical gradient.) Now let's add and subtract the exact error gradient at the 'known guess' w_n from the empirical gradient on the right-hand side and rewrite the above scheme as

$$\begin{aligned} w_{n+1} \quad = \quad & w_n + \ a(n)(E[\langle Y_n - f_{w_n}(X_n), \nabla^w f_{w_n}(X_n)\rangle | w_n]) \\ & + a(n)(\langle Y_n - f_{w_n}(X_n), \nabla^w f_{w_n}(X_n)\rangle \\ & - E[\langle Y_n - f_{w_n}(X_n), \nabla^w f_{w_n}(X_n)\rangle | w_n]). \end{aligned}$$

This is of the form

$$w_{n+1} = w_n + a(n)(-\nabla^w g(w_n) + M_{n+1}),$$

with $\{M_n\}$ a martingale difference sequence as in the previous example. One may then view this scheme as a noisy discretization of the o.d.e.

$$\dot{w}(t) = -\nabla^w g(w(t)).$$

This is a particularly well studied o.d.e. We know that it will converge to

$H \stackrel{\text{def}}{=} \{w : \nabla^w g(w) = 0\}$ in general, and if this set is discrete, to in fact one of the local minima of g for typical (i.e., *generic*: belonging to an open dense set) initial conditions. As before, we are interested in tracking the asymptotic behaviour of this o.d.e. Hence we must ensure that the discrete time steps $\{a(n)\}$ used in the 'noisy discretization' above do cover the entire time axis, i.e.,

$$\sum_n a(n) = \infty, \tag{1.0.1}$$

while retaining $a(n) \to 0$. (Recall from the previous example that $a(n) \to 0$ is needed for asymptotic negligibility of discretization errors.) At the same time, we also want the error due to noise to be asymptotically negligible a.s. The urn example above then suggests that we also impose

$$\sum_n a(n)^2 < \infty, \tag{1.0.2}$$

which asymptotically suppresses the noise variance.

One can show that with (1.0.1) and (1.0.2) in place, for reasonable g (e.g., with $\lim_{||w||\to\infty} g(w) = \infty$ and finite H, among other possibilities) the 'stochastic gradient scheme' above will converge a.s. to a local minimum of g.

Once again, what we have here is a special case – perhaps the most important one – of stochastic approximation, analyzed by invoking the 'o.d.e. method'.

What, after all, is stochastic approximation? Historically, stochastic approximation started as a scheme for solving a nonlinear equation $h(x) = 0$ given 'noisy measurements' of the function h. That is, we are given a black box which on input x, gives as its output $h(x) + \xi$, where ξ is a zero mean random variable representing noise. The stochastic approximation scheme proposed by Robbins and Monro (1951)† was to run the iteration

$$x_{n+1} = x_n + a(n)[h(x_n) + M_{n+1}], \tag{1.0.3}$$

where $\{M_n\}$ is the noise sequence and $\{a(n)\}$ are positive scalars satisfying (1.0.1) and (1.0.2) above. The expression in the square brackets on the right is the noisy measurement. That is, $h(x_n)$ and M_{n+1} are not separately available, only their sum is. We shall assume $\{M_n\}$ to be a martingale difference sequence, i.e., a sequence of integrable random variables satisfying

$$E[M_{n+1}|x_m, M_m, m \leq n] = 0.$$

This is more general than it appears. For example, an important special case is the d-dimensional iteration

$$x_{n+1} = x_n + a(n)f(x_n, \xi_{n+1}), \ n \geq 0, \tag{1.0.4}$$

† See Lai (2003) for an interesting historical perspective.

for an $f : \mathcal{R}^d \times \mathcal{R}^k \to \mathcal{R}^d$ with i.i.d. noise $\{\xi_n\}$. This can be put in the format of (1.0.3) by defining $h(x) = E[f(x, \xi_1)]$ and $M_{n+1} = f(x_n, \xi_{n+1}) - h(x_n)$ for $n \geq 0$.

Since its inception, the scheme (1.0.3) has been a cornerstone in scientific computation. This has been so largely because of the following advantages, already apparent in the above examples:

- It is designed to handle noisy situations, e.g., the stochastic gradient scheme above. One may say that it captures the average behaviour in the long run. The noise in practice may not only be from measurement errors or approximations, but may also be added deliberately as a probing device or a randomized action, as, e.g., in certain dynamic game situations.
- It is incremental, i.e., it makes small moves at each step. This typically leads to more graceful behaviour of the algorithm at the expense of its speed. We shall say more on this later in the book.
- In typical applications, the computation per iterate is low, making its implementation easy.

These features make the scheme ideal for applications where the key word is 'adaptive'. Thus the stochastic approximation paradigm dominates the fields of adaptive signal processing, adaptive control, and certain subdisciplines of soft computing / artificial intelligence such as neural networks and reinforcement learning – see, e.g., Bertsekas and Tsitsiklis (1997), Haykin (1991) and Haykin (1998). Not surprisingly, it is also emerging as a popular framework for modelling boundedly rational macroeconomic agents – see, e.g., Sargent (1993). The two examples above are representative of these two strands. We shall be seeing many more instances later in this book.

As noted in the preface, there are broadly two approaches to the theoretical analysis of such algorithms. The first, popular with statisticians, is the probabilistic approach based on the theory of martingales and associated objects such as 'almost supermartingales'. The second approach, while still using a considerable amount of martingale theory, views the iteration as a noisy discretization of a limiting o.d.e. Recall that the standard 'Euler scheme' for numerically approximating a trajectory of the o.d.e.

$$\dot{x}(t) = h(x(t))$$

would be

$$x_{n+1} = x_n + ah(x_n),$$

with $x_0 = x(0)$ and $a > 0$ a small time step. The stochastic approximation iteration differs from this in two aspects: replacement of the constant time step 'a' by a time-varying '$a(n)$', and the presence of 'noise' M_{n+1}. This qualifies it as a noisy discretization of the o.d.e. Our aim is to seek x for which $h(x) = 0$,

i.e., the equilibrium point(s) of this o.d.e. The o.d.e. would converge (if it does) to these only asymptotically unless it happens to start exactly there. Hence to capture this asymptotic behaviour, we need to track the o.d.e. over the infinite time interval. This calls for the condition $\sum_n a(n) = \infty$. The condition $\sum_n a(n)^2 < \infty$ will on the other hand ensure that the errors due to discretization of the o.d.e. and those due to the noise $\{M_n\}$ both become negligible asymptotically with probability one. (To motivate this, let $\{M_n\}$ be i.i.d. zero mean with a finite variance σ^2. Then by a theorem of Kolmogorov, $\sum_n a(n)M_n$ converges a.s if and only if $\sum_n a(n)^2$ converges.) Together these conditions try to ensure that the iterates do indeed capture the asymptotic behaviour of the o.d.e. We have already seen instances of this above.

Pioneered by Derevitskii and Fradkov (1974), this 'o.d.e. approach' was further extended and introduced to the engineering community by Ljung (1977). It is already the basis of several excellent texts such as Benveniste, Metivier and Priouret (1990), Duflo (1996), and Kushner and Yin (2003), among others†. The rendition here is a slight variation of the traditional one, with an eye on pedagogy so that the highlights of the approach can be introduced quickly and relatively simply. The lecture notes of Benaim (1999) are perhaps the closest in spirit to the treatment here, though at a much more advanced level. (Benaim's notes in particular give an overview of the contributions of Benaim and Hirsch, which introduced important notions from dynamical systems theory, such as internal chain recurrence, to stochastic approximation. These represent a major development in this field in recent years.)

While it is ultimately a matter of personal taste, the o.d.e. approach does indeed appeal to engineers because of the 'dynamical systems' view it takes, which is close to their hearts. Also, as we shall see at the end of this book, it can serve as a useful recipe for concocting new algorithms: any convergent o.d.e. is a potential source of a stochastic approximation algorithm that converges with probability one.

The organization of the book is as follows. Chapter 2 gives the basic convergence analysis for the stochastic approximation algorithm with decreasing stepsizes. This is the core material for the rest of the book. Chapter 3 gives some 'stability tests' that ensure the boundedness of iterates with probability one. Chapter 4 gives some refinements of the results of Chapter 2, viz., an estimate for probability of convergence to a specific attractor if the iterates fall in its domain of attraction. It also gives a result about avoidance with probability one of unstable equilibria. Chapter 5 gives the counterparts of the basic results of Chapter 2 for a more general iteration, which has a differential inclusion as a limit rather than an o.d.e. This is useful in many practical in-

† Wasan (1969) and Nevelson and Khasminskii (1976) are two early texts on stochastic approximation, though with a different flavour. See also Ljung et al. (1992).

stances, which are also described in this chapter. Chapter 6 analyzes the cases when more than one timescale is used. This chapter, notably the sections on 'averaging the natural timescale', is technically a little more difficult than the rest and the reader may skip the details of the proofs on a first reading. Chapter 7 describes the distributed asynchronous implementations of the algorithm. Chapter 8 describes the functional central limit theorem for fluctuations associated with the basic scheme of Chapter 2. All the above chapters use decreasing stepsizes. Chapter 9 briefly describes the corresponding theory for constant stepsizes which are popular in some applications.

Chapter 10 of the book has a different flavour: it collects together several examples from engineering, economics, etc., where the stochastic approximation formalism has paid rich dividends. Thus the general techniques of the first part of the book are specialized to each case of interest and the additional structure available in the specific problem under consideration is exploited to say more, depending on the context. It is a mixed bag, the idea being to give the reader a flavour of the various 'tricks of the trade' that may come in handy in future applications. Broadly speaking, one may classify these applications into three strands. The first is the stochastic gradient scheme and its variants wherein h above is either the negative gradient of some function or something close to the negative gradient. This scheme is the underlying paradigm for many adaptive filtering, parameter estimation and stochastic optimization schemes in general. The second is the o.d.e. version of fixed point iterations, i.e., successive application of a map from a space to itself so that it may converge to a point that remains invariant under it (i.e., a fixed point). These are important in a class of applications arising from dynamic programming. The third is the general collection of o.d.e.s modelling collective phenomena in economics etc., such as the urn example above. This classification is, of course, not exhaustive and some instances of stochastic approximation in practice do fall outside of this. Also, we do not consider the continuous time analog of stochastic approximation (see, e.g., Mel'nikov, 1996).

The background required for this book is a good first course on measure theoretic probability, particularly the theory of discrete parameter martingales, at the level of Breiman (1968) or Williams (1991) (though we shall generally refer to Borkar (1995), more out of familiarity than anything), and a first course on ordinary differential equations at the level of Hirsch, Smale and Devaney (2003). There are a few spots where something more than this is required, viz., the theory of weak (Prohorov) convergence of probability measures. The three appendices in Chapter 11 collect together the key aspects of these topics that are needed here.