5 The Stochastic Approximation Algorithm

5.1 Stochastic Processes – Some Basic Concepts

5.1.1 Random Variables and Random Sequences

Let $(\Omega, \mathcal{F}, P)$ be a probability space, namely:

- Ω is the sample space.
- $\mathcal{F}$ is the event space. Its elements are subsets of Ω, and it is required to be a σ-algebra (includes $\emptyset$ and Ω; includes all countable union of its members; includes all complements of its members).
- P is the probability measure (assigns a probability in [0,1] to each element of $\mathcal{F}$, with the usual properties: $P(\Omega) = 1$, countably additive).

A random variable (RV) X on $(\Omega, \mathcal{F})$ is a function $X : \Omega \to \mathbb{R}$, with values $X(\omega)$. It is required to be *measurable* on $\mathcal{F}$, namely, all sets of the form $\{\omega : X(\omega) \leq a\}$ are events in $\mathcal{F}$.

A vector-valued RV is a vector of RVs. Equivalently, it is a function $X : \Omega \to \mathbb{R}^d$, with similar measurability requirement.

A *random sequence*, or a discrete-time *stochastic process*, is a sequence $(X_n)_{n\geq 0}$ of $\mathbb{R}^d$-valued RVs, which are all defined on the same probability space.

5.1.2 Convergence of Random Variables

A random sequence may converge to a random variable, say to X. There are several useful notions of convergence:

1. Almost sure convergence (or: convergence with probability 1):

$$X_n \stackrel{a.s.}{\to} X \quad \text{if} \quad P\{\lim_{n\to\infty} X_n = X\} = 1\,.$$

2. Convergence in probability:

$$X_n \stackrel{p}{\to} X \quad \text{if} \quad \lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0\,, \forall \epsilon > 0\,.$$

3. Mean-squares convergence (convergence in L^2):

$$X_n \stackrel{L^2}{\to} X \quad \text{if} \quad E|X_n - X_\infty|^2 \to 0\,.$$

4. Convergence in Distribution:

$$X_n \stackrel{Dist}{\to} X \text{ (or } X_n \Rightarrow X) \quad \text{if} \quad Ef(X_n) \to Ef(X)$$

   for every bounded and continuous function f.

The following relations hold:

a. Basic implications: (a.s. or L^2) $\Longrightarrow p \Longrightarrow$ Dist

b. Almost sure convergence is equivalent to

$$\lim_{n\to\infty} P\{\sup_{k\geq n} |X_k - X| > \epsilon) = 0\,, \quad \forall\epsilon > 0\,.$$

c. A useful *sufficient* condition for a.s. convergence:

$$\sum_{n=0}^{\infty} P(|X_n - X| > \epsilon) < \infty\,.$$

5.1.3 Sigma-algebras and information

Sigma algebras (or σ-algebras) are part of the mathematical structure of probability theory. They also have a convenient interpretation as "information sets", which we shall find useful.

- Define $\mathcal{F}_X \triangleq \sigma\{X\}$, the σ-algebra generated by the RV X. This is the smallest σ-algebra that contains all sets of the form $\{X \leq a\} \equiv \{\omega \in \Omega : X(\omega) \leq a\}$.
- We can interpret $\sigma\{X\}$ as carrying all the information in X. Accordingly, we identify

  $$E(Z|X) \equiv E(Z|\mathcal{F}_X)\,.$$

  Also, "Z is measurable on $\sigma\{X\}$" is equivalent to: $Z = f(X)$ (with the additional technical requirement that f is a Borel measurable function).
- We can similarly define $\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}$, etc. Thus,

  $$E(Z|X_1, \ldots, X_n) \equiv E(Z|\mathcal{F}_n)\,.$$

- Note that $\mathcal{F}_{n+1} \supset \mathcal{F}_n$: more RVs carry more information, leading $\mathcal{F}_{n+1}$ to be finer, or "more detailed"

5.1.4 Martingales

A sequence $(X_k, \mathcal{F}_k)_{k\geq 0}$ on a given probability space $(\Omega, \mathcal{F}, P)$ is a martingale if

a. $(\mathcal{F}_k)$ is a "filtration" – an increasing sequence of σ-algebras in $\mathcal{F}$.

b. Each RV X_k is $\mathcal{F}_k$-measurable.

c. $E(X_{k+1}|\mathcal{F}_k) = X_k$ (P–a.s.).

Note that

- (a) Property is roughly equivalent to:

  $\mathcal{F}_k$ represents (the information in) some RVs $(Y_0, \ldots, Y_k)$,

  and (b) then means: X_k is a function of $(Y_0, \ldots, Y_k)$.

- A particular case is $\mathcal{F}_n = \sigma\{X_1, \ldots, X_n\}$ (a self-martingale).
- The central property is (c), which says that the conditional mean of X_{k+1} equals X_k. This is obviously stronger than $E(X_{k+1}) = E(X_k)$.
- The definition sometimes requires also that $E|X_n| < \infty$, we shall assume that below.
- Replacing (c) by $E(X_{k+1}|\mathcal{F}_k) \geq X_k$ gives a *submartingale*, while $E(X_{k+1}|\mathcal{F}_k) \leq X_k$ corresponds to a *supermartingale*.

Examples:

a. The simplest example of a martingale is

$$X_k = \sum_{\ell=0}^{k} \xi_\ell \,,$$

with $\{\xi_k\}$ a sequence of 0-mean independent RVs, and $\mathcal{F}_k = \sigma(\xi_0, \ldots, \xi_k)$.

b. $X_k = E(X|\mathcal{F}_k)$, where (F_k) is a given filtration and X a fixed RV.

Martingales play an important role in the convergence analysis of stochastic processes. We quote a few basic theorems (see, for example: A.N. Shiryaev, *Probability,* Springer, 1996).

Martingale Inequalities

Let$(X_k, \mathcal{F}_k)_{k\geq 0}$ be a martingale. Then for every $\lambda > 0$ and $p \geq 1$

$$P\left\{\max_{k\leq n} |X_k| \geq \lambda\right\} \leq \frac{E|X_n|^p}{\lambda^p}$$

and for $p > 1$

$$E[(\max_{k\leq n} |X_k|)^p] \leq (\tfrac{p}{p-1})^p E(|X_n|^p) \,.$$

Martingale Convergence Theorems

1. *Convergence with Bounded-moments*: Consider a martingale $(X_k, \mathcal{F}_k)_{k\geq 0}$. Assume that:

   $$E|X_k|^q \leq C \text{ for some } C < \infty,\ q \geq 1 \text{ and all } k.$$

   Then $\{X_k\}$ converges (a.s.) to a RV X_∞ (which is finite w.p. 1).

2. *Positive Martingale Convergence*: If $(X_k, \mathcal{F}_k)$ is a positive martingale (namely $X_n \geq 0$), then X_k converges (a.s.) to some RV X_∞.

Martingale Difference Convergence

The sequence $(\xi_k, \mathcal{F}_k)$ is a *martingale difference* sequence if property (c) is replaced by $E(\xi_{k+1}|\mathcal{F}_k) = 0$. In this case we have:

3. Suppose that for some $0 < q \leq 2$, $\sum_{k=1}^{\infty} \frac{1}{k^q} E(|\xi_k|^q | F_{k-1}) < \infty$ (a.s.).
   Then $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \xi_k = 0$ (a.s.).

For example, the conclusion holds if the sequence (ξ_k) is bounded, namely $|\xi_k| \leq C$ for some $C > 0$ (independent of k).

Note:

- It is trivially seen that $(\xi_n \triangleq X_n - X_{n-1})$ is a martingale difference if (X_n) is a martingale.
- More generally, for any sequence (Y_k) and filtration $(\mathcal{F}_k)$, where Y_k is measurable on $\mathcal{F}_k$, the following is a martingale difference:

  $$\xi_k \triangleq Y_k - E(Y_k|\mathcal{F}_{k-1}) .$$

  The conditions of the last theorem hold for this ξ_k if either:
  (i) $|Y_k| \leq M \ \forall k$ for some constant $M < \infty$,
  (ii) or, more generally, $E(|Y_k|^q|\mathcal{F}_{k-1}) \leq M$ (a.s.) for some $q > 1$ and a finite RV M.
  In that case we have

  $$\frac{1}{n}\sum_{k=1}^{n} \xi_k \;\equiv\; \frac{1}{n}\sum_{k=1}^{n}(Y_k - E(Y_k|\mathcal{F}_{k-1})) \to 0 \quad \text{(a.s.)}$$

5.2 The Basic SA Algorithm

The stochastic approximations (SA) algorithm essentially solves a system of (nonlinear) equations of the form

$$h(\theta) = 0$$

based on noisy measurements of $h(\theta)$.

More specifically, we consider a (continuous) function $h : \mathbb{R}^d \to \mathbb{R}^d$, with $d \geq 1$, which depends on a set of parameters $\theta \in \mathbb{R}^d$. Suppose that h is unknown. However, for each θ we can measure $Y = h(\theta) + \omega$, where ω is some 0-mean noise. The classical SA algorithm (Robbins-Monro, 1951) is of the form

$$\begin{aligned} \theta_{n+1} &= \theta_n + \alpha_n Y_n \\ &= \theta_n + \alpha_n [h(\theta_n) + \omega_n], \quad n \geq 0\,. \end{aligned}$$

Here α_n is the algorithm the step-size, or *gain*.

Obviously, with zero noise ($\omega_n \equiv 0$) the stationary points of the algorithm coincide with the solutions of $h(\theta) = 0$. Under appropriate conditions (on α_n, h and ω_n) the algorithm indeed can be shown to converge to a solution of $h(\theta) = 0$.

References:

H. Kushner and G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer, 1997.

V. Borkar, *Stochastic Approximation: A Dynamic System Viewpoint*, Hindustan, 2008.

J. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*, Wiley, 2003.

Some examples of the SA algorithm:

a. *Average of an i.i.d. sequence:* Let $(Z_n)_{\geq 0}$ be an i.i.d. sequence with mean $\mu = E(Z_0)$ and finite variance. We wish to estimate the mean.

The iterative algorithm

$$\theta_{n+1} = \theta_n + \frac{1}{n+1}[Z_n - \theta_n]$$

gives

$$\theta_n = \frac{1}{n}\theta_0 + \frac{1}{n}\sum_{k=0}^{n-1} Z_k \to \mu \quad \text{(w.p. 1)}, \quad \text{by the SLLN.}$$

This is a SA iteration, with $\alpha_n = \frac{1}{n+1}$, and $Y_n = Z_n - \theta_n$. Writing $Z_n = \mu + \omega_n$ (Z_n is considered a noisy measurement of μ, with zero-mean noise ω_n), we can identify $h(\theta) = \mu - \theta$.

b. *Function minimization:* Suppose we wish to minimize a (convex) function $f(\theta)$. Denoting $h(\theta) = -\nabla f(\theta) \equiv -\frac{\partial f}{\partial \theta}$, we need to solve $h(\theta) = 0$.

The basic iteration here is

$$\theta_{n+1} = \theta_n + \alpha_n[-\nabla f(\theta) + \omega_n].$$

This is a "noisy" gradient descent algorithm.

When ∇f is not computable, it may be approximated by finite differences of the form

$$\frac{\partial f(\theta)}{\partial \theta_i} \approx \frac{f(\theta + e_i\delta_i) - f(\theta - e_i\delta_i)}{2\delta_i}.$$

where e_i is the i-th unit vector. This scheme is known as the "Kiefer-Wolfowitz Procedure".

Some variants of the SA algorithm

- *A fixed-point formulation:* Let $h(\theta) = H(\theta) - \theta$. Then $h(\theta) = 0$ is equivalent to the fixed-point equation $H(\theta) = \theta$, and the algorithm is

  $$\theta_{n+1} = \theta_n + \alpha_n[H(\theta_n) - \theta_n + \omega_n] = (1 - \alpha_n)\theta_n + \alpha_n[H(\theta_n) + \omega_n] \,.$$

  This is the form used in the Bertsekas & Tsitsiklis (1996) monograph.

  Note that in the average estimation problem (example a. above) we get $H(\theta) = \mu$, hence $Z_n = H(\theta_n) + \omega_n$.

- *Asynchronous updates:* Different components of θ may be updated at different times and rates. A general form of the algorithm is:

  $$\theta_{n+1}(i) = \theta_n(i) + \alpha_n(i)Y_n(i), \quad i = 1, \cdots, d$$

  where each component of θ is updated with a different gain sequence $\{\alpha_n(i)\}$. These gain sequences are typically required to be of comparable magnitude.

  Moreover, the gain sequences may be allowed to be *stochastic*, namely depend on the entire history of the process up to the time of update. For example, in the TD(0) algorithm θ corresponds to the estimated value function $\hat{V} = (\hat{V}(s),\, s \in S)$, and we can define $\alpha_n(s) = 1/N_n(s)$, where $N_n(s)$ is the number of visits to state s up to time n.

- *Projections:* If is often known that the required parameter θ lies in some set $B \subset \mathbb{R}^d$. In that case we could use the projected iterates:

  $$\theta_{n+1} = Proj_B[\theta_n + \alpha_n Y_n]$$

  where $Proj_B$ is some projection onto B.

  The simplest case is of course when B is a box, so that the components of θ are simply truncated at their minimal and maximal values.

  If B is a bounded set then the estimated sequence $\{\theta_n\}$ is guaranteed to be bounded in this algorithm. This is very helpful for convergence analysis.

5.3 Assumptions

Gain assumptions

To obtain convergence, the gain sequence needs to decrease to zero. The following assumption is standard.

Assumption G1: $\alpha_n \geq 0$, and

$$\text{(i)} \quad \sum_{n=1}^{\infty} \alpha_n = \infty$$

$$\text{(ii)} \quad \sum_{n=1}^{\infty} \alpha_n^2 < \infty\,.$$

A common example is $\alpha_n = \dfrac{1}{n^a}$, with $\frac{1}{2} < a \leq 1$.

Noise Assumptions

In general the noise sequence $\{\omega_n\}$ is required to be "zero-mean", so that it will average out.

Since we want to allow dependence of ω_n on θ_n, the sequence $\{\omega_n\}$ cannot be assumed independent. The assumption below allows $\{\omega_n\}$ to be a martingale difference sequence.

Let

$$\mathcal{F}_{n-1} = \sigma\{\theta_0, \alpha_0, \omega_0, \cdots, \omega_{n-1}; \theta_n, \alpha_n\}$$

denote the (σ-algebra generated by) the history sequence up to step n. Note that ω_n is measurable on $\mathcal{F}_n$ by definition of the latter.

Assumption N1

(a) The noise sequence $\{\omega_n\}$ is a martingale difference sequence relative to the filtration $\{F_n\}$, namely

$$E(\omega_n | \mathcal{F}_{n-1}) = 0 \qquad \text{(a.s.)}.$$

(b) For some finite constants A, B and some norm $\|\cdot\|$ on $\mathbb{R}^d$,

$$E(\|\omega_n\|^2 | \mathcal{F}_{n-1}) \leq A + B\|\theta_n\|^2 \quad \text{(a.s.)}, \quad \forall n \geq 1\,.$$

Example: Let $\omega_n \sim N(0, \sigma_n)$, where σ_n may depend on θ_n, namely $\sigma_n = f(\theta_n)$. Formally,

$$
\begin{aligned}
E(\omega_n | F_n) &= 0 \\
E(\omega_n^2 | F_n) &= f(\theta_n)^2,
\end{aligned}
$$

and we require that $f(\theta)^2 \leq A + B\theta^2$.

Note: When $\{\theta_n\}$ is known to be bounded, then (b) reduces to

$$E(\|\omega_n\|^2 | \mathcal{F}_{n-1}) \leq C \quad \text{(a.s.)} \quad \forall n$$

for some $C < \infty$. It then follows by the martingale difference convergence theorem that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \omega_k = 0 \quad \text{(a.s.)}.$$

However, it is often the case that θ is not known to be bounded *a-priori*.

Markov Noise: The SA algorithm may converge under more general noise assumptions, which are sometimes useful. For example, for each fixed θ, ω_n may be a *Markov chain* such that its long-term average is zero (but $E(\omega_n | \mathcal{F}_{n-1}) \neq 0$). We shall not go into that generality here.

5.4 The ODE Method

The asymptotic behavior of the SA algorithm is closely related to the solutions of a certain ODE (Ordinary Differential Equation), namely

$$\frac{d}{dt}\theta(t) = h(\theta(t))$$

or $\dot{\theta} = h(\theta)$.

Given $\{\theta_n, \alpha_n\}$, we define a *continuous-time* process $\theta(t)$ as follows. Let

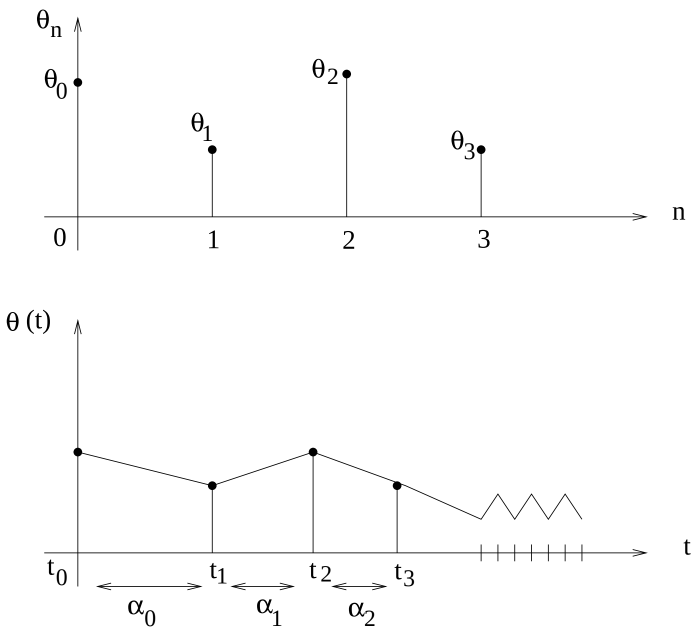
$$t_n = \sum_{k=0}^{n-1} \alpha_k \, .$$

Define

$$\theta(t_n) = \theta_n \, ,$$

and use linear interpolation in-between the t_n's.

Thus, the time-axis t is rescaled according to the gains $\{\alpha_n\}$.



Note that over a fixed Δt, the "total gain" is approximately constant:

$$\sum_{k \in K(t,\Delta t)} \alpha_k \simeq \Delta t \, ,$$

where $K(t, \Delta t) = \{k : t \leq t_k < t + \Delta t\}$.

Now:

$$\theta(t + \Delta t) = \theta(t) + \sum_{k \in k(t,\Delta t)} \alpha_k [h(\theta_n) + \omega_n] \,.$$

- For t large, α_k becomes small and the summation is over many terms; thus the noise term is approximately "averaged out": $\sum \alpha_k \omega_k \to 0$.
- For Δt small, θ_k is approximately constant over $K(t, \Delta t) : h(\theta_k) \simeq h(\theta(t))$.

We thus obtain:

$$\theta(t + \Delta t) \simeq \theta(t) + \Delta t \cdot h(\theta(t)) \,.$$

For $\Delta t \to 0$, this reduces to the ODE:

$$\dot{\theta}(t) = h(\theta(t)) \,.$$

To conclude:

- As $n \to \infty$, we "expect" that the estimates $\{\theta_n\}$ will follow a trajectory of the ODE $\dot{\theta} = h(\theta)$ (under the above time normalization).
- Note that the stationary point(s) of the ODE are given by $\theta^* : \quad h(\theta^*) = 0$.
- An obvious requirement for $\theta_n \to \theta^*$ is $\theta(t) \to \theta^*$ (for any $\theta(0)$). That is: θ^* is a *globally asymptotically stable* equilibrium of the ODE.

  This may this be viewed as a necessary condition for convergence of θ_n. It is also sufficient under additional assumptions on h (continuity, smoothness), and boundedness of $\{\theta_n\}$.

5.5 Some Convergence Results

A typical convergence result for the (synchronous) SA algorithm is the following:

Theorem 1 Assume G1, N1, and furthermore:

(i) h is Lipschitz continuous.

(ii) The ODE $\dot{\theta} = h(\theta)$ has a unique equilibrium point θ^*, which is globally asymptotically stable.

(iii) The sequence (θ_n) is bounded (with probability 1).

Then $\theta_n \to \theta^*$ (w.p. 1), for any initial conditions θ_0.

Remarks:

1. More generally, even if the ODE is not globally stable, θ_n can be shown to converge to an *invariant set* of the ODE (e.g., a limit cycle).
2. Corresponding results exist for the asynchronous versions, under suitable assumptions on the relative gains.
3. A major assumption in the last result in the boundedness of (θ_n). In general this assumption has to be verified independently. However, there exist several results that rely on further properties of h to deduce boundedness, and hence convergence.

The following convergence result from B. &T. (1996) relies on on contraction properties of H, and applies to the asynchronous case. It will directly apply to some of our learning algorithms. We start with a few definitions.

- Let $H(\theta) = h(\theta) + \theta$, so that $h(\theta) = H(\theta) - \theta$.
- Recall that $H(\theta)$ is a *contraction operator* w.r.t. a norm $\|\cdot\|$ if

$$\|H(\theta_1) - H(\theta_2)\| \leq \alpha \|\theta_1 - \theta_2\|$$

  for some $\alpha < 1$ and all θ_1, θ_2.
- $H(\theta)$ is a *pseudo-contraction* if the same holds for a fixed $\theta_2 = \theta^*$. It easily follows then that θ^* is a unique fixed point of H.

- Recall that the *max-norm* is given by $\|\theta\|_\infty = \max_i |\theta(i)|$. The *weighted* max-norm, with a weight vector w, $w(i) > 0$, is given by

$$\|\theta\|_w = \max_i \{ \frac{|\theta(i)|}{w(i)} \} .$$

Theorem 2 (Prop. 4.4. in B.&T). Let

$$\theta_{n+1}(i) = \theta_n(i) + \alpha_n(i)[H(\theta_n) - \theta_n + \omega_n]_i \,, \quad i = 1, \cdots, d \,.$$

Assume N1, and:

(a) Gain assumption: $\alpha_n(i) \geq 0$, measurable on the "past", and satisfy

$$\sum_n \alpha_n(i) = \infty, \quad \sum_n \alpha_n(i)^2 < \infty \quad \text{(w.p. 1)} \,.$$

(b) H is a pseudo-contraction w.r.t. some weighted max-norm.

Then $\theta_n \to \theta^*$ (w.p. 1), where θ^* is the unique fixed point of H.

Remark on "Constant Gain" Algorithms

As noted before, in practice it is often desirable to keep a non-diminishing gain. A typical case is $\alpha_n(i) \in [\underline{\alpha}, \bar{\alpha}]$.

Here we can no longer expect "w.p. 1" convergence results. What can be expected is a statement of the form:

- For $\bar{\alpha}$ small enough, we have for all $\epsilon > 0$

$$\limsup_{n \to \infty} P(\|\theta_n - \theta^*\| > \epsilon) \leq b(\epsilon) \cdot \bar{\alpha} \,,$$

  with $b(\epsilon) < \infty$.

This is related to "convergence in probability", or "weak convergence". We shall not give a detailed account here.