# Part VI
# Appendix

This part puts together five appendices on (a) convergence notions for a sequence of random vectors, (b) results on martingales and their convergence, (c) ordinary differential equations, (d) the Borkar and Meyn stability result, and (e) a result on convergence of projected stochastic approximation due to Kushner and Clark.Some of the background material as well as the main results used in other chapters have been summarized here.

# Appendix A
# Convergence Notions for a Sequence of Random Vectors

We briefly discuss here the various notions of convergence for random vectors. Let $(\Omega, \mathscr{F}, P)$ denote the underlying probability space, where $\Omega$ is the sample set, $\mathscr{F}$ the sigma field and $P$ the probability measure, see for instance, [1] for a good account of probability theory. Let $X_n, n \geq 0$ denote a sequence of $\mathbb{R}^N$–valued random vectors on $(\Omega, \mathscr{F}, P)$. Suppose $X$ is another $\mathbb{R}^N$–valued random vector on $(\Omega, \mathscr{F}, P)$. Further, let $x \in \mathbb{R}^N$ be an $N$-dimensional vector. Let $F_{X_n}(\cdot)$, $F_X(\cdot)$, $n \geq 0$ denote the corresponding distribution functions associated with the random vectors $X_n$, $X$, $n \geq 0$. Suppose $X_n = (X_n^1, \ldots, X_n^N)^T$, $X = (X^1, \ldots, X^N)^T$ and $x = (x^1, \ldots, x^N)^T$, respectively, where $X_n^i$, $X^i$, $x^i$, $i = 1, \ldots, N$ are $\mathbb{R}$-valued. Then $F_{X_n}(x) = P(X_n^i \leq x^i, i = 1, \ldots, N)$ and $F_X(x) = P(X^i \leq x^i, i = 1, \ldots, N)$, respectively. The following are standard notions of convergence:

1. **Deterministic Convergence:** We say that $X_k \to X$ as $k \to \infty$ deterministically if $X_k(w) \to X(w)$ as $k \to \infty$ for all $w \in \Omega$.
2. **Uniformly:** $X_k \to X$ uniformly as $k \to \infty$ if for all $\varepsilon > 0$ there exists an $N > 1$ such that $\forall n \geq N, \forall w \in \Omega$, $\|X_k(w) - X(w)\| < \varepsilon$.
3. **Almost Sure (a.s.) or With Probability One (w.p.1) Convergence:** We say that $X_k \to X$ as $k \to \infty$ almost surely (a.s.) or with probability one (w.p.1) if
$$P\left(w \in \Omega \mid \lim_{k \to \infty} \| X_k(w) - X(w) \| = 0\right) = 1.$$
4. **Probabilistic or In Probability Convergence:** We say that $X_k \to X$ as $k \to \infty$ probabilistically or in probability if
$$\lim_{k \to \infty} P(w \in \Omega \mid \| X_k(w) - X(w) \| \geq \varepsilon) = 0 \; \forall \varepsilon > 0.$$
5. **Convergence in $L^p$:** Let for $p \geq 1$,
$$L^p(\Omega, \mathscr{F}, P) = \{X | E|X|^p < \infty\},$$

denote a set of all $\mathbb{R}^N$-valued random variables on $(\Omega, \mathscr{F}, P)$ which have finite $p^{\text{th}}$ moment. We say that $X_k \to X$ as $k \to \infty$ in $L^p$ for $p \geq 1$, if

$$\lim_{k\to\infty} E\left(\| X_k(w) - X(w) \|^p\right) = 0.$$

When $p = 2$, the $L^p$ convergence is also referred to as *mean-square convergence*.

6. **In Distribution Convergence:** We say that $X_k \to X$ as $k \to \infty$ in distribution if

$$\lim_{k\to\infty} F_{X_k}(x) = F_X(x) \text{ at all points } x \text{ of continuity of } F_X(x).$$

7. **Nearly uniformly:** $X_k \to X$ nearly uniformly as $k \to \infty$ if $\forall \varepsilon > 0, \exists A \in F$ such that $P(A) < \varepsilon$ and on $A^c, X_k \to X$ uniformly.

**Theorem A.1 (Egorov).** *If $X_k \xrightarrow{a.s.} X$, then, $X_k \xrightarrow{n.u.} X$. The result is true for any measure $\mu$ with $\mu(\Omega) < \infty$.*

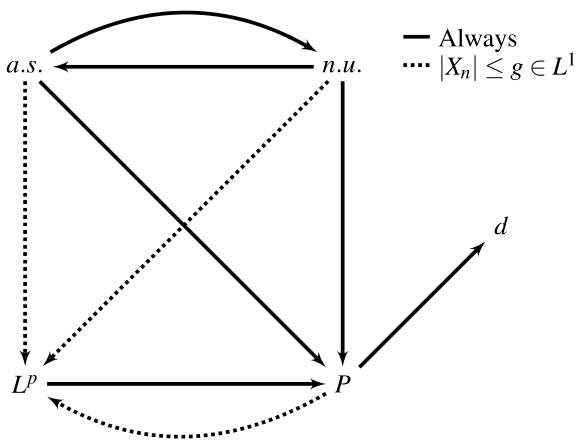


**Fig. A.1** Relationship between the various notions of convergence. We use the following abbreviations - a.s. to denote “almost surely”, n.u. for “nearly uniformly”, P for “convergence in probability”, d for “convergence in distribution”, $L^p$ for “convergence in $L^p$”.

A general relationship between the various notions of convergence is shown in Figure A.1. In the figure, a directed arrow from “A” to “B” i.e., $A \to B$ indicates that “A is stronger than B”. Further, we assume that a transitivity property holds in that $A \to B$ and $B \to C$ implies that $A \to C$, even when an arrow from “A” to “C” is not explicitly shown. Note that a.s. or w.p.1 convergence implies that there exists a set of zero probability on which the said convergence does not hold. Deterministic convergence can be viewed as a special case of a.s. convergence as here the above zero-probability set is in fact empty. Also, while in Figure A.1, there is no arrow between a.s. convergence and m.s. convergence, the former implies the latter under certain conditions on the random vectors $X_n, X$, $n \geq 1$. As an example, if the said random vectors are uniformly bounded by a $L^1$ function, then $L^p$ convergence follows from a.s. convergence.

## Reference

1. Chow, Y.S., Teicher, H.: Probability Theory: Independence, Interchangeability and Martingales, 3rd edn. Springer, New York (1997)

# Appendix B
# Martingales

As before, let $(\Omega, \mathscr{F}, P)$ be a given probability space. Let $\{\mathscr{F}_n\}$ be a family of increasing sub-$\sigma$−fields of $\mathscr{F}$ (also called a filtration), i.e.,

$$\mathscr{F}_0 \subset \mathscr{F}_1 \subset \mathscr{F}2_{\subset} \cdots \subset \mathscr{F}_n \subset \mathscr{F}_{n+1} \subset \cdots \subset \mathscr{F}.$$

**Definition B.1.** 1. A sequence of $\mathscr{R}$-valued random variables $X_n, n \geq 0$ defined on $(\Omega, \mathscr{F}, P)$ is said to be a martingale w.r.t. the filtration $\{\mathscr{F}_n\}$ if each $X_n$ is integrable and measurable with respect to $\mathscr{F}_n$.
2. Further,

$$E[X_{n+1} \mid \mathscr{F}_n] = X_n \text{ w.p.1 } \forall n \geq 0. \tag{B.1}$$

**Definition B.2.** A sequence of random variables $X_n, n \geq 0$ as in Definition 1 is said to be a submartingale w.r.t. the filtration $\{\mathscr{F}_n\}$ if the first part in Definition 1 holds. In addition, the equality in (B.1) is replaced with "$\geq$".

**Definition B.3.** A sequence of random variables $X_n, n \geq 0$ as in Definition 1 is said to be a submartingale w.r.t. the filtration $\{\mathscr{F}_n\}$ if the first part in Definition 1 holds. In addition, the equality in (B.1) is replaced with "$\leq$".

Many times, one identifies the martingale (alternatively, sub- or super-martingale) with the sequence of tuples $(X_n, \mathscr{F}_n)$, $n \geq 0$ instead of just $\{X_n\}$ itself.

**Definition B.4.** For a martingale sequence $X_n, n \geq 0$, the sequence $M_{n+1}$, $n \geq 0$ obtained as $M_{n+1} = (X_{n+1} - X_n)$, $n \geq 0$ with $M_0 = X_0$, is called a martingale difference sequence.

Note that

$$\begin{aligned} E[M_{n+1} \mid \mathscr{F}_n] &= E[(X_{n+1} - X_n) \mid \mathscr{F}_n] \\ &= (E[X_{n+1} \mid \mathscr{F}_n] - X_n) = 0 \text{ w.p.1}, \end{aligned}$$

from (B.1).

**Definition B.5.** A vector martingale (also many times referred to as a martingale) is a sequence of $\mathscr{R}^N-$valued random vectors $X_n = (X_n^1, \ldots, X_n^N)$ such that each of its component processes $X_n^i, n \geq 0$ $(i = 1, \ldots, N)$ is a martingale.

We recall the following important result due to Doob, see, for instance, [1, Theorem 3.2.2 on pp. 49].

**Theorem B.1 (Doob decomposition).** *A submartingale $(X_n, \mathscr{F}_n), n \geq 0$, can be decomposed as $X_n = Y_n + A_n$, $n \geq 0$, where $(Y_n, \mathscr{F}_n)$, $n \geq 0$, is a zero-mean martingale and $A_n, n \geq 0$ is a non-decreasing process, i.e., $A_n \leq A_{n+1}$ almost surely for all $n \geq 0$. Further, $A_n$ is $\mathscr{F}_{n-1}$-measurable for all $n \geq 0$, where $\mathscr{F}_{-1} = \{\phi, \Omega\}$. This decomposition is almost surely unique.*

There are various convergence results for martingales but the one that we often use in this book is based on the convergence of the quadratic variation process associated with the martingale $X_n, n \geq 0$ (see below). Let $X_n$, $n \geq 0$ be a square integrable (scalar) martingale, i.e., it is a martingale for which $E[X_n^2] < \infty$ for all $n \geq 0$. It is easy to see that in this case, $(X_n^2, \mathscr{F}_n)$, $n \geq 0$ forms a submartingale. Hence from the Doob decomposition theorem (cf. Theorem B.1), it follows that $X_n^2 = Y_n + A_n$, $n \geq 0$, where $\{Y_n\}$ and $\{A_n\}$ satisfying the properties in Theorem B.1. It is easy to see that

$$\begin{aligned} A_n &= \sum_{m=1}^{n} \left( E\left[X_m^2 \mid \mathscr{F}_{m-1}\right] - X_{m-1}^2 \right) + E\left[X_0^2\right] \\ &= \sum_{m=0}^{n-1} E\left[ (X_{m+1} - X_m)^2 \mid \mathscr{F}_m \right] + E\left[X_0^2\right], \end{aligned} \tag{B.2}$$

$\forall n \geq 0$. As mentioned above, $A_n, n \geq 0$ is called the quadratic variation process associated with the martingale $X_n, n \geq 0$.

**Theorem B.2 (Martingale Convergence Theorem).** *Let $(X_n, \mathscr{F}_n)$, $n \geq 0$ be a square-integrable martingale with $A_n, n \geq 0$ as its quadratic variation process. Let $A_\infty = \lim_{n\to\infty} A_n$. Then $\{X_n\}$ converges with probability one on the set $\{A_\infty < \infty\}$ and $X_n = o(f(A_n))$ on $\{A_\infty = \infty\}$ for every increasing $f : [0, \infty) \to [0, \infty)$ satisfying $\int_0^\infty (1 + f(t))^{-2} dt < \infty$.*

The proof of this result is available for instance on pp. 53-54 of [1] (cf. Theorem 3.3.4). Detailed treatments of martingales can be found, for instance, in the texts of Breiman [2], Neveu [3] and Borkar [1].

## References

1. Borkar, V.S.: Probability Theory: An Advanced Course. Springer, New York (1995)
2. Breiman, L.: Probability. Addison-Wesley, Reading (1968)
3. Neveu, J.: Discrete Parameter Martingales. North Holland, Amsterdam (1975)

# Appendix C
# Ordinary Differential Equations

We begin with a definition of the $O$ and $o$ notation as this has been used at various places the text.

**Definition C.1.** Let $\{a_n\}$ and $\{b_n\}$ be two sequences of real numbers such that $b_n \geq 0, \forall n$.

1. We say $a_n = O(b_n)$ if there exists a constant $L > 0$ such that $|a_n| \leq Lb_n$ for all $n$.
2. We say $a_n = o(b_n)$ if $\lim_{n\to\infty} \frac{a_n}{b_n} = 0$.

**Definition C.2.** A function $h : \mathscr{R}^d \to \mathscr{R}^d$ is said to be Lipschitz continuous if $\exists M > 0$ such that

$$\| h(x) - h(y) \| \leq M \| x - y \|, \ \forall x, y \in \mathscr{R}^d.$$

The Gronwall inequality plays an important role in the proof of convergence of stochastic approximation algorithms. We give the result below, whose proof can be found in several texts, see for instance, Appendix B of [1].

**Lemma C.1 (Gronwall inequality).** *For continuous functions* $f(\cdot), g(\cdot) \geq 0$ *and scalars* $K_1, K_2, T \geq 0$,

$$f(t) \leq K_1 + K_2 \int_0^t f(s)g(s)ds \ \ \forall t \in [0,T], \tag{C.1}$$

*implies*

$$f(t) \leq K_1 e^{K_2 \int_0^T g(s)ds}, \ t \in [0,T].$$

Consider the ODE given by

$$\dot{\theta}(t) = L(\theta(t)), \ \theta(0) = \theta_0. \tag{C.2}$$

**Definition C.3.** The ODE (C.2) is said to be well-posed if starting from any $\theta(0) = \theta_0$, the trajectory $\theta(\cdot) = \{\theta(t), t \geq 0\}$ of (C.2) is unique. Further, the map $\theta_0 \to \theta(\cdot)$ is continuous.

**Theorem C.2.** *A sufficient condition for (C.2) to be well-posed is if the function $L : \mathbb{R}^N \to \mathbb{R}^N$ is Lipschitz continuous.*

*Proof.* See Theorem 5 on pp.143 of [1]. □

**Definition C.4.** A closed set $H \subset \mathbb{R}^N$ is called an invariant set for the ODE (C.2) if whenever the initial point $\theta(0) \in H$, then $\theta(t) \in H$ for all $t \geq 0$, i.e., if the ODE trajectory is initiated in $H$, it stays in $H$ for all time.

**Definition C.5.** A closed set $H \subset \mathbb{R}^N$ is called an attractor for the ODE (C.2) if

(i) $H$ is an invariant set, and
(ii) there is an open set $M$ containing $H$ (i.e., $M$ is an open neighborhood of $H$) such that if the ODE trajectory is initiated in $M$, it stays in $M$ and converges to $H$.

**Definition C.6.** The largest possible open set $M$ that is an open neighborhood of $H$ such that any ODE trajectory initiated in $M$ stays in $M$ and converges to $H$ is called the Domain of Attraction of $H$.

Given $\eta > 0$, let

$$H^\eta = \{\theta \in \mathbb{R}^N \mid \| \theta - \bar{\theta} \| < \eta,$$

denote the $\eta$-neighborhood of $H$, i.e., the set of all points within a distance $\eta$ from the set $H$.

**Definition C.7.** A closed invariant set $H$ is Lyapunov stable if for any $\varepsilon > 0$, there exists $\delta > 0$ such that every trajectory initiated in $H^\delta$ stays in $H^\varepsilon$ for all time (i.e., if $\theta(0) \in H^\delta$, then $\theta(t) \in H^\varepsilon$ for all $t$).

**Definition C.8.** A closed invariant set $H$ is asymptotically stable if it is both Lyapunov stable and an attractor.

**Definition C.9.** A closed invariant set $H$ is globally asymptotically stable if $H$ is asymptotically stable and an attractor. All trajectories of the ODE in this case converge to $H$. Thus, the domain of attraction of $H$ when it is globally asymptotically stable is $\mathbb{R}^N$.

The following theorem gives a criterion to verify asymptotic stability of the set $H$.

**Theorem C.3.** *The set $H$ is asymptotically stable for the ODE (C.2) if one can find a function $V : \mathbb{R}^N \to \mathbb{R}$ such that the following hold:*

*(i)* $V(\theta) \geq 0 \; \forall \theta \in \mathbb{R}^N$,

*(ii) There exists an open neighborhood M of H such that* $V(\theta) \to \infty$ *as* $\theta \to \partial O$ *(i.e., the boundary of M),*

*(iii)* $\dfrac{dV(\theta(t))}{dt} = \nabla V(\theta(t))^T \dot{\theta}(t) = \nabla V(\theta(t))^T L(\theta(t)) \leq 0, \; \forall \theta(\cdot) \in M.$

*In particular,* $\dfrac{dV(\theta(t))}{dt} = 0$ *if and only if* $\theta(t) \in H$.

The following result on convergence of an ODE trajectory is due to Lasalle [3].

**Theorem C.4 (Lasalle Invariance Theorem).** *Let H be the globally asymptotically stable attractor set for the ODE (C.2). Let* $V : \mathbb{R}^N \to \mathbb{R}$ *be a function such that* $V(\theta) \geq 0 \; \forall \theta \in \mathbb{R}^N$. *Further,* $V(\theta) \to \infty$ *as* $\| \theta \| \to \infty$ *and* $\nabla V(\theta)^T L(\theta) \leq 0 \; \forall \theta$. *Then any trajectory* $\theta(\cdot)$ *must converge to the largest invariant set contained in*

$$\{\theta \mid \nabla V(\theta)^T L(\theta) = 0\}.$$

**Definition C.10 (($T,\Delta$)-perturbation).** Given $T$, $\Delta > 0$, we call a bounded, measurable $y(\cdot) : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^N$, a $(T,\Delta)$-perturbation of (C.2) if there exist $0 = T_0 < T_1 < T_2 < \cdots < T_r \uparrow \infty$ with $T_{r+1} - T_r \geq T \; \forall r$ and solutions $\theta^r(t)$, $t \in [T_r, T_{r+1}]$ of (C.2) for $r \geq 0$, such that

$$\sup_{t \in [T_r, T_{r+1}]} \| \theta^r(t) - y(t) \| < \Delta.$$

Again let $H$ be the globally asymptotically stable attractor set for (C.2) and $H^\varepsilon$ be the $\varepsilon$-neighborhood of $H$. The following result due to Hirsch [2] (Theorem 1, pp.339) describes convergence to $H^\varepsilon$ of a function that closely approximates the ODE trajectory.

**Lemma C.5 (Hirsch Lemma).** *Given* $\varepsilon$, $T > 0$, $\exists \bar{\Delta} > 0$ *such that for all* $\Delta \in (0, \bar{\Delta})$, *every* $(T,\Delta)$*-perturbation of (C.2) converges to* $H^\varepsilon$.

## References

1. Borkar, V.S.: Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press and Hindustan Book Agency (Jointly Published), Cambridge and New Delhi (2008)
2. Hirsch, M.W.: Convergent activation dynamics in continuous time networks. Neural Networks 2, 331–349 (1989)
3. Lasalle, J.P., Lefschetz, S.: Stability by Liapunov's Direct Method with Applications. Academic Press, New York (1961)

# Appendix D
# The Borkar-Meyn Theorem for Stability and Convergence of Stochastic Approximation

While there are various techniques to show stability of stochastic iterates, we review below the one by Borkar and Meyn [2] (see also [1], Chapter 3) as it is seen to be widely applicable in a large number of settings. They analyze the $N$-dimensional stochastic recursion

$$X_{n+1} = X_n + a(n)(h(X_n) + M_{n+1}),$$

under the following assumptions:

**Assumption D.1.**

(i) The function $h : \mathscr{R}^N \to \mathscr{R}^N$ is Lipschitz continuous and there exists a function $h_\infty : \mathscr{R}^N \to \mathscr{R}^N$ such that

$$\lim_{r\to\infty} \frac{h(rx)}{r} = h_\infty(x), x \in \mathscr{R}^N.$$

(ii) The origin in $\mathscr{R}^N$ is an asymptotically stable equilibrium for the ODE

$$\dot{x}(t) = h_\infty(x(t)). \tag{D.1}$$

(iii) There is a unique globally asymptotically stable equilibrium $x^* \in \mathscr{R}^N$ for the ODE D.1.

**Assumption D.2.** The sequence $\{M_n, \mathscr{G}_n, n \geq 1\}$ with $\mathscr{G}_n = \sigma(X_i, M_i, i \leq n)$ is a martingale difference sequence. Further for some constant $C_0 < \infty$ and any initial condition $X_0 \in \mathscr{R}^N$,

$$E[\| M_{n+1} \|^2 | \mathscr{G}_n] \leq C_0(1 + \| X_n \|^2),\ n \geq 0.$$

Further, the step-sizes $a(n), n \geq 0$ satisfy

$$a(n) > 0 \forall n, \ \sum_n a(n) = \infty, \ \sum_n a(n)^2 < \infty.$$

The main result of [2] (see Theorems 2.1(i)-2.2 of [2]) is the following:

**Theorem D.1 (Borkar and Meyn Theorem).** *Suppose Assumptions D.1 and D.2 hold. For any initial condition* $X_0 \in \mathscr{R}^N$, $\sup_n \parallel X_n \parallel < \infty$ *almost surely (a.s.). Further,* $X_n \to x^*$ *a.s. as* $n \to \infty$.

[2] also contains a result for bounded step-size sequences (not tapering to zero). However, for our purposes, we only require the result for diminishing step-sizes. Assumptions D.1 and D.2 are seen to be satisfied in many cases, for instance, in reinforcement learning algorithms.

## References

1. Borkar, V.S.: Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press and Hindustan Book Agency (Jointly Published), Cambridge and New Delhi (2008)
2. Borkar, V.S., Meyn, S.P.: The O.D.E. method for convergence of stochastic approximation and reinforcement learning. Journal of Control and Optimization 38(2), 447–469 (2000)

# Appendix E
# The Kushner-Clark Theorem for Convergence of Projected Stochastic Approximation

We review here an important result due to Kushner and Clark [3] (cf. Theorem 5.3.1 on pp. 191-196 of [3]) that shows the convergence of projected stochastic approximations. While the result, as stated in [3], is more generally applicable, we present its adaptation here that is relevant to the setting that we consider.

Let $C \subset \mathscr{R}^N$ be a compact and convex set and $\Gamma : \mathscr{R}^N \to C$ denote a projection operator that projects any $x = (x_1, \ldots, x_N)^T \in \mathscr{R}^N$ to its nearest point in $C$. Thus, if $x \in C$, then $\Gamma(x) \in C$ as well. For instance, if $C$ is an $N$-dimensional rectangle having the form $C = \prod_{i=1}^{N} [a_{i,\min}, a_{i,\max}]$, where $-\infty < a_{i,\min} < a_{i,\max} < \infty$, $\forall i = 1, \ldots, N$, then a convenient way to identify $\Gamma(x)$ is according to $\Gamma(x) = (\Gamma_1(x_1), \ldots, \Gamma_N(x_N))^T$, where the individual operators $\Gamma_i : \mathscr{R} \to \mathscr{R}$ are defined by $\Gamma_i(x_i) = \min(a_{i,\max}, \max(a_{i,\min}, x))$, $i = 1, \ldots, N$.

Consider the following the $N$-dimensional stochastic recursion

$$X_{n+1} = \Gamma(X_n + a(n)(h(X_n) + \xi_n + \beta_n)), \tag{E.1}$$

under the assumptions listed below. Also, consider the following ODE associated with (E.1):

$$\dot{X}(t) = \bar{\Gamma}(h(X(t))). \tag{E.2}$$

Let $\mathscr{C}(C)$ denote the space of all continuous functions from $C$ to $\mathscr{R}^N$. The operator $\bar{\Gamma} : \mathscr{C}(C) \to \mathscr{C}(\mathscr{R}^N)$ is defined according to

$$\bar{\Gamma}(v(x)) = \lim_{\eta \to 0} \left( \frac{\Gamma(x + \eta v(x)) - x}{\eta} \right), \tag{E.3}$$

for any continuous $v : C \to \mathscr{R}^N$. The limit in (E.3) exists and is unique since $C$ is a convex set. In case this limit is not unique, one may consider the set of all limit points of (E.3). Note also that from its definition, $\bar{\Gamma}(v(x)) = v(x)$ if $x \in C^o$ (the interior of $C$). This is because for such an $x$, one can find $\eta > 0$ sufficiently small so that $x + \eta v(x) \in C^o$ as well and hence $\Gamma(x + \eta v(x)) = x + \eta v(x)$. On the other hand, if $x \in \partial C$ (the boundary of $C$) is such that $x + \eta v(x) \notin C$, for any small $\eta > 0$, then $\bar{\Gamma}(v(x))$ is the projection of $v(x)$ to the tangent space of $\partial C$ at $x$.

Consider now the assumptions listed below.

**Assumption E.1.** The function $h : \mathscr{R}^N \to \mathscr{R}^N$ is continuous.

**Assumption E.2.** The step-sizes $a(n), n \geq 0$ satisfy

$$a(n) > 0 \forall n, \ \sum_n a(n) = \infty, \ a(n) \to 0 \text{ as } n \to \infty.$$

**Assumption E.3.** The sequence $\beta_n, n \geq 0$ is a bounded random sequence with $\beta_n \to 0$ almost surely as $n \to \infty$.

Let $t(n), n \geq 0$ be a sequence of positive real numbers defined according to $t(0) = 0$ and for $n \geq 1$, $t(n) = \sum_{j=0}^{n-1} a(j)$. By Assumption E.2, $t(n) \to \infty$ as $n \to \infty$. Let $m(t) = \max\{n \mid t(n) \leq t\}$. Thus, $m(t) \to \infty$ as $t \to \infty$.

**Assumption E.4.** There exists $T > 0$ such that $\forall \varepsilon > 0$,

$$\lim_{n \to \infty} P\left( \sup_{j \geq n} \max_{t \leq T} \left| \sum_{i=m(jT)}^{m(jT+t)-1} a(i)\xi_i \right| \geq \varepsilon \right) = 0.$$

**Assumption E.5.** The ODE (E.2) has a compact subset $K$ of $\mathscr{R}^N$ as its set of asymptotically stable equilibrium points.

[3, Theorem 5.3.1 (pp. 191-196)] essentially says the following:

**Theorem E.1 (Kushner and Clark Theorem).** *Under Assumptions E.1–E.5, almost surely, $X_n \to K$ as $n \to \infty$.*

**Remark E.1.** We comment here on the validity of Assumptions E.1–E.5. Note that Assumptions E.1, E.2 and E.5 are essentially standard requirements. In particular, the ODE (E.2) turns out to be well-posed as a consequence of Assumption E.1. The requirement on the sequence of step-sizes summing to infinity in Assumption E.2 ensures that the algorithm does not converge prematurely since $t(n) \to \infty$ as $n \to \infty$, even though the difference between successive time points (in the algorithm's trajectory) $t(n) - t(n-1) \to 0$. Assumption E.5 holds because $C$ is a compact set and $K$ being a closed subset of $C$ is also compact.

In the type of algorithms that we consider in this book, $\xi_n$ will typically correspond to the martingale difference term $M_{n+1}$. In such a case, the process $N_n, n \geq 0$ defined according to $N_0 = 0$ and $N_n = \sum_{m=0}^{n-1} \xi_m, n \geq 1$ will correspond to a martingale with respect to an appropriate filtration. If this martingale is convergent (that can perhaps be shown using say a martingale convergence theorem based argument), then Assumption E.4 can be seen to easily hold as well.

Finally, the sequence $\beta_n, n \geq 0$ in (E.1) will correspond in many cases to a measurement error term. For instance, if say $h(X_n) = -\nabla J(X_n)$, where $X_n$ is the $n$th parameter update and $\nabla J(X_n)$ is being estimated, i.e., is not known precisely, then $\beta_n$ could correspond to the error in the gradient estimate. As an example, consider the SPSA algorithm (with projection), see Chapter 5).

$$X_{n+1} = \Gamma\left(X_n + a(n)\left(\frac{J(X_n - \delta(n)\Delta(n)) - J(X_n + \delta(n)\Delta(n))}{2\delta(n)}(\Delta(n))^{-1}\right)\right), \tag{E.4}$$

where $\Delta(n) = (\Delta_1(n), \ldots, \Delta_N(n))^T$ with $\Delta_j(n), n \geq 0, j = 1, \ldots, N$ being independent random variables with (say) $\Delta_j(n) = \pm 1$ w. p. 1/2. Also, $(\Delta(n))^{-1} = (1/\Delta_1(n), \ldots, 1/\Delta_N(n))$. Now (E.4) can be rewritten in the form (E.1) with $h(X_n) = -\nabla J(X_n)$. Also,

$$\xi_n = \frac{J(X_n - \delta(n)\Delta(n)) - J(X_n + \delta(n)\Delta(n))}{2\delta(n)}(\Delta(n))^{-1}$$

$$-E\left[\frac{J(X_n - \delta(n)\Delta(n)) - J(X_n + \delta(n)\Delta(n))}{2\delta(n)}(\Delta(n))^{-1} \mid \mathscr{F}_n\right],$$

and

$$\beta_n = E\left[\frac{J(X_n + \delta(n)\Delta(n)) - J(X_n - \delta(n)\Delta(n))}{2\delta(n)}(\Delta(n))^{-1} \mid \mathscr{F}_n\right] - \nabla J(X_n),$$

respectively, where $\mathscr{F}_n = \sigma(X_m, m \leq n; \Delta(m), m < n), n \geq 1$. Assuming that $\delta(n) \to 0$, it can be seen that $\beta_n \to 0$ as $n \to \infty$. Further, if one assumes in addition to Assumption E.2 that $\sum_n \left(\frac{a(n)}{\delta(n)}\right)^2 < \infty$, then the martingale sequence $\sum_{m=0}^{n-1} a(m)\xi_m, n \geq 1$ can be seen to be convergent. Assumption E.4 is seen to hold in such a case.

**Remark E.2.** Note that stability of the iterates (E.1) is guaranteed by the fact that the operator $\Gamma$ projects each iterate of (E.1) to the set $C$ that is a compact subset of $\mathscr{R}^N$. The result as stated in Theorem 5.3.1 of [3] is in fact more general than that described in Theorem E.1. The latter however suffices for our purposes. In applications where it is usually difficult to prove that the iterates of the stochastic recursion are stable, projection is a commonly used technique to enforce stability of the iterates. By choosing the constraint region $C$ to be large enough, one can also ensure in many cases, that $C$ contains all the asymptotically stable attractors of the unprojected ODE $\dot{X}(t) = h(X(t))$. In such a case, it might actually be useful to apply a projection based scheme since then the algorithm would not spend its resources in searching the portion of the parameter space that is known not to contain the stable attractors. In the case when there are no stable fixed points of the unprojected ODE that lie inside the constraint set $C$ (i.e., in $C^o$), the algorithm will converge to a boundary point of $C$ that is the closest to an asymptotically stable attractor of the unprojected ODE. There could also be spurious fixed points that get introduced because of the projection operation. All such points however lie on the boundary of the constraint region $C$ (see for instance pp. 79 of [4]).

**Remark E.3.** As described in Assumption E.5, the set $K \subset \mathscr{R}^N$ corresponds to the set of asymptotically stable equilibria of the ODE (E.2). The set of fixed points, say $\hat{K}$, of the ODE (E.2) would contain $K$ in addition to other fixed points that would however be unstable. A stochastic approximation procedure would typically converge to the set $\hat{K}$. It has however been shown, for instance, in [2], [5] and [1] (Chapter 4) that with a sufficiently rich noise sequence, the stochastic update in fact converges to the stable attractor set and avoids the unstable portion of $\hat{K}$ altogether. Further, in practice, it is ususally the case that the stochastic algorithm converges to the stable set (and not the unstable portion) even when no extra conditions are imposed on the noise sequence.

## References

1. Borkar, V.S.: Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press and Hindustan Book Agency (Jointly Published), Cambridge and New Delhi (2008)
2. Brandiere, O.: Some pathological traps for stochastic approximation. SIAM J. Contr. and Optim. 36, 1293–1314 (1998)
3. Kushner, H.J., Clark, D.S.: Stochastic Approximation Methods for Constrained and Unconstrained Systems. Springer, New York (1978)
4. Kushner, H.J., Yin, G.G.: Stochastic Approximation Algorithms and Applications. Springer, New York (1997)
5. Pemantle, R.: Nonconvergence to unstable points in urn models and stochastic approximations. Annals of Prob. 18, 698–712 (1990)

# Index