Appendix B

Stochastic approximation

This appendix summarizes a number of fundamental results on stochastic approximation that play a key role in understanding the convergence properties of several algorithms discussed in Chapter 8. The main results are provided without proof, and the reader is referred to a proper reference on stochastic approximation for a full technical treatment of this subject.

The classical reference is, perhaps, the book of Kushner and Yin (2003). The books of Benveniste, Métevier, and Priouret (1990) and Duflo (1997) are alternative references that discuss both the theory and practice of stochastic approximation, while Borkar (2008) provides a more succinct and up-to-date account that includes recent developments, such as two-time-scale stochastic approximation algorithms.

B.1 Stochastic approximation algorithms

Stochastic approximation algorithms were originally proposed to address the problem of computing the zero w^* of some real-valued function F, when the function is not known but noise-corrupted samples thereof are available at any desired point $w \in \mathbb{R}$, and several estimation and optimization problems of interest can be framed as stochastic approximation.

Example B.1 Consider the problem of identifying the mean w^* of an unknown distribution with bounded second moment. Since

$$w^* = \operatorname*{argmin}_w \mathbb{E}\left[(\mathrm{x} - w)^2\right]$$

we can describe w^* as the solution of

$$\nabla_w \mathbb{E}\left[(\mathrm{x} - w^*)^2\right] = 0.$$

Then, letting

$$F(w) = -\frac{1}{2}\nabla_w \mathbb{E}\left[(\mathrm{x} - w)^2\right] = \mathbb{E}\left[\mathrm{x}\right] - w,$$

finding the mean of the unknown distribution is equivalent to finding the zero of F.

Given a mapping $F : \mathbb{R}^P \to \mathbb{R}^P$, let $\boldsymbol{w}^* \in \mathbb{R}^P$ be such that $F(\boldsymbol{w}^*) = 0$. A stochastic approximation algorithm to compute $\boldsymbol{w}^*$ takes the general form

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \boldsymbol{\alpha}_t(F_t(\boldsymbol{w}_t) + \mathbf{m}_t + \boldsymbol{\varepsilon}_t), \tag{B.1}$$

where

- $\boldsymbol{w}_t$ is a vector in $\mathbb{R}^P$ representing the current estimate for $\boldsymbol{w}^*$.
- The mapping $F_t : \mathbb{R}^P \to \mathbb{R}^P$ is an approximate version of F.
- $\boldsymbol{\alpha}_t$ is a step-size, with $\alpha_{t,p} \geq 0$ for $p = 1, \ldots, P$.
- $\mathbf{m}_t$ is a zero-mean noise term.
- ε_t is an asymptotically negligible disturbance.

In the context of this book, we are mostly interested in using stochastic approximation for fixed point computation. In other words, we want to determine $\boldsymbol{w}^*$ such that

$$H(\boldsymbol{w}^*) = \boldsymbol{w}^*,$$

for some unknown mapping $H : \mathbb{R}^P \to \mathbb{R}^P$. Letting

$$F(\boldsymbol{w}) = H(\boldsymbol{w}^*) - \boldsymbol{w}^*,$$

it becomes apparent that finding the fixed point of H can be framed as finding the zero of F. The update (B.1) can then be written equivalently as

$$\boldsymbol{w}_{t+1} = (1 - \boldsymbol{\alpha}_t)\boldsymbol{w}_t + \boldsymbol{\alpha}_t(H_t(\boldsymbol{w}_t) + \mathbf{m}_t + \boldsymbol{\varepsilon}_t), \tag{B.2}$$

where now the mapping $H_t : \mathbb{R}^P \to \mathbb{R}^P$ is an approximate version of H.

Example B.2 Let us return to the example of computing the mean of an unknown distribution. Given a set of points $\{x_1, \ldots, x_{t+1}\}$ sampled i.i.d. from the distribution of interest, it is a well-known result from probability theory that the desired mean can be estimated as

$$\begin{aligned} w_{t+1} &= \frac{1}{t+1}\sum_{\tau=1}^{t+1} x_\tau = \frac{1}{t+1}\sum_{\tau=1}^{t} x_\tau + \frac{1}{t+1}x_{t+1} \\ &= \frac{t}{t+1}w_t + \frac{1}{t+1}x_{t+1} \\ &= \left(1 - \frac{1}{t+1}\right)w_t + \frac{1}{t+1}x_{t+1}. \end{aligned} \tag{B.3}$$

Letting

- $H_t(w_t) = H(w_t) = \mathbb{E}\,[x_{t+1}]$;
- $\alpha_t = \frac{1}{t+1}$;
- $\mathrm{m}_t = \mathrm{x}_{t+1} - \mathbb{E}\,[x_{t+1}]$,

we can rewrite update (B.3) as

$$w_{t+1} = (1 - \alpha_t)w_t + \alpha_t(H(w_t) + \mathrm{m}_t),$$

it is clear that the computation of the mean from i.i.d. samples can be solved as a stochastic approximation problem. Moreover, by the law of large numbers, the sequence $\{w_t, t \in \mathbb{N}\}$ thus generated converges to w^* with probability 1.

As seen in the example, the convergence of a simple update such as that in (B.3) follows from the law of large numbers. Therefore, the following result can be seen as a generalization of the law of large numbers and can, in particular, be used to establish the convergence of (B.3) with probability 1 (w.p.1).

Lemma B.1 (Averaging Lemma). *Let w_t be a scalar updated according to the recursion*

$$w_{t+1} = (1 - \alpha_t)w_t + \alpha_t \mathrm{u}_t,$$

and assume that

1. *The step sizes α_t are nonnegative and satisfy*

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t^2 \leq \infty.$$

2. *The noise term u_t satisfies*

$$E[\mathrm{u}_t \mid \mathcal{F}_t] = 0 \qquad E[\mathrm{u}_t^2 \mid \mathcal{F}_t] \leq A + Br_t^2,$$

for some constants A and B.

Then w_t converges to 0 with probability 1.

In the formulation of the theorem, $\mathcal{F}_t$ corresponds to the σ-algebra generated by the history of the process up to time-step t, i.e., $\{\mathrm{u}_0, \ldots, \mathrm{u}_{t-1}, \alpha_0, \ldots, \alpha_{t-1}\}$. The following example illustrates the application of Lemma B.1 to the stochastic approximation algorithm in Example B.2.

Example B.3 Let us use Lemma B.1 to establish the convergence of w to w^*. Let $\delta_t \stackrel{\text{def}}{=} w_t - w^*$. If $\delta_t \to 0$ w.p.1, it follows that $w_t \to w^*$ w.p.1. Subtracting w^* on both sides of (B.3), we get

$$\delta_{t+1} = (1 - \alpha_t)\delta_t + \alpha_t(x_{t+1} - w^*).$$

We now verify the conditions of Lemma B.1.

1. We have that
$$\sum_{t=0}^{\infty} \frac{1}{t+1} = \infty$$
and
$$\sum_{t=0}^{\infty} \frac{1}{(t+1)^2} = \frac{\pi^2}{6} - 1 < \infty.$$

2. We also have that
$$\mathbb{E}\left[\mathrm{u}_t \mid \mathcal{F}_t\right] = \mathbb{E}\left[\mathrm{x}_{t+1} \mid \mathcal{F}_t\right] - w^* = \mathbb{E}_{\mathrm{x}_{t+1}\sim p}\left[\mathrm{x}_{t+1}\right] - \mathbb{E}_{\mathrm{x}\sim p}\left[\mathrm{x}\right] = 0,$$
since x_{t+1} is independent of the history, by assumption. On the other hand,
$$\mathbb{E}\left[\mathrm{u}_t^2 \mid \mathcal{F}_t\right] = \mathrm{var}\left[\mathrm{x}_t\right].$$
Since, by assumption, the distribution of interest has bounded second moment, there is $A < \infty$ such that $\mathrm{var}\left[\mathrm{x}\right] < A$.

It follows that $\delta_t \to 0$ w.p.1 and, consequently, that $w_t \to w^*$.

In the next section, we introduce the first of two main results on the convergence of stochastic approximation algorithms

B.2 Convergence under pseudo-contraction properties

Our first result, that we provide without proof, establishes convergence w.p.1 of the update in (B.2) when the mappings H_t are *pseudo-contractions*.

Theorem B.2. *Let $\{\boldsymbol{w}_t, t \in \mathbb{N}\}$ be the sequence generated by the update* (B.2)*, and suppose that*

(Ass. A.1) *The stepsizes $\alpha_{t,p}$ are nonnegative and satisfy, with probability 1,*

$$\sum_{t=0}^{\infty} \alpha_{t,p} = \infty, \qquad \sum_{t=0}^{\infty} \alpha_{t,p}^2 \leq \infty,$$

for $p = 1, \ldots, P$.

(Ass. A.2) *The noise term $\mathbf{m}_t$ is such that, for every $p = 1, \ldots, P$ and $t \in \mathbb{N}$,*

$$\mathbb{E}\left[\mathrm{m}_{t,p} \mid \mathcal{F}_t\right] = 0.$$

Moreover, there are constants $A, B > 0$ such that

$$\mathbb{E}\left[\mathrm{m}_{t,p}^2 \mid \mathcal{F}_t\right] \leq A + B\left\|\boldsymbol{w}_t\right\|^2,$$

where $\|\cdot\|$ is a suitable norm in $\mathbb{R}^P$.

(Ass. A.3) *There exists a vector $\boldsymbol{w}^* \in \mathbb{R}^P$, a positive vector $\mu \in \mathbb{R}^P$, and a scalar $\gamma \in [0, 1)$ such that, for all $t \in \mathbb{N}$,*

$$\left\|H_t(\boldsymbol{w}_t) - \boldsymbol{w}^*\right\|_\mu \leq \gamma \left\|\boldsymbol{w}_t - \boldsymbol{w}^*\right\|_\mu,$$

where $\|\cdot\|_\mu$ denotes a weighted norm induced by μ.

(Ass. A.4) *There exists a nonnegative random sequence c_t that converges to zero with probability 1, and is such that, for all t,*

$$|\boldsymbol{\varepsilon}_{t,p}| \leq c_t(\|\boldsymbol{w}_t\|_\mu + 1).$$

Then, $\boldsymbol{w}_t$ converges to $\boldsymbol{w}^*$ with probability 1.

As before, $\mathcal{F}_t$ denotes the σ-algebra generated by the history of the process up to time step t.

Theorem B.2 prompts several important observations. First of all, it is immediate to see that the Averaging Lemma (Lemma B.1) is a direct corollary of Theorem B.2.

A second observation is that the step size sequence $\{\boldsymbol{\alpha}_t, t \in \mathbb{N}\}$ may be non-deterministic and takes values in $\mathbb{R}^P$, which means that different components of $\boldsymbol{w}_t$ may experience different updates (that is not the case if the step-sizes are scalar). We refer to such differing updates as *asynchronous updates*.

One final observation is concerned with the assumptions of Theorem B.2.

- Assumptions A.1 is a standard assumption, and ensures that the step-sizes are sufficiently large to allow arbitrary initial estimates $\boldsymbol{w}_0$ to reach the target $\boldsymbol{w}^*$ but sufficiently small to ensure convergence.

- Assumptions A.2 and A.4 are also relatively standard, and ensure that the noise and disturbance sequences are "well-behaved".

- Assumption A.3 is the key assumption of Theorem B.2. It requires that the mappings H_t are *pseudo-contractions* around the target $\boldsymbol{w}^*$ (the zero of H that we want to compute).

We can now use Theorem B.2 to establish several simpler results. The first result is a generalization of the averaging lemma.

Proposition B.3. *Consider the update*

$$\boldsymbol{w}_{t+1} = (1 - \boldsymbol{\alpha}_t)\boldsymbol{w}_t + \boldsymbol{\alpha}_t \mathbf{u}_t. \tag{B.4}$$

Assume that the following hold:

1. *The step sizes $\alpha_{t,p}$ are nonnegative and satisfy, with probability 1,*

$$\sum_{t=0}^{\infty} \alpha_{t,p} = \infty, \qquad \sum_{t=0}^{\infty} \alpha_{t,p}^2 \leq \infty.$$

2. *There is a positive vector $\mu \in \mathbb{R}^P$ such that*

$$\left\| \mathbb{E}\left[\mathbf{u}_t \mid \mathcal{F}_t\right] \right\|_\mu \leq \gamma \left\| \boldsymbol{w}_t \right\|_\mu + c_t,$$

   where $\gamma \in [0, 1)$ and c_t is a sequence converging to zero w.p.1.

3. *There are constants $A, B > 0$ such that*

$$\operatorname{var}\left[\mathrm{u}_{t,p} \mid \mathcal{F}_t\right] \leq A + B \left\| \boldsymbol{w}_t \right\|_\mu^2.$$

Then, $\boldsymbol{w}_t$ converges to zero w.p.1.

Proof. Let

$$\tilde{\mathbf{u}}_t = \begin{cases} \mathbf{u}_t & \text{if } \left\| \mathbb{E}\left[\mathbf{u}_t \mid \mathcal{F}_t\right] \right\|_\mu \leq \gamma \left\| \boldsymbol{w}_t \right\|_\mu \\ \gamma \left\| \boldsymbol{w}_t \right\|_\mu & \text{otherwise,} \end{cases}$$

By construction, $\left\| \mathbb{E}\left[\tilde{\mathbf{u}}_t \mid \mathcal{F}_t\right] \right\|_\mu \leq \gamma \left\| \boldsymbol{w}_t \right\|_\mu$. We can now write the update (B.2) as

$$\boldsymbol{w}_{t+1} = (1 - \boldsymbol{\alpha}_t)\boldsymbol{w}_t + \boldsymbol{\alpha}_t \left(\mathbb{E}\left[\tilde{\mathbf{u}}_t \mid \mathcal{F}_t\right] + \mathbf{u}_t - \mathbb{E}\left[\mathbf{u}_t \mid \mathcal{F}_t\right] + \mathbb{E}\left[\mathbf{u}_t - \tilde{\mathbf{u}}_t \mid \mathcal{F}_t\right] \right),$$

and set

$$H_t(\boldsymbol{w}_t) = \mathbb{E}\left[\tilde{\mathbf{u}}_t \mid \mathcal{F}_t\right], \qquad \mathbf{m}_t = \mathbf{u}_t - \mathbb{E}\left[\mathbf{u}_t \mid \mathcal{F}_t\right], \qquad \boldsymbol{\varepsilon}_t = \mathbb{E}\left[\mathbf{u}_t - \tilde{\mathbf{u}}_t \mid \mathcal{F}_t\right].$$

Then,

1. Assumption A.1 of Theorem B.2 follows from 1.

2. We have that

$$\mathbb{E}\left[\mathbf{m}_t \mid \mathcal{F}_t\right] = \mathbb{E}\left[\mathbf{m}_t - \mathbf{m}_t \mid \mathcal{F}_t\right] = 0.$$

   Moreover, from 3,

$$\mathbb{E}\left[\mathrm{m}_{t,p}^2 \mid \mathcal{F}_t\right] = \operatorname{var}\left[\mathrm{u}_{t,p} \mid \mathcal{F}_t\right] \leq A + B \left\| \boldsymbol{w}_t \right\|_\mu^2.$$

3. Letting $\boldsymbol{w}^* = \mathbf{0}$,
$$\|H_t(\boldsymbol{w}_t)\|_\mu = \|\mathbb{E}\left[\tilde{\mathbf{u}}_t \mid \mathcal{F}_t\right]\|_\mu \leq \gamma \|\boldsymbol{w}_t\|_\mu ,$$
for $\gamma \in [0, 1)$.
4. Finally,
$$|\varepsilon_{t,p}| = |\mathbb{E}\left[\mathrm{m}_{t,p} - \tilde{\mathrm{m}}_{t,p} \mid \mathcal{F}_t\right]| \leq c_t,$$
by construction.

The conclusion follows. □

The next result considers stochastic approximation from a functional perspective. We start with some preliminary definitions. Given an arbitrary mapping $H : \mathbb{R}^P \to \mathbb{R}^P$, let $\mathcal{H} = \{\mathrm{H}_t, t \in \mathbb{N}\}$ denote a sequence of (possibly stochastic) mappings $\mathrm{H}_t : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}^P$. We say that $\mathcal{H}$ *approximates* H *at* $\boldsymbol{w}^* \in \mathbb{R}^P$ for initial values from $U \subset \mathbb{R}^P$ if the sequence
$$\boldsymbol{w}_{t+1} = \mathrm{H}_t(\boldsymbol{w}_t, \boldsymbol{w}^*) \tag{B.5}$$
converges to $H(\boldsymbol{w}^*)$ with probability 1.

Example B.4 Let us consider, once again, the computation of the mean from an unknown distribution. As seen in Example B.2, it is possible to use the update
$$w_{t+1} = (1 - \alpha_t)w_t + +\alpha_t(H(w_t) + \mathrm{m}_t),$$
where m_t is a zero-mean random variable. Define
$$H_t(w, w') = (1 - \alpha_t)w + \alpha(H(w') + \mathrm{m}_t).$$
As seen in Example B.2, it follows from the law of large number that $\mathcal{H}$ thus defined approximates H at w^* for initial values from $\mathbb{R}^P$.

Let us then consider a mapping $H : \mathbb{R}^P \to \mathbb{R}^P$ with fixed point $\boldsymbol{w}^*$. Moreover, let $\mathcal{H} = \{\mathrm{H}_t, t \in \mathbb{N}\}$ approximate H at $\boldsymbol{w}^*$ for initial values in some set $U \subset \mathbb{R}^P$, where U is *invariant under* $\mathcal{H}$, i.e., for all $\boldsymbol{w}, \boldsymbol{w}' \in U$, $\mathrm{H}_t(\boldsymbol{w}, \boldsymbol{w}') \in U$. Given $\boldsymbol{w}_0 \in U$, define the sequence
$$\boldsymbol{w}_{t+1} = \mathrm{H}_t(\boldsymbol{w}_t, \boldsymbol{w}_t).$$
We have the following result.

Proposition B.4. *Given a mapping $H : \mathbb{R}^P \to \mathbb{R}^P$ with fixed point $\boldsymbol{w}^*$, let $\mathcal{H} = \{\mathrm{H}_t, t \in \mathbb{N}\}$ approximate H at $\boldsymbol{w}^*$ for initial values in some set $U \subset \mathbb{R}^P$ and assume that U is* invariant under *$\mathcal{H}$, i.e., for all $\boldsymbol{w}, \boldsymbol{w}' \in U$, $\mathrm{H}_t(\boldsymbol{w}, \boldsymbol{w}') \in U$. Given $\boldsymbol{w}_0 \in U$, define the sequence*
$$\boldsymbol{w}_{t+1} = \mathrm{H}_t(\boldsymbol{w}_t, \boldsymbol{w}_t).$$

If there exist random sequences $\{\mathbf{u}_t, t \in \mathbb{N}\}$ *and* $\{\mathbf{v}_t, t \in \mathbb{N}\}$*, with* $\mathbf{u}_t, \mathbf{v}_t \in \mathbb{R}^P$*, such that*

1. *For every* $t \in \mathbb{N}$*,* $0 \leq \mathrm{u}_{t,p} \leq 1$ *and* $0 \leq \mathrm{v}_{t,p} \leq 1$*,* $p = 1, \ldots, P$*.*
2. *For every* $\boldsymbol{w}_1, \boldsymbol{w}_2 \in U$ *and* $t \in \mathbb{N}$*,*

$$|\mathrm{H}_{t,p}(\boldsymbol{w}_1, \boldsymbol{w}^*) - \mathrm{H}_{t,p}(\boldsymbol{w}_2, \boldsymbol{w}^*)_p| \leq \mathrm{v}_{t,p} \, |w_{1,p} - w_{2,p}| \,, \qquad p = 1, \ldots, P,$$

   where we write $\mathrm{H}_{t,p}(\boldsymbol{w}, \boldsymbol{w}')$ *to denote the* p*th component of the vector* $\mathrm{H}_t(\boldsymbol{w}, \boldsymbol{w}')$*.*

3. *For every* $\boldsymbol{w}_1, \boldsymbol{w}_2 \in U$ *and* $t \in \mathbb{N}$*,*

$$|\mathrm{H}_{t,p}(\boldsymbol{w}_1, \boldsymbol{w}^*) - \mathrm{H}_{t,p}(\boldsymbol{w}_1, \boldsymbol{w}_2)| \leq \mathrm{u}_{t,p}(\|\boldsymbol{w}^* - \boldsymbol{w}_2\| + \lambda_t), \quad p = 1, \ldots, P,$$

   where $\{\lambda_t, t \in \mathbb{N}\}$ *is a sequence such that* $\lambda_t \to 0$ *w.p.1.*

4. *For* $p = 1, \ldots, P$*,*

$$\sum_{t=0}^{\infty} (1 - \mathrm{v}_{t,p}) = \infty \qquad\qquad \sum_{t=0}^{\infty} (1 - \mathrm{v}_{t,p})^2 < \infty.$$

5. *There is* $0 \leq \gamma < 1$ *and large enough* t

$$\mathrm{u}_{t,p} \leq \gamma(1 - \mathrm{v}_{t,p}), \qquad p = 1, \ldots, P.$$

then $\boldsymbol{w}^{(k)}$ *converges to* $\boldsymbol{w}^*$ *w.p.1.*

Proof. Consider the sequence

$$\boldsymbol{\omega}_0 = \boldsymbol{w}_0, \qquad\qquad \boldsymbol{\omega}_{t+1} = H_t(\boldsymbol{\omega}_t, \boldsymbol{w}^*).$$

Since $\mathcal{H}$ approximates H at $\boldsymbol{w}^*$, $\boldsymbol{\omega}_t \to \boldsymbol{w}^*$ w.p.1. Moreover, defining $\Delta_{t,p} = |\omega_{t,p} - w_p^*|$, we have that $\Delta_t \to 0$ w.p.1.

Define $\delta_{t,p} = |\omega_{t,p} - w_{t,p}|$, where $\boldsymbol{w}_t$ is the sequence in the statement of the theorem. If $\delta_t \to 0$ w.p.1, then $\boldsymbol{w}_t \to \boldsymbol{w}^*$ w.p.1. We have that

$$\begin{aligned}
\delta_{t+1,p} &= |\omega_{t+1,p} - w_{t+1,p}| \\
&= |H_{t,p}(\boldsymbol{\omega}_t, \boldsymbol{w}^*) - H_{t,p}(\boldsymbol{w}_t, \boldsymbol{w}_t)| \\
&\leq |H_{t,p}(\boldsymbol{\omega}_t, \boldsymbol{w}^*) - H_{t,p}(\boldsymbol{w}_t, \boldsymbol{w}^*)| + |H_{t,p}(\boldsymbol{w}_t, \boldsymbol{w}^*) - H_{t,p}(\boldsymbol{w}_t, \boldsymbol{w}_t)| \\
&\leq \mathrm{v}_{t,p}|\omega_{t,p} - w_{t,p}| + \mathrm{u}_{t,p}(\|\boldsymbol{w}_t - \boldsymbol{w}^*\| + \lambda_t) \\
&\leq \mathrm{v}_{t,p}\delta_{t,p} + \mathrm{u}_{t,p}(\|\boldsymbol{w}_t - \boldsymbol{\omega}_t\| + \|\boldsymbol{\omega}_t - \boldsymbol{w}^*\| + \lambda_t) \\
&= \mathrm{v}_{t,p}\delta_{t,p} + \mathrm{u}_{t,p}(\|\delta_t\| + \|\Delta_t\| + \lambda_t) \\
&\leq \mathrm{v}_{t,p}\delta_{t,p} + (1 - \mathrm{v}_{t,p})\gamma(\|\delta_t\| + \|\Delta_t\| + \lambda_t)
\end{aligned}$$

Consider the update

$$z_{t+1,p} = (1 - \alpha_{t,p})z_{t,p} + \alpha_{t,p}\gamma(\|\boldsymbol{z}_t\| + \|\Delta_t\| + \lambda_t),$$

where $\alpha_{t,p} = (1 - \mathrm{u}_{t,p})$. We can now apply Proposition B.3 to conclude that $\boldsymbol{z}_t$ converges to 0 w.p.1.

We now have that $z_{t,p} \geq 0$ and $\delta_{t,p} \geq 0$ for all t and all x. Setting $\boldsymbol{z}_0 = \delta_0$, it follows that $z_{t,p} \geq \delta_{t,p}$ for all t. To see why this is so, suppose that $z_{t,p} \geq \delta_{t,p}$ for some t. Then $\|\boldsymbol{z}_t\| \geq \|\delta_t\|$ and, therefore,

$$\begin{aligned} z_{t+1,p} &= \mathrm{v}_{t,p} z_{t,p} + (1 - \mathrm{v}_{t,p})\gamma(\|\boldsymbol{z}_t\| + \|\Delta_t\| + \lambda_t) \\ &\geq \mathrm{v}_{t,p}\delta_{t,p} + (1 - \mathrm{v}_{t,p})\gamma(\|\delta_t\| + \|\Delta_t\| + \lambda_t) \\ &\geq \delta_{t+1,p}. \end{aligned}$$

It follows by induction that $z_{t,p} \geq \delta_{t,p}$ for all t. But then, since $0 \leq \delta_{t,p} \leq z_{t,p}$ and $z_{t,p} \to 0$ w.p.1, it follows that $\delta_{t,p} \to 0$ w.p.1 and the conclusion follows. □

B.3 Dynamical systems approach to convergence

Theorem B.2 establishes the convergence of stochastic approximation algorithms where the mapping H_t in (B.2) is a so-called *pseudo-contraction*. In other words, we require that $H_t(\boldsymbol{w})$ is closer to $\boldsymbol{w}^*$ than $\boldsymbol{w}$. This is a strong assumption, ant it is not surprising that the algorithm converges to $\boldsymbol{w}^*$, provided that the noise and disturbances are "well-behaved", in the sense of Theorem B.2.

In this section we provide without proof an alternative result that alleviates the pseudo-contraction properties required of H_t. This generalization, however, comes at a cost. First, the convergence result assumes scalar step sizes and, as such, is only applicable to algorithms with *synchronous updates*. Second, boundedness of the iterates of the algorithm must be established independently.

Consider the update rule

$$\boldsymbol{w}_{t+1} = (1 - \alpha_t)\boldsymbol{w}_t + \alpha_t(H(\boldsymbol{w}_t) + \mathbf{m}_t + \boldsymbol{\varepsilon}_t), \tag{B.6}$$

where the step size α_t is now a positive scalar, and H is time-independent.

Theorem B.5. *Let $\{\boldsymbol{w}_t, t \in \mathbb{N}\}$ be the sequence generated by the update* (B.6)*, and suppose that*

(Ass. B.1) *The stepsizes α_t are nonnegative and satisfy, with probability 1,*

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t^2 \leq \infty.$$

(Ass. B.2) *The noise term $\mathbf{m}_t$ is such that, for every $p = 1, \ldots, P$ and $t \in \mathbb{N}$,*

$$\mathbb{E}\left[\mathrm{m}_{t,p} \mid \mathcal{F}_t\right] = 0.$$

Moreover, there are constants $A, B > 0$ such that

$$\mathbb{E}\left[\mathrm{m}_{t,p}^2 \mid \mathcal{F}_t\right] \leq A + B\left\|\boldsymbol{w}_t\right\|^2,$$

where $\|\cdot\|$ is a suitable norm in $\mathbb{R}^P$.

(Ass. B.3) *The mapping $H : \mathbb{R}^P \to \mathbb{R}^P$ is* Lipschitz continuous*, i.e., there is a scalar $C > 0$ such that*

$$\|H(\boldsymbol{w}) - H(\boldsymbol{w}')\| \leq C\left\|\boldsymbol{w} - \boldsymbol{w}'\right\|,$$

for all $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^P$.

(Ass. B.4) *There exists a nonnegative random sequence c_t that converges to zero with probability 1, and is such that, for all t,*

$$|\boldsymbol{\varepsilon}_{t,p}| \leq c_t(\left\|\boldsymbol{w}_t\right\|_\mu + 1).$$

(Ass. B.5) *The iterations of the algorithm remain bounded w.p.1., i.e.,*

$$\sup_{t\in\mathbb{N}} \|\boldsymbol{w}_t\| < \infty$$

with probability 1.

(Ass. B.6) *There exists a continuously differentiable function $V : \mathbb{R}^P \to \mathbb{R}$ such that*

- $V(\boldsymbol{w}) \geq 0$*;*
- $V(\boldsymbol{w}) \to \infty$ *as* $\|\boldsymbol{w}\| \to \infty$*—i.e., V is* radially unbounded*;*
- $\nabla_{\boldsymbol{w}} V(\boldsymbol{w})^\top (H(\boldsymbol{w}) - \boldsymbol{w}) \leq 0$.

Then, $V(\boldsymbol{w}_t)$ converges to 0 with probability 1.

Let us consider Theorem B.5 in detail. Assumptions B.1, B.2 and B.4 are essentially similar to those encountered in Theorem B.2. Assumption B.3 alleviates the pseudo-contraction requirement of Theorem B.2, requiring H to be only Lipschitz continuous.

Theorem B.2 includes two additional assumptions. Assumption B.5 has already been discussed, and requires the sequence $\{\boldsymbol{w}_t, t \in \mathbb{N}\}$ generated by the algorithm to remain bounded. Such requirement is often non-trivial to establish, and in the continuation we provide a result that is often useful to establish Assumption B.5.

From Assumptions B.1-B.5 it is possible to show that the sequence $\{\boldsymbol{w}_t, t \in \mathbb{N}\}$ generated by the update (B.6) are "close" to those of the ordinary differential equation (o.d.e.)

$$\dot{\boldsymbol{w}}_t = H(\boldsymbol{w}_t) - \boldsymbol{w}_t. \tag{B.7}$$

Therefore, convergence of the algorithm can be established by showing that (B.7)

is asymptotically stable. Assumption B.6 ensures precisely that: the function V in the text of the theorem is a *Lyapunov function* for the o.d.e. (B.7), and its existence ensures that the trajectories of the o.d.e. eventually settle in an equilibrium point Khalil, 2001.

The next result, which we provide without proof, identifies conditions under which Assumption B.5 can be asserted.

Theorem B.6. *Let $\{\boldsymbol{w}_t, t \in \mathbb{N}\}$ be the sequence generated by the update* (B.6)*, and suppose that Assumptions B.1 through B.3 hold. Consider the mapping $H_\rho : \mathbb{R}^P \to \mathbb{R}^P$ defined as*

$$H_\rho(\boldsymbol{w}) = \frac{1}{\rho} H(\rho \boldsymbol{w}),$$

for some positive scalar ρ, and let $H_\infty : \mathbb{R}^P \to \mathbb{R}^P$ be defined as

$$H_\infty(\boldsymbol{w}) = \lim_{\rho \to \infty} H_\rho(\boldsymbol{w}).$$

If the $\mathbf{0}$ is the unique, globally asymptotically stable equilibrium of the o.d.e.

$$\dot{\boldsymbol{w}}_t = H_\infty(\boldsymbol{w}_t) - \boldsymbol{w}_t,$$

then $\sup_t \|\boldsymbol{w}_t\| < \infty$ w.p.1.

From Theorem B.6 it follows that both Assumptions B.5 and B.6 can be established by analyzing the behavior of a suitable dynamical system.

We conclude this appendix by remarking that Assumption A.3 of Theorem B.2 implies both Assumptions B.5 and B.6 of Theorem B.5. We start with Assumption B.6.

Lemma B.7. *Suppose that Assumption A.3 of Theorem B.2 holds. Then, Assumption B.6 holds with the function*

$$V(\boldsymbol{w}) = \frac{1}{2} \|\boldsymbol{w} - \boldsymbol{w}^*\|_\mu^2 .$$

Proof. By construction, $V(\boldsymbol{w}) \geq 0$, with equality only if $\boldsymbol{w} = \boldsymbol{w}^*$. On the other hand, letting $\boldsymbol{M}$ denote the diagonal matrix with μ in its main diagonal,

$$V(\boldsymbol{w}) = \frac{1}{2} (\boldsymbol{w} - \boldsymbol{w}^*)^\top \boldsymbol{M} (\boldsymbol{w} - \boldsymbol{w}^*).$$

Therefore,

$$\begin{aligned}\nabla_{\boldsymbol{w}} V(\boldsymbol{w})^\top (H(\boldsymbol{w}) - \boldsymbol{w}) &= (\boldsymbol{w} - \boldsymbol{w}^*)^\top \boldsymbol{M}(H(\boldsymbol{w}) - \boldsymbol{w}) \\ &= (\boldsymbol{w} - \boldsymbol{w}^*)^\top \boldsymbol{M}(H(\boldsymbol{w}) - \boldsymbol{w}^* + \boldsymbol{w}^* - \boldsymbol{w}) \\ &= -\|\boldsymbol{w} - \boldsymbol{w}^*\|_\mu^2 + (\boldsymbol{w} - \boldsymbol{w}^*)^\top \boldsymbol{M}(H(\boldsymbol{w}) - \boldsymbol{w}^*).\end{aligned}$$

Using the Cauchy-Schwartz inequality (see Appendix A), we get

$$\begin{aligned}\nabla_{\boldsymbol{w}} V(\boldsymbol{w})^\top (H(\boldsymbol{w}) - \boldsymbol{w}) &\leq -\|\boldsymbol{w} - \boldsymbol{w}^*\|^2 + \|\boldsymbol{w} - \boldsymbol{w}^*\|_\mu \|H(\boldsymbol{w}) - \boldsymbol{w}^*\|_\mu \\ &\leq -\|\boldsymbol{w} - \boldsymbol{w}^*\|^2 + \gamma \|\boldsymbol{w} - \boldsymbol{w}^*\|_\mu^2 \\ &\leq \boldsymbol{0},\end{aligned}$$

and the proof is complete. □

The next lemma establishes that, under Assumption A.3, we can use Theorem B.6 to assert that the iterations of the algorithm remain bounded.

Lemma B.8. *Suppose that Assumption A.3 of Theorem B.2 holds. Then,*

$$\sup \|\boldsymbol{w}_t\| < \infty$$

with probability 1.

Proof. As stated above, we show that the iterations of the algorithm remain bounded by using Theorem B.6. To show that the origin is a globally asymptotically stable equilibrium of the o.d.e.

$$\dot{\boldsymbol{w}}_t = H_\infty(\boldsymbol{w}_t) - \boldsymbol{w}_t,$$

we let

$$V(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|_\mu^2$$

and show that it is a Lyapunov function for the o.d.e. In other words, V should verify

- $V(\boldsymbol{w}) \geq \boldsymbol{0}$, and $V(\boldsymbol{w}) = 0$ only if $\boldsymbol{w} = \boldsymbol{0}$;
- V is radially unbounded, i.e., $\lim_{\|\boldsymbol{w}\| \to \infty} V(\boldsymbol{w}) = \infty$:
- $\dot{V}(\boldsymbol{w}) \leq \boldsymbol{0}$, with $\dot{V}(\boldsymbol{w}) = 0$ only if $\boldsymbol{w} = \boldsymbol{0}$.

The two first properties hold by construction. As for the third,

$$\begin{aligned}\dot{V}(\boldsymbol{w}) &= \boldsymbol{w}^\top \boldsymbol{M}(H_\infty(\boldsymbol{w}) - \boldsymbol{w}) \\ &= -\|\boldsymbol{w}\|_\mu^2 + \boldsymbol{w}^\top \boldsymbol{M} H_\infty(\boldsymbol{w}).\end{aligned}$$

Replacing H_∞ by its definition, we have

$$\begin{aligned}\dot{V}(\boldsymbol{w}) &= -\left\|\boldsymbol{w}\right\|_\mu^2 + \lim_{\rho\to\infty}\frac{1}{\rho}\boldsymbol{w}^\top \boldsymbol{M} H(\rho\boldsymbol{w}) \\ &= -\left\|\boldsymbol{w}\right\|_\mu^2 + \lim_{\rho\to\infty}\frac{1}{\rho}\boldsymbol{w}^\top \boldsymbol{M}(H(\rho\boldsymbol{w}) - \boldsymbol{w}^*),\end{aligned}$$

since $\lim_{\rho\to\infty}\boldsymbol{w}^*/\rho = \mathbf{0}$. Using the Cauchy-Schwartz inequality, we get

$$\begin{aligned}\dot{V}(\boldsymbol{w}) &= -\left\|\boldsymbol{w}\right\|_\mu^2 + \lim_{\rho\to\infty}\frac{\gamma}{\rho}\left\|\boldsymbol{w}\right\|_\mu \left\|\rho\boldsymbol{w} - \boldsymbol{w}^*\right\|\mu \\ &= -\left\|\boldsymbol{w}\right\|_\mu^2 + \gamma\left\|\boldsymbol{w}\right\|_\mu^2 .\end{aligned}$$

The conclusion follows. □