Asynchronous stochastic-approximation theorem

Asynchronous stochastic-approximation theorem. Consider a finite table of errors $e_t: \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a finite set of table indices. A **table index** is one element of $\mathcal{X}$. It names one cell of the table. At time t, the function e_t assigns one real number to each cell; this number is the current error stored in that cell.

In the Q-learning application, the convergence proof in the extended version of Chapter 7 defines the error table in Equation 7E.24 before deriving the error recursion in Equation 7E.25. A table index is an observation-action pair. If $x = (o, a)$, Equation 7E.24 means, cell by cell,

$$e_t(x) = e_t(o, a) \stackrel{\text{def}}{=} Q_t(o, a) - Q^*(o, a).$$

For example, suppose there are two observations and two actions,

$$\mathcal{O} = \{\text{near,far}\}, \qquad \mathcal{A} = \{\text{wait,move}\}.$$

Then the finite set of table indices is

$$\mathcal{X} = \mathcal{O} \times \mathcal{A} = \{(\text{near,wait}), (\text{near,move}), (\text{far,wait}), (\text{far,move})\}.$$

An error table at time t could then be

index $x = (o, a)$	error $e_t(x)$
(near,wait)	0.20
(near,move)	-0.50
(far,wait)	-0.10
(far,move)	0.40

For this table,

$$e_t(\text{near,move}) = -0.50.$$

The max norm is the largest absolute value among the table entries. In this example,

$$\|e_t\|_\infty = \max\{0.20, 0.50, 0.10, 0.40\} = 0.50.$$

Thus $\|e_t\|_\infty$ measures the largest remaining cell error, not an average error.

At each time t, one entry $x_t \in \mathcal{X}$ is chosen for update. The chosen entry x_t is called the **visited entry** at time t. Every entry $y \in \mathcal{X}$ with $y \neq x_t$ is called **unvisited** at time t. Suppose that every entry is visited infinitely often.

At time t, the update has already selected the entry x_t and the learning rate α_t, but the new noise w_{t+1} has not yet been observed. Let H_t denote the **history** available at that moment. In the abstract theorem, H_t is a name for the information with respect to which the already chosen quantities in the update are fixed. Thus the current error table e_t, the visited entry x_t, the learning rate α_t, and the past noise terms $w_1, \ldots, w_t$ are determined once H_t is fixed. The new noise term w_{t+1} is not included in H_t.

Let $g_t(x_t)$ and w_{t+1} be the two scalar terms that appear in the update target for the visited entry. Suppose $0 \leq \alpha_t \leq 1$ and suppose the visited entry satisfies

$$e_{t+1}(x_t) = (1 - \alpha_t)e_t(x_t) + \alpha_t[g_t(x_t) + w_{t+1}]. \qquad (A.1)$$

In Equation A.1, the bracketed quantity $g_t(x_t) + w_{t+1}$ is the second value in a weighted sum. The new value $e_{t+1}(x_t)$ is formed from the old value $e_t(x_t)$ with coefficient $1 - \alpha_t$ and from the bracketed quantity with coefficient α_t. Since $0 \leq \alpha_t \leq 1$, these two coefficients are nonnegative and sum to 1.

The theorem assumes that the second value in this weighted sum is written as a sum of two scalar parts. The term $g_t(x_t)$ is called the **non-noise term**. It may depend on the history H_t, so its value is fixed once H_t is fixed. The term w_{t+1} is called the **noise term**. It is the new random part observed after H_t.

> In the Q-learning application, when $x_t = (o_t, a_t)$, Equation 7E.25 has the same form as Equation A.1. The term $g_t(x_t)$ is the Bellman-error term defined in Equation 7E.26, and its bound is shown in Equation 7E.27. The term w_{t+1} is the sampling-noise term defined in Equation 7E.22.

Every other entry is unchanged at that time:

$$e_{t+1}(y) = e_t(y), \qquad y \neq x_t. \tag{A.2}$$

Suppose also that, for each entry x, the learning rates along the visits to x satisfy the Robbins-Monro conditions

$$\sum_k \alpha_k(x) = \infty, \qquad \sum_k \alpha_k(x)^2 < \infty \tag{A.3}$$

where k counts visits to that entry. Suppose that the noise terms satisfy the **martingale-difference condition** with respect to the histories H_t:

$$\mathbb{E}[w_{t+1} \mid H_t] = 0 \tag{A.4}$$

for every t. The expectation in Equation A.4 is taken over the random outcome that produces the next noise term w_{t+1}, after the history H_t has been fixed. Since x_t and α_t are included in H_t, the expectation is not averaging over which entry is chosen or which learning rate is used at time t. It averages only over the remaining randomness in the next update. In the Q-learning application, this is the randomness in the next sampled transition and reward.

> In the Q-learning proof, the corresponding zero-conditional-mean statement is Equation 7E.23. There the conditioning is written as h_t, a_t. In the theorem, H_t is the abstract name for the same kind of pre-transition information. For the Q-learning application, take H_t to contain h_t, the selected action a_t, and only quantities that are already fixed once h_t, a_t, and the learning rule are fixed. These include Q_t, e_t, $x_t = (o_t, a_t)$, and $\alpha_t(o_t, a_t)$. They do not include the next reward, the next observation, the sampled target y_t, or the new noise w_{t+1}. Under this choice of H_t, Equation 7E.23 is the Q-learning instance of Equation A.4.

A sequence of noise terms satisfying Equation A.4 is called a **martingale difference sequence**.

Also suppose that the noise is bounded. The word **bounded** means that there is a finite constant C such that

$$|w_{t+1}| \leq C \tag{A.5}$$

for every t. Suppose, finally, that there is a constant $0 \leq \gamma < 1$ such that the non-noise term in Equation A.1 satisfies

$$|g_t(x_t)| \leq \gamma \|e_t\|_\infty \tag{A.6}$$

for every t.

> In the Q-learning application, with $x_t = (o_t, a_t)$ and $g_t(x_t)$ equal to the Bellman-error term from Equation 7E.26, this bound is exactly the bound proved in Equation 7E.27.

Then

$$\|e_t\|_\infty \to 0$$

with probability 1.

Proof.

Let

$$M_t \overset{\text{def}}{=} \|e_t\|_\infty.$$

The bound used below is the hypothesis displayed in Equation A.6. Since $M_t = \|e_t\|_\infty$, that hypothesis says

$$|g_t(x_t)| \le \gamma M_t.$$

First observe that the error process is bounded. This boundedness is a consequence of the theorem hypotheses; it is not a new assumption. We will construct one finite constant D and prove that

$$M_t = \|e_t\|_\infty \le D$$

for every integer time $t \ge 0$.

The bounded-noise hypothesis in Equation A.5 gives

$$|w_{t+1}| \le C$$

for all t. Define

$$D \stackrel{\text{def}}{=} \max\left\{M_0,\ \frac{C}{1-\gamma}\right\}.$$

We will prove by induction that

$$M_t \le D$$

for every time t. The definition of D is chosen to make both parts of that induction work.

By the definition of D,

$$M_0 \le D.$$

This establishes the base case of the induction. It says that the initial error table is already inside the bound D.

Second, again by the definition of D,

$$\frac{C}{1-\gamma} \le D.$$

Because $\gamma < 1$, the number $1-\gamma$ is positive. Multiplying the last inequality by $1-\gamma$ gives

$$C \le (1-\gamma)D. \qquad (A.7)$$

Equation A.7 will be used in the induction step to prove the implication

$$M_t \le D \quad \Rightarrow \quad M_{t+1} \le D.$$

Suppose the current table satisfies $M_t \le D$. Then every current error entry has magnitude at most D; in particular, $|e_t(x_t)| \le D$. The theorem hypothesis in Equation A.6 gives

$$|g_t(x_t)| \le \gamma M_t.$$

Combining this with the induction hypothesis $M_t \le D$ gives

$$|g_t(x_t)| \le \gamma M_t \le \gamma D. \qquad (A.8)$$

So the non-noise term has magnitude at most γD because of two facts used together: the theorem hypothesis in Equation A.6 and the current induction hypothesis $M_t \le D$.

The bracketed target in Equation A.1 is $g_t(x_t) + w_{t+1}$. By the triangle inequality, Equation A.8, the bounded-noise hypothesis in Equation A.5, and Equation A.7,

$$|g_t(x_t) + w_{t+1}| \le |g_t(x_t)| + |w_{t+1}| \le \gamma D + C \le \gamma D + (1-\gamma)D = D.$$

So, whenever $M_t \le D$, the bracketed target has magnitude at most D. Equation A.1 then forms $e_{t+1}(x_t)$ as a weighted average of two numbers that both lie in $[-D, D]$: the old entry $e_t(x_t)$, with weight $1-\alpha_t$, and the bracketed target, with weight α_t. Since the two weights are nonnegative and sum to one, the average also lies in $[-D, D]$—the same convex-combination argument that bounds $\|Q_t\|_\infty$ in the Q-learning convergence proof.

Therefore $|e_{t+1}(x_t)| \leq D$. All other entries are unchanged by Equation A.2, so $M_{t+1} \leq D$. Together with the base case $M_0 \leq D$, induction gives $M_t \leq D$ for every t.

Before stating the next lemma, we introduce one piece of notation. For a real sequence $b_0, b_1, b_2, ...$, the **limit superior** $\limsup_{k\to\infty} b_k$ is the value that describes the largest level reached by the sequence in the long run. The statement

$$\limsup_{k\to\infty} b_k \leq A$$

means that, for every $\varepsilon > 0$, there is an index K such that

$$b_k \leq A + \varepsilon$$

for all $k \geq K$. In words, after finitely many early terms, the sequence stays below $A + \varepsilon$. Below, the shorter notation $\limsup_k$ means $\limsup_{k\to\infty}$.

We now use the **Robbins-Monro averaging lemma**. Its full statement and step-by-step proof, together with a pathwise remark that we will need below, are given in the companion note of the same name. The lemma states that if a scalar sequence satisfies

$$z_{k+1} = (1 - \beta_k)z_k + \beta_k(c_k + \xi_{k+1}),$$

where $0 \leq \beta_k \leq 1$,

$$\sum_k \beta_k = \infty, \qquad \sum_k \beta_k^2 < \infty,$$

(ξ_{k+1}) is a bounded martingale difference sequence, and $\limsup_k |c_k| \leq A$, then

$$\limsup_k |z_k| \leq A$$

with probability 1.

Now fix one table entry x. Look only at the subsequence of times when x is updated. Since every entry is updated infinitely often, this subsequence is infinite; write $t_1 < t_2 < t_3 < \cdots$ for its elements. Along this subsequence, the update of $e_t(x)$ has exactly the scalar form in Equation A.1, with learning rates satisfying Equation A.3 for that entry.

The corresponding noise terms are still martingale differences with respect to the histories at those update times. The reason is that the update times are themselves determined by the histories: whether the k^{th} update of x happens at time t is decided by information contained in H_t. On that event, conditioning on the history at the k^{th} update time is the same as conditioning on H_t, and Equation A.4 then gives zero conditional mean for the next noise term.

To make the correspondence with the scalar lemma explicit, define, for $k = 1,2,3, ...$,

$$z_k = e_{t_k}(x), \qquad \beta_k = \alpha_{t_k}, \qquad c_k = g_{t_k}(x), \qquad \xi_{k+1} = w_{t_k+1}.$$

Because x is not updated between times t_k and t_{k+1}, Equation A.2 gives

$$e_{t_{k+1}}(x) = e_{t_k+1}(x).$$

Therefore Equation A.1 becomes

$$z_{k+1} = (1 - \beta_k)z_k + \beta_k(c_k + \xi_{k+1}),$$

which is exactly the scalar recursion in the Robbins-Monro averaging lemma.

Let

$$L \overset{\text{def}}{=} \limsup_{t\to\infty} M_t.$$

Because $M_t \leq D$ for all t, L is finite. Note that L is not a number known in advance: it depends on the realized sequence of errors, so different runs of the algorithm can produce different values of L.

For the fixed entry x, the non-noise term along its update subsequence satisfies the bound in Equation A.6:

$$|g_{t_k}(x)| \leq \gamma M_{t_k}.$$

A limit superior along a subsequence is at most the limit superior along the full sequence, so

$$\limsup_{k\to\infty} |g_{t_k}(x)| \leq \gamma \limsup_{k\to\infty} M_{t_k} \leq \gamma \limsup_{t\to\infty} M_t = \gamma L.$$

We now apply the Robbins-Monro averaging lemma to this entry. Because the bound γL depends on the realization, we use the lemma in the pathwise form stated in the remark at the end of its proof: on the probability-one event where the accumulated noise series for this entry converges, the realized input sequence has limit superior at most γL, and the lemma's conclusion holds for that same realization. This gives

$$\limsup_{k\to\infty} |e_{t_k}(x)| \leq \gamma L$$

with probability 1.

The limit superior along the update times is also the limit superior over all times. Between consecutive updates of x, the entry does not change, by Equation A.2; since x is updated infinitely often, every late value of $|e_t(x)|$ equals a late value of the update-time subsequence, held constant until the next update. Therefore

$$\limsup_{t\to\infty} |e_t(x)| \leq \gamma L$$

with probability 1.

This holds for every entry $x \in \mathcal{X}$. There are finitely many entries, so the finitely many probability-one events—one per entry—intersect in a single probability-one event on which the bound holds for all entries simultaneously:

$$\max_{x\in\mathcal{X}} \limsup_{t\to\infty} |e_t(x)| \leq \gamma L.$$

To turn this into a statement about L itself, we use the fact that, for finitely many sequences, the limit superior of their maximum equals the maximum of their limit superiors. One direction is immediate: at every time, the maximum over entries is at least each individual $|e_t(x)|$, so its limit superior is at least each individual limit superior, and hence at least their maximum. For the other direction, pick a subsequence of times along which $\max_x |e_t(x)|$ converges to its limit superior. At each of those times, the maximum is attained by some entry; since $\mathcal{X}$ is finite, at least one entry $\bar{x}$ attains it infinitely often along the subsequence, and along those times $|e_t(\bar{x})|$ converges to $\limsup_t \max_x |e_t(x)|$. Hence $\limsup_t |e_t(\bar{x})|$ is at least that value, and so is the maximum of the per-entry limit superiors. This interchange is where the finiteness of the table enters this part of the argument; with infinitely many entries the equality can fail. Therefore

$$L = \limsup_{t\to\infty} \|e_t\|_\infty = \max_{x\in\mathcal{X}} \limsup_{t\to\infty} |e_t(x)| \leq \gamma L.$$

But $L \geq 0$ and $\gamma < 1$, so the only possibility is

$$L = 0.$$

Therefore

$$\limsup_{t\to\infty} \|e_t\|_\infty = 0.$$

Since $\|e_t\|_\infty$ is nonnegative, this implies

$$\|e_t\|_\infty \to 0$$

with probability 1.

This proves the theorem.