Note on Stochastic Approximation

Extending Tsitsiklis 1994

Longye Tian
longye.tian@anu.edu.au

Australian National University
School of Economics

May 16th, 2025

Big Picture

- We want to estimate some unknown function x^*
- We have an estimate $x(t)$ at time t
- At each time t, we have noisy observation $F(x(t)) + w(t)$
  - $F(x(t))$: we can think this as the fixed point equation

$$F(x^*) = x^*$$

  - $w(t)$: a random noise comes with the observation
- Stochastic approximation algorithm ($x(t) \to x^*$ as $t \to \infty$)

$$x(t+1) = (1 - \alpha(t))x(t) + \alpha(t)(F(x(t)) + w(t))$$

or

$$x(t+1) = x(t) + \alpha(t)\left[F(x(t)) + w(t) - x(t)\right]$$

Motivating example: Q-learning

- We want to estimate the unknown $Q^*(s,a)$
  - $Q^*(s,a)$ is the maximal expected lifetime rewards given state s and action a.
  - with **known** reward function $r(s,a)$ and transition probability, we can use DP method to compute Q^*

$$Q^*(s,a) = r(s,a) + \beta \sum_{s'} \max_{a'} Q^*(s',a') P(s'|s,a)$$

  - Sometime, we only observe
    - one realization of the random variable for reward R
    - one next state s' not all possible next states
    - use current estimate of Q not Q^*
- at each time t, we observe

$$R(t) + \beta \max_{a'} Q(s',a')$$

where $F(Q(s,a)) = \mathbb{E}(R + \beta \max_{a'} Q(s',a')|s,a)$

Motivating example: Q-learning

The stochastic approximation algorithm in Q-learning

$$Q(s,a) \leftarrow Q(s,a) + \alpha(t)\left[R + \beta \max_{a'} Q(s',a') - Q(s,a)\right]$$

Outline

- Simplified setup compared to Tsitsiklis 1994
- Lemma 1 and Robbins-Siegmund Theorem
- Theorem 1 in Tsitsiklis 1994
- Theorem 3 in Tsitsiklis 1994
- Extension to Eventual contraction assumption

Simplified Setup

Let $x(t)$ denote the state at discrete time $t \in \mathbb{N}$ with component $x_i(t)$. For each component, we have

$$x_i(t+1) = (1 - \alpha_i(t))x_i(t) + \alpha_i(t)(F_i(x(t)) + w_i(t))$$

where

- $\alpha_i(t) \in [0, 1]$ is the stepsize parameter
- $w_i(t)$ is a noise term

All variables are defined on a probability space $(\Omega, \mathcal{F}, P)$ with an increasing sequence of σ-fields $\{\mathcal{F}(t)\}_{t=0}^{\infty}$ representing the algorithm's history.
For any positive vector $v = (v_1, \ldots, v_n)$, we define the weighted maximum norm:

$$\|x\|_v = \max_i \frac{|x_i|}{v_i}, \quad x \in \mathbb{R}^n \tag{1}$$

Simplified Setup - Assumption 1 - need for all theorems

We assume

(a) $x(0)$ is $\mathcal{F}(0)$-measurable;

(b) For every i and t, $w_i(t)$ is $\mathcal{F}(t+1)$-measurable;

(c) For every i and t, $\alpha_i(t)$ is $\mathcal{F}(t)$-measurable;

(d) For every i and t, we have $\mathbb{E}[w_i(t) \mid \mathcal{F}(t)] = 0$;

(e) There exist constants A and B such that $\mathbb{E}[w_i^2(t) \mid \mathcal{F}(t)] \leq A + B \max_j \max_{\tau \leq t} |x_j(\tau)|^2, \forall i, t.$

Assumption 2 - need for all theorems

We assume

(a) For every i, $\sum_{t=0}^{\infty} \alpha_i(t) = \infty$, w.p.1;

(b) There exists a constant C such that for every i, $\sum_{t=0}^{\infty} \alpha_i^2(t) \leq C$, w.p.1.

Assumption 3 - contraction

There exists a vector $x^* \in \mathbb{R}^n$, a positive vector v, and a scalar $\beta \in [0, 1)$, such that

$$\|F(x) - x^*\|_v \leq \beta \|x - x^*\|_v, \quad \forall x \in \mathbb{R}^n. \tag{2}$$

Assumption 4 - boundedness

There exists a positive vector v, a scalar $\beta \in [0, 1)$, and a scalar D such that

$$\|F(x)\|_v \leq \beta \|x\|_v + D, \quad \forall x \in \mathbb{R}^n. \tag{3}$$

Remark: Assumption 3 implies Assumption 4

Notice that Assumption 3 implies Assumption 4:

$$\begin{aligned}
\|F(x)\|_v &\leq \|F(x) - x^*\|_v + \|x^*\|_v && (\Delta \text{ ineq.}) \\
&\leq \beta\|x - x^*\|_v + \|x^*\|_v && (\text{Assumption 3}) \\
&\leq \beta\|x\|_v + (1+\beta)\|x^*\|_v && (\Delta \text{ ineq.})
\end{aligned}$$

Let $D := (1+\beta)\|x^*\|_v$

Robbins-Siegmund Theorem (Almost supermartingale)

Theorem 1 (Robbins-Siegmund)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* $\{\mathcal{F}_n\}_{n=0}^{\infty}$ *be a filtration. Let* $\{V_n, \beta_n, \xi_n, \zeta_n\}_{n=0}^{\infty}$ *be sequences of non-negative random variables adapted to* $\{\mathcal{F}_n\}_{n=0}^{\infty}$ *such that:*

$$\mathbb{E}[V_{n+1} \mid \mathcal{F}_n] \leq (1+\beta_n)V_n + \xi_n - \zeta_n \quad a.s. \text{ for all } n \geq 0$$

where

- $\sum_{n=0}^{\infty} \beta_n < \infty$ *almost surely*
- $\sum_{n=0}^{\infty} \xi_n < \infty$ *almost surely*

Then:

- $\lim_{n\to\infty} V_n = V_\infty$ *exists and is finite almost surely*
- $\sum_{n=0}^{\infty} \zeta_n < \infty$ *almost surely*

Lemma 2

Let $\{\mathcal{F}(t)\}$ be an increasing sequence of σ-fields. For each t, let $\alpha(t)$, $w(t-1)$, and $B(t)$ be $\mathcal{F}(t)$-measurable scalar random variables. Let C be a deterministic constant. Suppose that the following hold with probability 1:

(a) $\mathbb{E}[w(t) \mid \mathcal{F}(t)] = 0$;

(b) $\mathbb{E}[w^2(t) \mid \mathcal{F}(t)] \leq B(t)$;

(c) $\alpha(t) \in [0, 1]$;

(d) $\sum_{t=0}^{\infty} \alpha(t) = \infty$;

(e) $\sum_{t=0}^{\infty} \alpha^2(t) \leq C$.

Suppose that the sequence $\{B(t)\}$ is bounded with probability 1. Let $W(t)$ satisfy the recursion

$$W(t+1) = (1 - \alpha(t))W(t) + \alpha(t)w(t). \tag{4}$$

Then $\lim_{t \to \infty} W(t) = 0$, with probability 1.

Proof Sketch for Lemma 1

The proof is based on Robbins-Siegmund Theorem

1. We use the squared process $V(t) = W^2(t)$ and show that the squared process fits the condition of Robbins-Siegmund Theorem

$$\mathbb{E}[V(t+1) \mid \mathcal{F}(t)] \leq V(t) + \alpha^2(t)K - \alpha(t)V(t)$$

$$\mathbb{E}[V_{n+1} \mid \mathcal{F}_n] \leq (1+\beta_n)V_n + \xi_n - \zeta_n \quad \text{a.s. for all } n \geq 0$$

2. Use Robbins-Siegmund Theorem to get convergence $V(t) \to V_\infty$ and $\sum_{t=0}^{\infty} \zeta_t = \sum_{t=0}^{\infty} \alpha(t)V(t) < \infty$ almost surely.
3. Prove $V_\infty = 0$ almost surely by contradiction, hence the original process converges to zero almost surely.

$$\begin{aligned} P\{V_\infty \geq 2\epsilon\} > \delta &\implies P(V(t) \geq \epsilon, t \geq T) > \delta \\ &\implies P\left(\sum_{t=0}^{\infty} \alpha(t)V(t) = \infty\right) > \delta \end{aligned}$$

Main Theorem 1 in Tsitsiklis 1994

Let $(\Omega, \mathcal{F}, P)$ be a probability space with filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let $x(t)$ denote the state at discrete time $t \in \mathbb{N}$ with component $x_i(t)$. For each component, we have

$$x_i(t+1) = (1 - \alpha_i(t))x_i(t) + \alpha_i(t)(F_i(x(t)) + w_i(t))$$

If Assumption 1,2,4 holds, then, the sequence $x(t)$ is bounded with probability 1.

Proof Sketch

1. Create a growing envelope $G(t)$ to track the growth of $x(t)$
2. Use this tracking and growing envelope to normalize the noise and this normalized noise fits the condition of lemma 1.
3. We use lemma 1 to show that the normalized noise converges to 0
4. Setup the contradiction by selecting a time t_0 that the noise is very small for all $t \geq t_0$
5. Derive the contradiction by showing the growing envelope is stablized after t_0 by induction

Main Theorem 2 in Tsitsiklis 1994

Let $(\Omega, \mathcal{F}, P)$ be a probability space with filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$. Let $x(t)$ denote the state at discrete time $t \in \mathbb{N}$ with component $x_i(t)$. For each component, we have

$$x_i(t+1) = (1 - \alpha_i(t))x_i(t) + \alpha_i(t)(F_i(x(t)) + w_i(t))$$

If Assumption 1,2,3 holds,then, the sequence $x(t)$ converges to x^* with probability 1.

Proof Sketch

1. Show that $x(t)$ is bounded using Main theorem 1
2. Create a sequence of decreasing bounds $D_0, D_1, D_2, \cdots$ that converges to zero
3. Prove using induction that for each k, the proess eventually stays within the bounds given by D_k, this is the outer induction.
4. To prove the induction step in the outer induction, we use an inner induction to show that the process eventually moves to D_{k+1}.

Extension to Eventual Contraction

Assumption 3 - Contraction:
There exists a vector $x^* \in \mathbb{R}^n$, a positive vector v, and a scalar $\beta \in [0, 1)$, such that

$$\|F(x) - x^*\|_v \leq \beta \|x - x^*\|_v, \quad \forall x \in \mathbb{R}^n. \tag{5}$$

Assumption 3+ - Eventual contraction:
There exists a vector $x^* \in \mathbb{R}^n$, and positive linear operator K with spectral radius $\rho(K) < 1$ such that

$$|F(x) - x^*| \leq K|x - x^*|, \quad \forall x \in \mathbb{R}^n. \tag{6}$$

Lemma - Perturbed nonnegative matrix

Lemma 3

Let A be a n-dimensional nonnegative square matrix with spectral radius $\rho(A) < 1$. Then there exists a strictly positive matrix B such that

$$A < B \text{ and } \rho(B) < 1$$

Remark: One way to show this is via eigenvalue is continuous function of the matrix. But I prove this lemma using Gelfand's formula.

Gelfand's formula

Lemma 4 (Gelfand's formula)

If B is any square matrix and $\|\cdot\|$ is any matrix norm, then

$$\rho(B)^k \leq \|B^k\| \quad \textit{for all } k \in \mathbb{N}$$

$$\|B^k\|^{1/k} \to \rho(B) \ \textit{as } k \to \infty$$

Corollary 5

If B is any square matrix and $\|\cdot\|$ is any matrix form, then if there exists $n \in \mathbb{N}$ such that

$$\|B^n\| < 1$$

this implies $\rho(B) < 1$.

Proof of the lemma

Let J denote the n-dimensional square matrix with every entry equals to 1. We construct $B = A + \epsilon J$. We show that there exists $0 < \epsilon < 1$ such that $\rho(B) < 1$.
Using the Gelfand's formula, we have there exists $N \in \mathbb{N}$ such that for all $n \geq N$, $\|A^n\| < 1$. Fix $n \geq N$. We set $\delta := 1 - \|A^n\|$.
Moreover, we have

$$\begin{aligned} \|B^n\| &= \|(A + \epsilon J)^n\| \\ &= \|A^n + \epsilon(\Gamma_{1,1} + \cdots + \Gamma_{1,C_1^n}) + \cdots + \epsilon^{n-1}(\Gamma_{n-1,1} + \cdots \Gamma_{n-1,C_{n-1}^n}) + \epsilon^n J^n\| \end{aligned}$$

for some square matrix $\Gamma_{i,j}$ and C_j^i be the number of combinations of choosing j objects from i objects.

Remark on the expansion

Moreover, we have

$$\begin{aligned}\|B^n\| &= \|(A+\epsilon J)^n\| \\ &= \|A^n + \epsilon(\Gamma_{1,1} + \cdots + \Gamma_{1,C_1^n}) + \cdots + \epsilon^{n-1}(\Gamma_{n-1,1} + \cdots \Gamma_{n-1,C_{n-1}^n}) + \epsilon^n J^n\|\end{aligned}$$

for some square matrix $\Gamma_{i,j}$ and C_j^i be the number of combinations of choosing j objects from i objects. To motive this step, we have for $n = 2$,

$$\begin{aligned}(A+\epsilon J)^2 &= A^2 + \epsilon AJ + \epsilon JA + \epsilon^2 J^2 \\ &= A^2 + \epsilon(AJ + JA) + \epsilon^2 J^2\end{aligned}$$

Hence, we have $\Gamma_{1,1} = AJ$ and $\Gamma_{1,2} = JA$ with $C_1^2 = 2$.

Proof

Moreover, we have

$$\begin{aligned}\|B^n\| &= \|(A+\epsilon J)^n\| \\ &= \|A^n + \epsilon(\Gamma_{1,1} + \cdots + \Gamma_{1,C_1^n}) + \cdots + \epsilon^{n-1}(\Gamma_{n-1,1} + \cdots \Gamma_{n-1,C_{n-1}^n}) + \epsilon^n J^n\|\end{aligned}$$

for some square matrix $\Gamma_{i,j}$ and C_j^i be the number of combinations of choosing j objects from i objects. Then by triangle inequality, we have

$$\|B^n\| \leq \|A^n\| + \sum_{k=1}^{n-1} \epsilon^k \left(\sum_{j=1}^{C_k^n} \|\Gamma_{k,j}\| \right) + \epsilon^n \|J^n\|$$

Proof

Let

$$M := \max_{1 \le k,j \le n} \{\|\Gamma_{k,j}\|, \|J^n\|\}$$

$$\gamma := \max_{1 \le k \le n} C_k^n$$

By finite dimension, we have M and γ is well-defined and finite. This gives

$$\begin{aligned} \|B^n\| &\le \|A^n\| + \gamma M \sum_{k=1}^{n} \epsilon^k \\ &< \|A^n\| + \gamma M n \epsilon \qquad (0 < \epsilon < 1) \end{aligned}$$

Let $0 < \epsilon < \frac{\delta}{\gamma M n}$. Then, we have

$$\|B^n\| = \|(A + \epsilon J)^n\| < \|A^n\| + \delta < 1$$

By the previous corollary, this implies $\rho(B) < 1$.

Main extension proof - Eventual contraction implies contraction with a specific weighted maximum norm

Suppose there exists a vector $x^* \in \mathbb{R}^n$ and a positive linear operator K with spectral radius $\rho(K) < 1$ such that

$$|F(x) - x^*| \leq K|x - x^*|, \quad \forall x \in \mathbb{R}^n$$

Then, this implies there exists a positive vector $v \in \mathbb{R}^n$ and a scalar $\beta \in [0, 1)$, such that

$$\|F(x) - x^*\|_v \leq \beta \|x - x^*\|_v$$

In other words, eventual contraction assumption implies contraction assumption.

Proof

First, since K is a positive linear operator in a finite dimensional space, it can be represented by a nonnegative matrix with spectral radius $\rho(K) < 1$.
By lemma on perturbed nonnegative matrix, there exists a strictly positive matrix $\tilde{K} > K$ such that $\rho(\tilde{K}) < 1$.
Using the Perron-Frobenius theorem, we know

- the spectral radius $\beta := \rho(\tilde{K}) = \frac{(\tilde{K}v)_i}{v_i} < 1$ is a positive real simple eigenvalue of $\tilde{K}$
- Its corresponding eigenvector v is uniquely positive up to positive scaling.

Proof

Hence, we have pointwise

$$|F_i(x) - x_i^*| \leq (K|x - x^*|)_i \leq (\tilde{K}|x - x^*|)_i, \quad i = 1, 2, \cdots, n$$

as $K < \tilde{K}$. Using the matrix representation, we have

$$(\tilde{K}|x - x^*|)_i = \sum_{j=1}^{n} \tilde{K}_{ij}|x_j - x_j^*|$$

We define

$$\|z\|_v := \max_{1 \leq i \leq n} \frac{|z_i|}{v_i}, \quad \forall z \in \mathbb{R}^n$$

as the weighted maximum norm using v. Hence, this implies

$$\frac{|z_j|}{v_j} \leq \max_{1 \leq i \leq n} \frac{|z_i|}{v_i}, \quad j = 1, 2, \cdots, n$$

proof

Hence,

$$|z_j| \leq v_j \|z\|_v, \quad j = 1, 2, \cdots, n$$

We can apply this to $|x_j - x_j^*|$, we get

$$\begin{aligned}
(\tilde{K}|x - x^*|)_i &= \sum_{j=1}^{n} \tilde{K}_{ij} |x_j - x_j^*| \\
&\leq \sum_{j=1}^{n} \tilde{K}_{ij} v_j \|x - x^*\|_v \\
&= \|x - x^*\|_v \sum_{j=1}^{n} \tilde{K}_{ij} v_j \\
&= \|x - x^*\|_v (\tilde{K} v)_i
\end{aligned}$$

proof

This implies

$$|F_i(x) - x_i^*| \leq \|x - x^*\|_v (\tilde{K}v)_i$$

Now we divide both sides by v_i, we get

$$\frac{|F_i(x) - x_i^*|}{v_i} \leq \frac{(\tilde{K}v)_i}{v_i} \|x - x^*\|_v = \beta \|x - x^*\|_v$$

for all $i = 1, 2, \cdots, n$. Hence, we have

$$\|F(x) - x^*\|_v = \max_{1 \leq i \leq n} \frac{|F_i(x) - x_i^*|}{v_i} \leq \beta \|x - x^*\|_v$$

This completes the proof.

Appendix - direct comparison with Tsitsiklis 1994 setup

We consider iterative updates of a vector $x \in \mathbb{R}^n$ to solve the fixed-point equation $F(x^*) = x^*$, where $F : \mathbb{R}^n \mapsto \mathbb{R}^n$ with component mappings $F_i : \mathbb{R}^n \mapsto \mathbb{R}$.
Let $x(t)$ denote the state at discrete time $t \in \mathbb{N}$, with components $x_i(t)$. For each component i, we have:

$$x_i(t+1) = \begin{cases} x_i(t), & t \notin T^i \\ x_i(t) + \alpha_i(t)(F_i(x^i(t)) - x_i(t) + w_i(t)), & t \in T^i \end{cases} \tag{7}$$

where:

- $T^i \subset \mathbb{N}$ is the set of update times for component i
- $\alpha_i(t) \in [0, 1]$ is the stepsize parameter
- $w_i(t)$ is a noise term
- $x^i(t) = (x_1(\tau_1^i(t)), \ldots, x_n(\tau_n^i(t)))$ contains possibly outdated information with $0 \leq \tau_j^i(t) \leq t$

All variables are defined on a probability space $(\Omega, \mathcal{F}, P)$ with an increasing sequence of σ-fields $\{\mathcal{F}(t)\}_{t=0}^{\infty}$ representing the algorithm's history.

Simplified setup notation

Let $x(t)$ denote the state at discrete time $t \in \mathbb{N}$ with component $x_i(t)$. For each component, we have

$$x_i(t+1) = (1 - \alpha_i(t))x_i(t) + \alpha_i(t)(F_i(x(t)) + w_i(t)) \tag{8}$$

$$\mathbf{x}(t+1) = (I - \mathbf{A}(t))\mathbf{x}(t) + \mathbf{A}(t)(\mathbf{F}(\mathbf{x}(t)) + \mathbf{w}(t))$$

where

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \mathbf{A}(t) = \begin{pmatrix} \alpha_1(t) & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & \cdots & \alpha_n(t) \end{pmatrix}, \mathbf{F}(\mathbf{x}(t)) = \begin{pmatrix} F_1(\mathbf{x}(t)) \\ \vdots \\ F_n(\mathbf{x}(t)) \end{pmatrix}, \mathbf{w}(t) = \begin{pmatrix} w_1(t) \\ \vdots \\ w_n(t) \end{pmatrix}$$

Martingale, sub- and super-martingale

Definition 6

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with filtration $\mathbb{F} = (\mathcal{F}(t))_{t\geq 0}$. A stochastic process $X = (X(t))_{t\geq 0}$ is called a martingale with respect to the filtration $\mathbb{F}$ if

1. X is adaped to $\mathbb{F}$
2. $\mathbb{E}_{\mathbb{P}}|X(t)| < \infty$ for all $t \geq 0$
3. For $s \leq t$, $\mathbb{E}_{\mathbb{P}}(X(t)|\mathcal{F}(s)) = X(s)$

A stochastic process $X(t)$ is called a submartingale if the third condition becomes

$$s \leq t, \mathbb{E}_{\mathbb{P}}(X(t)|\mathcal{F}(s)) \geq X(s)$$

A stochastic process $X(t)$ is called a supermartingale if the third condition becomes

$$s \leq t, \mathbb{E}_{\mathbb{P}}(X(t)|\mathcal{F}(s)) \leq X(s)$$

Full proof for Lemma 1

Proof.

Let us first note that, without loss of generality, we can assume that $B(t) \leq K$ for some constant K almost surely, since the sequence $\{B(t)\}$ is bounded with probability 1.

Step 1: Use the squared process

We analyze the evolution of the squared process $V(t) = W^2(t)$. From the recursion for $W(t)$, we have:

$$W(t+1) = (1-\alpha(t))W(t) + \alpha(t)w(t)$$

Squaring both sides yields:

$$\begin{aligned} W^2(t+1) &= ((1-\alpha(t))W(t) + \alpha(t)w(t))^2 \\ &= (1-\alpha(t))^2 W^2(t) + 2(1-\alpha(t))\alpha(t)W(t)w(t) + \alpha^2(t)w^2(t) \end{aligned}$$

$\square$

Full proof for lemma 1 part 2

Proof.

Taking the conditional expectation with respect to $\mathcal{F}(t)$:

$$\mathbb{E}[W^2(t+1) \mid \mathcal{F}(t)] = (1-\alpha(t))^2 W^2(t) + 2(1-\alpha(t))\alpha(t)W(t)\mathbb{E}[w(t) \mid \mathcal{F}(t)] + \alpha^2(t)\mathbb{E}[w^2(t) \mid$$

Using the conditions $\mathbb{E}[w(t) \mid \mathcal{F}(t)] = 0$ and $\mathbb{E}[w^2(t) \mid \mathcal{F}(t)] \leq B(t) \leq K$, we obtain:

$$\begin{aligned}
\mathbb{E}[V(t+1) \mid \mathcal{F}(t)] &\leq (1-\alpha(t))^2 V(t) + \alpha^2(t)K \\
&= (1 - 2\alpha(t) + \alpha^2(t))V(t) + \alpha^2(t)K \\
&= V(t) - 2\alpha(t)V(t) + \alpha^2(t)V(t) + \alpha^2(t)K \\
&= V(t) - \alpha(t)V(t)(2-\alpha(t)) + \alpha^2(t)K
\end{aligned}$$

□

Full proof of lemma 1

Proof.

Since $\alpha(t) \in [0,1]$, we have $(2-\alpha(t)) \geq 1$, which gives:

$$\begin{aligned}
\mathbb{E}[V(t+1) \mid \mathcal{F}(t)] &\leq V(t) - \alpha(t)V(t) + \alpha^2(t)K \\
&= (1-\alpha(t))V(t) + \alpha^2(t)K \\
&= V(t) + \alpha^2(t)K - \alpha(t)V(t)
\end{aligned}$$

□

Full proof of lemma 1

Proof.

Step 2: Use 1

Now, we let

- $\xi_t = \alpha^2(t)K$, we have

$$\sum_{t=0}^{\infty} \xi_t = \sum_{t=0}^{\infty} \alpha(t)^2 K = K \sum_{t=0}^{\infty} \alpha^2(t) < \infty$$

  by our assumption.

- $\zeta_t = \alpha(t)V(t)$ is nonnegative and adapted to the filtration.

Hence, we use 1, we get

- $\lim_{t\to\infty} V(t) = V_\infty$ exists and is finite almost surely
- $\sum_{t=0}^{\infty} \zeta_t = \sum_{t=0}^{\infty} \alpha(t)V(t) < \infty$ almost surely.

$\square$

Full proof of lemma 1

Proof.

Step 3: Prove $V_\infty = 0$ almost surely by contradiction

Suppose that $P(V_\infty \geq 2\epsilon) > \delta$ for some $\epsilon, \delta > 0$. Then we have on the set $\{\omega : V_\infty(\omega) \geq 2\epsilon\}$, by the definition of limit, for every $\omega \in \{\omega : V_\infty(\omega) \geq 2\epsilon\}$, there exists $T(\omega) \in \mathbb{N}$ such that for all $t \geq T(\omega)$, $V(t,\omega) \geq \epsilon$. Hence for all $\omega \in \{V_\infty \geq \epsilon\}$:

$$\sum_{t=0}^{\infty} \zeta_t(\omega) = \sum_{t=0}^{\infty} \alpha(t) V(t,\omega) \geq \sum_{t=T(\omega)}^{\infty} \alpha(t) V(t,\omega) \geq \epsilon \sum_{t=T(\omega)}^{\infty} \alpha(t)$$

By $\sum_{t=0}^{\infty} \alpha(t) = \infty$, we have $\sum_{t=T(\omega)}^{\infty} \alpha(t) = \infty$. Hence

$$\sum_{t=0}^{\infty} \zeta_t(\omega) \geq \epsilon \sum_{t=T(\omega)}^{\infty} \alpha(t) = \infty$$

□

Full proof of lemma 1

Proof.

This implies

$$\left\{\omega : \sum_{t=0}^{\infty} \zeta_t(\omega) = \infty\right\} \supseteq \{\omega : V_\infty(\omega) \geq 2\epsilon\}$$

Hence

$$P\left(\sum_{t=0}^{\infty} \zeta_t = \infty\right) \geq P\left(V_\infty \geq 2\epsilon\right) > \delta$$

This contradicts to $\sum_{t=0}^{\infty} \zeta_t < \infty$ almost surely.
Hence, this contradiction gives $V_\infty = 0$ almost surely. $\square$

Assumption of noise variance in Q-learning

In short, in Q-learning, the assumption there exists constant A and B such that

$$\mathbb{E}[w_i^2(t)|\mathcal{F}(t)] \leq A + B \max_j \max_{\tau \leq t} |x_j(\tau)|^2, \quad \forall i, t$$

is equivalent to assume the reward process has bounded variance.

Q-learning

In finite dimension, we have the following update equation for the Q-learning:

$$Q(s, a; t+1) = Q(s, a, t) + \alpha(s, a, t)[R(s, a) + \beta \min_{a'} Q(s', a', t) - Q(s, a, t)]$$

This gives us the $F(Q(s, a)) = \mathbb{E}[R(s, a)] + \beta \mathbb{E}\left[\min_{a'} Q(S'(s, a), a')\right]$ Hence, we have

$$Q(s, a; t+1) = Q(s, a, t) + \alpha(s, a, t)[F(Q(s, a, t)) + w(s, a, t) - Q(s, a, t)]$$

Q-learning noise

For the noise part, we have

$$w(s,a,t) = r(s,a) - \mathbb{E}(R(s,a)) + \min_{a'} Q(s',a',t) - \mathbb{E}\left[\min_{a'} Q(S'(s,a),a',t)\right]$$

Hence, the variance is
$\mathbb{E}[w(s,a,t)^2] = \mathbb{E}[(r(s,a) - \mathbb{E}[R(s,a)])^2] + \mathbb{E}[(\min_{a'} Q(s',a',t) - \mathbb{E}[\min_{a'} Q(S'(s,a),a',t)])^2]$
The first term

$$\mathbb{E}[(r(s,a) - \mathbb{E}[R(s,a)])^2] = Var(R(s,a))$$

The second term is

$$Var(\min_{a'} Q(S'(s,a),a',t)) \leq \mathbb{E}[(\min_{a'} Q(S'(s,a),a',t))^2] \leq \max_{s\in S} \max_{a\in A} Q(s,a,t)^2$$

Q-learning

Hence, we have

$$\mathbb{E}[w(s,a,t)^2] \leq Var(R(s,a)) + \max_{s\in S}\max_{a\in A} Q(s,a,t)^2$$

And compare to

$$\mathbb{E}[w_i^2(t)|\mathcal{F}(t)] \leq A + B \max_{j}\max_{\tau\leq t} |x_j(\tau)|^2, \quad \forall i, t$$

We need $Var(R(s,a))$ to be constant and bounded.