Convergence of Q-learning Algoritms.

I_ Stochastic Approximation.

Consider algorithm $x_{n+1} = x_n + a_n\left(h(x_n) + \xi_{n+1}\right)$

where $\mathbb{E}\left[\xi_{n+1} \mid \mathcal{F}_n\right] = 0$ a.s. $\forall n$

$\mathcal{F}_n = \sigma(x_i, i \leq n)$.

Assume h is L-libschitz.

$$\sum_n a_n = \infty, \quad \sum_n a_n^2 < \infty$$

$$\sup_n \|x_n\| < \infty \quad \text{a.s.}$$

$$\mathbb{E}\left[\|\xi_{n+1}\|^2 \mid \mathcal{F}_n\right] \leq K\left(1 + \|x_n\|^2\right) \quad \text{a.s. } \forall n$$

Then the algorithm trajectories look like those of

$$\dot{x} = h(x) \qquad (1)$$

If (1) has a unique globally asymptotic stable point x^* then $\lim_{n\to\infty} x_n = x^*$.

II_ Q_learning.

Let $F : \mathbb{R}^{S\times A} \longrightarrow \mathbb{R}^{S\times A}$

$$F(q)_{sa} = r(s,a) + \gamma \underset{j\sim p(\cdot|s,a)}{\mathbb{E}}\left[\max_b q(j,b)\right]$$

F is a contraction (1-lipschitz)

Q - learning algorithm :

For any pair (s,a), if we look at the n-th times, state s and action a are observed :

$$q_{n+1}(s,a) = q_n(s,a) + \alpha_n \Big[r(s,a) + \gamma \max_b q_n\big(S_{n+1}(s,a), b\big) - q_n(s,a) \Big]$$

⚠ $q_n(s,a)$ is the Q-value after the pair (s,a) has been observed n-times.

If $\sum_n \alpha_n = \infty$, $\sum_n \alpha_n^2 < \infty$, we have a stochastic approximation algorithm, mimicking the dynamical system :

$$\dot{q} = F(q) - q$$

This system is globally stable because is contractive (Exercice), and so Q-learning converges to the unique fixed point of F

\#