Master MVA: Reinforcement Learning **Lecture: 3**

Reinforcement Learning Algorithms

Lecturer: *Alessandro Lazaric* *http://researchers.lille.inria.fr/∼lazaric/Webpage/Teaching.html*

Objectives of the lecture

1. **Understand:** Stochastic approximation.

2. **Use:** TD(λ), Q-learning.

1 Solving MPDs with Unknown Dynamics and Rewards

Remark: the standard DP algorithms explicitly assume that both the transition probability p and the reward function r are known. On the other hand, in the general reinforcement learning setting, this information is not available to the agent which need a direct interaction with the environment (i.e., the MDP) in order to solve it. This is a very common scenario in all those problems where the dynamics is very difficult to formalize precisely (e.g., wind) and the reward function is not explicitly known in advance (e.g., in human-computer interaction problems).

Depending on the level of knowledge and interaction available with the environment we define:

- **Online learning:** At each time t the agent is at state x_t, it takes action a_t, it observes a transition to state x_{t+1}, and it receives a reward r_t. We still assume that $x_{t+1} \sim p(\cdot|x_t, a_t)$ and $r_t = r(x_t, a_t)$ (i.e., MDP assumption) but p and r are unknown. In order to guarantee that an agent could converge to the optimal policy it is critical that the MDP is such that all the states can be experienced multiple times.

- **Episodic learning:** It is possible to generate different trajectories over multiple *episodes*. In each episode the agent is place in an arbitrary state and a policy is followed. An episode terminates after a possibly random amount of time, after which the agent is *reset* and the next episode is started.

- **Learning with generative model:** Although p and r are not known in closed form, a *black-box simulator* of the environment is available, so that for any arbitrary pair (x, a) it is possible to compute the corresponding next state $y \sim p(\cdot|x, a)$ and reward $r = r(x, a)$.

2 Policy Evaluation with Monte-Carlo Algorithms

2.1 Definition

We consider the undiscounted setting with infinite horizon and terminal state. Let π be a proper policies, then the value function is defined as

$$V^\pi(x) = \mathbb{E}\big[\sum_{t=0}^{T-1} r^\pi(x_t) \,|\, x_0 = x; \pi\big],$$

where $r^\pi(x_t) = r(x_t, \pi(x_t))$ and T is the random time when the terminal state is achieved.

A simple estimation of $V^\pi(x)$ can be obtained as the average of the returns of independent trajectories.

Algorithm Definition 1 (Monte-Carlo). Let $(x_0^i = x, x_1^i, \dots, x_{T_i}^i = 0)_{i \le n}$ be a set of n **independent** trajectories all starting from the initial state x and terminating after T_i steps. For any $t < T_i$, we denote by

$$\widehat{R}^i(x_t^i) = \big[r^\pi(x_t^i) + r^\pi(x_{t+1}^i) + \cdots + r^\pi(x_{T_i-1}^i)\big]$$

the **return** of the i-th trajectory at state x_t^i. Then the Monte-Carlo estimator of $V^\pi(x)$ is defined as

$$V_n(x) = \frac{1}{n}\sum_{i=1}^{n}\big[r^\pi(x_0^i) + r^\pi(x_1^i) + \cdots + r^\pi(x_{T_i-1}^i)\big] = \frac{1}{n}\sum_{i=1}^{n}\widehat{R}^i(x)$$

Guarantees. Since the MC estimator is just the empirical mean of n independent random variables all with the same mean $V^\pi(x)$ (i.e., $\mathbb{E}[R^i(x)] = V^\pi(x)$), then according to the *strong law of large numbers* we obtain that

$$V_n(x) \xrightarrow{a.s.} V^\pi(x).$$

2.2 First-visit and Every-visit Monte-Carlo

Remark: any trajectory $(x_0, x_1, x_2, \dots, x_T)$ contains also the sub-trajectory $(x_t, x_{t+1}, \dots, x_T)$ whose return $\widehat{R}(x_t) = r^\pi(x_t) + \cdots + r^\pi(x_{T-1})$ could be used to build an estimator of $V^\pi(x_t)$. Thus one single trajectory provides a sample for the estimation of all the $\{V^\pi(x_t)\}_t$ over all the states traversed by the policy. Furthermore, a trajectory may visit the same state multiple times.

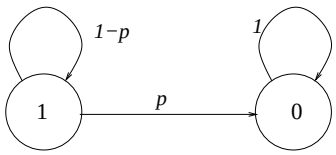
First-visit Monte-Carlo. For each state x we only consider the sub-trajectory when x is first achieved. Although the estimator is unbiased, we only have one sample per trajectory.

Every-visit Monte-Carlo. Given a trajectory $(x_0 = x, x_1, x_2, \dots, x_T)$, we list all the m sub-trajectories starting from x up to x_T. We denote by $(x_0^j = x, x_1^j, x_2^j, \dots, x_{T_j}^j)_{j=1}^m$ the j-th sub-trajectory obtained when state x is observed for the j-th time within the original trajectory $(x_0 = x, x_1, x_2, \dots, x_T)$. In this case the sub-trajectories are not longer independent samples. Furthermore the number of sub-trajectories obtained from one trajectory is itself a random variable. As a result, although from one trajectory we can obtain more than one sample, the average of the sum of rewards over different sub-trajectories could be biased.

In order to compare the two methods, we measure the mean squared error (MSE) of an estimator $\widehat{V}$ w.r.t. V (we omit the dependency on the state x and the policy π since it is clear from the context) as

$$\mathbb{E}\big[(\widehat{V} - V)^2\big] = \underbrace{\big(\mathbb{E}[\widehat{V}] - V\big)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\big[\big(\widehat{V} - \mathbb{E}[\widehat{V}]\big)^2\big]}_{\text{Variance}}$$

Example: 2-state Markov Chain We consider the Markov chain (notice that once π is fixed any MDP reduces to a Markov chain):



The reward is 1 while in state 1 (while is 0 in the terminal state). All the trajectories $(x_0, x_1, \dots, x_T)$ are such that for any $0 \le t \le T-1$, $x_t = 1$ and $x_T = 0$. By using Bellman equations we have that the value function in state 1 can be simply computed as

$$V(1) = 1 + (1-p)V(1) + p0 = \frac{1}{p},$$

since $V(0) = 0$.

First-visit Monte-Carlo. Since all the trajectories start from state 1, then the reward cumulate over one single trajectory is exactly T, i.e., $\widehat{V} = T$. The time-to-end T is a *geometric* random variable since it is the probability of obtaining a 0 from a Bernoulli random variable of probability p. Thus its expectation is

$$\mathbb{E}[\widehat{V}] = \mathbb{E}[T] = \frac{1}{p} = V^\pi(1) \Rightarrow \textbf{unbiased estimator.}$$

Thus the MSE of $\widehat{V}$ coincides with the variance of T, which is

$$\mathbb{E}\Big[\big(T - \frac{1}{p}\big)^2\Big] = \frac{1}{p^2} - \frac{1}{p}.$$

Every-visit Monte-Carlo. Given one single trajectory of length T, we can construct as many sub-trajectories as the number of times state 1 has been visited. Thus the estimator is the average of the T sub-trajectories each of them with a cumulative reward $T - t$, that is

$$\widehat{V} = \frac{1}{T}\sum_{t=0}^{T-1}(T-t) = \frac{1}{T}\sum_{t'=1}^{T} t' = \frac{T+1}{2}.$$

The corresponding expectation is

$$\mathbb{E}\Big[\frac{T+1}{2}\Big] = \frac{1+p}{2p} \neq V^\pi(1) \Rightarrow \textbf{ biased estimator}.$$

Let now consider the case we n independent trajectories are generated, each of length T_i. In this case the total number of samples available is $\sum_{i=1}^n T_i$ and the estimator $\widehat{V}_n$ is

$$\begin{aligned}\widehat{V}_n &= \frac{\sum_{i=1}^n \sum_{t=0}^{T_i-1}(T_i - t)}{\sum_{i=1}^n T_i} = \frac{\sum_{i=1}^n T_i(T_i+1)}{2\sum_{i=1}^n T_i} \\ &= \frac{1/n\sum_{i=1}^n T_i(T_i+1)}{2/n\sum_{i=1}^n T_i} \xrightarrow{a.s.} \frac{\mathbb{E}[T^2] + \mathbb{E}[T]}{2\mathbb{E}[T]} = \frac{1}{p} = V^\pi(1) \Rightarrow \textbf{consistent estimator}.\end{aligned}$$

We now compute the MSE of this estimator:

$$\mathbb{E}\Big[\big(\frac{T+1}{2} - \frac{1}{p}\big)^2\Big] = \frac{1}{2p^2} - \frac{3}{4p} + \frac{1}{4},$$

which is smaller than the MSE of the first-visit Monte-Carlo estimator.

Summing up we have that:

- Every-visit Monte-Carlo: biased but consistent estimator.
- First-visit Monte-Carlo: unbiased estimator with potentially bigger MSE.

Remark: when the state space is large the probability of visiting multiple times the same state is low, then the performance of the two methods tends to be the same.

3 Policy Evaluation with Stochastic Approximation

3.1 The TD(1) Algorithm

The algorithm is similar to the MC algorithm but applies the stochastic approximation idea.[1]

Algorithm Definition 2 (TD(1)). Let $(x_0^n = x, x_1^n, \dots, x_{T_n}^n)$ be the n-th trajectory and $\widehat{R}^n$ be the corresponding return. For all x_t with $t \leq T-1$ observed along the trajectory, we update the value function estimate as

$$V_n(x_t^n) = (1 - \eta_n(x_t^n))V_{n-1}(x_t^n) + \eta_n(x_t^n)\widehat{R}^n(x_t^n). \tag{1}$$

In the following we drop the dependency of the trajectory on the index n for sake of clarity.

Guarantees. Since each sample is an unbiased estimator of the actual value function, i.e.,

$$\mathbb{E}\big[r^\pi(x_t) + r^\pi(x_{t+1}) + \cdots + r^\pi(x_{T-1})|x_t\big] = V^\pi(x_t),$$

then we can directly apply the convergence result in Proposition 7 and obtain that if all the states are visited in an infinite number of trajectories and for all $x \in X$

$$\sum_n \eta_n(x) = \infty$$
$$\sum_n \eta_n(x)^2 < \infty,$$

then $V_n(x) \overset{a.s.}{\to} V^\pi(x)$. Notice that this statement can be extended to any $x \in X$ if all the states are traversed infinitely often.

3.2 The TD(0) Algorithm

The TD(1) algorithm relies on the fact that in its original definition, the state value function V^π is an expected value over random variables, that is

$$V^\pi(x) = \mathbb{E}\big[\sum_{t=0}^{T-1} r^\pi(x_k) \,|\, x_0 = x; \pi\big].$$

[1] Here we assume that no state is not observed multiple times along the trajectory. In practice, the update is done recursively at each state observed along the trajectory, thus the subscript n in V_n is not accurate anymore, since a value function could be update more than once for each new trajectory.

On the other hand, we recall that V^π can be also defined recursively through the Bellman equations and can viewed as the fixed point of the operator $\mathcal{T}^\pi$, so that

$$V^\pi(x) = r(x, \pi(x)) + \sum_{y \in X} p(y|x, \pi(x) V^\pi(x) = \mathcal{T}^\pi V^\pi(x).$$

Then we can apply the stochastic approximation algorithm for contraction operators and obtain the TD(0) algorithm.

In order to apply stochastic approximation for contraction operators, we need a noisy observation of the operator $\mathcal{T}^\pi$. We notice that for any $x \in X$ and any V,

$$\widehat{\mathcal{T}}^\pi V(x_t) = r^\pi(x_t) + V(x_{t+1}), \text{ with } x_t = x,$$

is an unbiased estimator of $\mathcal{T}^\pi V(x)$ since

$$\mathbb{E}[\widehat{\mathcal{T}}^\pi V(x_t)|x_t = x] = \mathbb{E}[r^\pi(x_t) + V(x_{t+1})|x_t = x] = r(x, \pi(x)) + \sum_y p(y|x, \pi(x)) V(y) = \mathcal{T}^\pi V(x).$$

Traditionally, the corresponding algorithm is not defined with $\widehat{V}$ but with the notion of temporal difference.

Definition 1. *At iteration n, given the estimator V_{n-1}, for any observed transition from state x_t to state x_{t+1} we define its corresponding* ***temporal difference*** *as*

$$d_t^n = r^\pi(x_t) + V_{n-1}(x_{t+1}) - V_{n-1}(x_t).$$

Remark: Recalling the definition of Bellman equation for state value function, the temporal difference d_t^n provides a measure of *coherence* of the estimator V_{n-1} w.r.t. the transition $x_t \to x_{t+1}$.

Algorithm Definition 3 (TD(0)). Let $(x_0^n = x, x_1^n, \ldots, x_{T_n}^n)$ be the n-th trajectory, $\{\widehat{\mathcal{T}}^\pi V_{n-1}(x_t^n)\}_t$ the noisy observation of the operator $\mathcal{T}^\pi$, and $(d_t^n)_{t=1}^{T_n}$ be the corresponding temporal differences. For all x_t with $t \leq T-1$ observed along the trajectory, we update the value function estimate as

$$\begin{aligned} V_n(x_t^n) &= (1 - \eta_n(x_t^n)) V_{n-1}(x_t^n) + \eta_n(x_t^n) \widehat{\mathcal{T}}^\pi V_{n-1}(x_t^n) \\ &= (1 - \eta_n(x_t^n)) V_{n-1}(x_t^n) + \eta_n(x_t^n) \big(r^\pi(x_t) + V_{n-1}(x_{t+1})\big) \\ &= V_{n-1}(x_t^n) + \eta_n(x_t^n) d_t^n. \end{aligned} \tag{2}$$

Guarantees. The previous algorithm is a direct application of the stochastic approximation algorithm for fixed point operators, then from Proposition 8 we have that $V_n(x) \overset{a.s}{\to} V^\pi(x)$. Notice that this statement can be extended to any $x \in X$ if all the states are traversed infinitely often.

(**TODO: Need to show the definition of the noise and its boundedness.**)

3.3 Temporal Differences TD(λ)

Comparison between TD(1) and TD(0). Notice that the update scheme in eq. 2 can be written as

$$V_n(x_t) \quad = \quad V_{n-1}(x_t) + \eta_n(x_t)[d_t^n + d_{t+1}^n + \cdots + d_{T-1}^n]$$

where $d_t^n = r^\pi(x_t) + V_{n-1}(x_{t+1}) - V_{n-1}(x_t)$ is the *temporal difference* in the evaluation of V_n when the transition $x_t \to x_{t+1}$ is observed. Thus algorithms $TD(0)$ and $TD(1)$ differ in the temporal difference elements used in the update. The following algorithm proposes an intermediate approach between the two extremes of $TD(1)$ and $TD(0)$ by using a *discount* over temporal differences.

Definition 2 (The $\mathcal{T}_\lambda^\pi$ Bellman operator)**.** *Let $\lambda < 1$ be a fixed parameter, then the Bellman operator $\mathcal{T}_\lambda^\pi$ is a convex combination of the m-step Bellman operators $(\mathcal{T}^\pi)^m$ weighted by a sequences of coefficients defined as a function of a λ as*

$$\mathcal{T}_\lambda^\pi = (1-\lambda)\sum_{m\geq 0}\lambda^m(\mathcal{T}^\pi)^{m+1}. \tag{3}$$

Proposition 1. If π is a proper policy and the Bellman operator $\mathcal{T}^\pi$ is a β-contraction in a weighted $L_{\mu,\infty}$ norm, then Bellman operator $\mathcal{T}_\lambda^\pi$ of parameter λ is a contraction of factor

$$\frac{(1-\lambda)\beta}{1-\beta\lambda} \in [0,\beta].$$

Proof. Let P^π be the transition matrix of the Markov chain induced by the policy π then

$$\begin{aligned}
\mathcal{T}_\lambda^\pi V &= (1-\lambda)\Big[\sum_{m\geq 0}\lambda^m\sum_{i=0}^{m}(P^\pi)^i\Big]r^\pi + (1-\lambda)\sum_{m\geq 0}\lambda^m(P^\pi)^{m+1}V \\
&= \Big[\sum_{m\geq 0}\lambda^m(P^\pi)^m\Big]r^\pi + (1-\lambda)\sum_{m\geq 0}\lambda^m(P^\pi)^{m+1}V \\
&= (I-\lambda P^\pi)^{-1}r^\pi + (1-\lambda)\sum_{m\geq 0}\lambda^m(P^\pi)^{m+1}V.
\end{aligned}$$

Since $\mathcal{T}^\pi$ is a β-contraction in $L_{\mu,\infty}$ then $||P^\pi V||_\mu \leq \beta||V||_\mu$ and $||(P^\pi)^m V||_\mu \leq \beta^m||V||_\mu$. Thus

$$||(1-\lambda)\sum_{m\geq 0}\lambda^m(P^\pi)^{m+1}V||_\mu \leq (1-\lambda)\sum_{m\geq 0}\lambda^m||(P^\pi)^{m+1}V||_\mu \leq \frac{(1-\lambda)\beta}{1-\beta\lambda}||V||_\mu,$$

which implies that $\mathcal{T}_\lambda^\pi$ is a contraction in $L_{\infty,\mu}$ as well. □

Similarly to $TD(1)$ and $TD(0)$ we now apply a stochastic approximation algorithm to the $\mathcal{T}_\lambda^\pi$ Bellman operator and obtain the following algorithm.

Algorithm Definition 4 (Sutton, 1988). Let $(x_0^n = x, x_1^n, \ldots, x_{T_n}^n)$ be the n-th trajectory, and $(d_t^n)_{t=1}^{T_n}$ be the corresponding temporal differences. For all x_t with $t \leq T-1$ observed along the trajectory, we update the value function estimate as

$$V_n(x_t^n) = V_{n-1}(x_t^n) + \eta_n(x_t^n)\sum_{s=t}^{T_n-1}\lambda^{s-t}d_s^n. \tag{4}$$

Guarantees. The previous algorithm is based on the observation that $\sum_{s=t}^{T-1} \lambda^{s-t} d_s | x_t = x$ is an unbiased estimator $\mathcal{T}_\lambda^\pi V_{n-1}(x) - V_{n-1}(x)$, since for any $s \geq t$, (we omit the dependency on n)

$$\begin{aligned}
\mathbb{E}[d_s | x_t = x] &= \mathbb{E}\Big[r^\pi(x_s) + V_{n-1}(x_{s+1}) - V_{n-1}(x_s)\big|x_t = x\Big] \\
&= \mathbb{E}\Big[\sum_{i=t}^{s} r^\pi(x_i) + V_{n-1}(x_{s+1})\big|x_t = x\Big] - \mathbb{E}\Big[\sum_{i=k}^{s-1} r^\pi(x_i) + V_{n-1}(x_s)\big|x_t = x\Big] \\
&= (\mathcal{T}^\pi)^{s-t+1} V_{n-1}(x) - (\mathcal{T}^\pi)^{s-t} V_{n-1}(x),
\end{aligned}$$

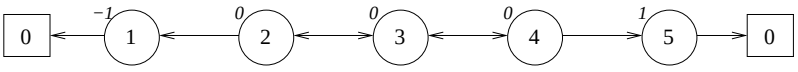
and thus

$$\begin{aligned}
\mathbb{E}[\sum_{s=t}^{T-1} \lambda^{s-t} d_s | x_t = x] &= \sum_{s=t}^{T-1} \lambda^{s-t} \Big[(\mathcal{T}^\pi)^{s-t+1} V_{n-1}(x) - (\mathcal{T}^\pi)^{s-t} V_{n-1}(x)\Big] \\
&= \sum_{m \geq 0} \lambda^m \Big[(\mathcal{T}^\pi)^{m+1} V_{n-1}(x) - (\mathcal{T}^\pi)^m V_{n-1}(x)\Big] \\
&= \sum_{m \geq 0} \lambda^m (\mathcal{T}^\pi)^{m+1} V_{n-1}(x) - \Big[V_{n-1}(x) + \sum_{m>0} \lambda^m (\mathcal{T}^\pi)^m V_{n-1}(x)\Big] \\
&= \sum_{m \geq 0} \lambda^m (\mathcal{T}^\pi)^{m+1} V_{n-1}(x) - \Big[V_{n-1}(x) + \lambda \sum_{m>0} \lambda^{m-1} (\mathcal{T}^\pi)^m V_{n-1}(x)\Big] \\
&= \sum_{m \geq 0} \lambda^m (\mathcal{T}^\pi)^{m+1} V_{n-1}(x) - \Big[V_{n-1}(x) + \lambda \sum_{m \geq 0} \lambda^m (\mathcal{T}^\pi)^{m+1} V_{n-1}(x)\Big] \\
&= (1-\lambda) \sum_{m \geq 0} \lambda^m (\mathcal{T}^\pi)^{m+1} V_{n-1}(x) - V_{n-1}(x) = \mathcal{T}_\lambda^\pi V_{n-1}(x) - V_{n-1}(x).
\end{aligned}$$

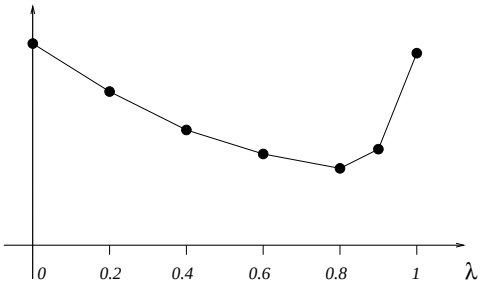
Then from Proposition 8 it follows that $V_n \xrightarrow{a.s.} V^\pi$.

(**TODO: Need to show the definition of the noise and its boundedness.**)

Sensitivity to λ. Simple linear chain example



The MSE of V_n w.r.t. V^π after $n = 100$ trajectories:



- $\lambda < 1$: smaller *variance* w.r.t. $\lambda = 1$ (MC/TD(1)).
- $\lambda > 0$: faster *propagation* of rewards w.r.t. $\lambda = 0$.

Implementation of TD(λ). Although in the previous paragraphs we described the TD algorithms as incremental over trajectories, in practice we apply the update rule at each time step after observing an transition. Thus, in the following we drop the transition index n, since at each step t the current value function estimate V is updated. In particular, TD algorithm are often implemented using the **eligibility traces**$z \in \mathbb{R}^N$. For every transition $x_t \to x_{t+1}$ we first compute the temporal difference $d_t = r^\pi(x_t) + V_n(x_{t+1}) - V_n(x_t)$ and then we update the eligibility traces as

$$z(x) = \begin{cases} \lambda z(x) & \text{if } x \neq x_t \\ 1 + \lambda z(x) & \text{if } x = x_t \\ 0 & \text{if } x_t = 0 \text{ (reset the traces)} \end{cases}$$

and finally for all states x we update the value function estimate as

$$V(x) = V(x) + \eta_t(x) z(x) d_t.$$

TD(λ) in discounted reward MDPs. The Bellman operator $\mathcal{T}_\lambda^\pi$ is defined as

$$\begin{aligned} \mathcal{T}_\lambda^\pi V(x_0) &= (1-\lambda)\mathbb{E}\big[\sum_{t\geq 0} \lambda^t \big(\sum_{i=0}^{t} \gamma^i r^\pi(x_i) + \gamma^{t+1} V(x_{t+1})\big)\big] \\ &= \mathbb{E}\big[(1-\lambda)\sum_{i\geq 0} \gamma^i r^\pi(x_i) \sum_{t\geq i} \lambda^t + \sum_{t\geq 0} \gamma^{t+1} V(x_{t+1})(\lambda^t - \lambda^{t+1})\big] \\ &= \mathbb{E}\big[\sum_{i\geq 0} \lambda^i \big(\gamma^i r^\pi(x_i) + \gamma^{i+1} V(x_{i+1}) - \gamma^i V(x_i)\big)\big] + V_n(x_0) \\ &= \mathbb{E}\big[\sum_{i\geq 0} (\gamma\lambda)^i d_i\big] + V(x_0), \end{aligned}$$

with the temporal difference $d_i = r^\pi(x_i) + \gamma V(x_{i+1}) - V(x_i)$.

The corresponding TD(λ) algorithm becomes

$$V_{n+1}(x_t) = V_n(x_t) + \eta_n(x_t) \sum_{s\geq t} (\gamma\lambda)^{s-t} d_t.$$

4 Q-learning

Remark: all previous algorithms allow to have an (asymptotically) accurate estimation of the value function for any fixed policy π. Nonetheless, in a policy iteration structure, at iteration k, given a policy π_k and an estimate V_n of its value function V^π, the *greedy* policy step requires to compute

$$\pi_{k+1}(x) \in \arg\max_a \big[r(x,a) + \sum_y p(y|x,a) V_n(y)\big].$$

In an online setting, the transition probabilities p are unknown, then it is not possible to compute π_{k+1}. On the other hand, we notice that if an estimate of the *action value function* Q_n is available, the greedy policy step simplifies to

$$\pi_{k+1}(x) \in \arg\max_a Q_n(x,a),$$

which does not require any knowledge about the transition probabilities.

The Q-learning algorithm follows a *value iteration* scheme, whose objective is to directly provide an approximation of the optimal action value function Q^*.

Algorithm Definition 5 (Watkins, 1989). We build a sequence of Q-functions Q_n in such a way that for every observed transition (x, a, y, r), the Q-function in (x, a) is updated as

$$Q_{n+1}(x,a) = (1-\eta_n(x,a))Q_n(x,a) + \eta_n(x,a)\big[r + \max_{b\in A} Q_n(y,b)\big].$$

Proposition 2. *[Watkins et Dayan, 1992]* Let assume that all the policies π are proper and that all the state-action pairs are visited **infinitely often**. If the learning steps satisfy the condition that $\forall x, a$, $\sum_{n\geq 0}\eta_n(x,a) = \infty$, $\sum_{n\geq 0}\eta_n^2(x,a) < \infty$. Then for any $x \in X$, $a \in A$,

$$Q_n(x,a) \xrightarrow{a.s.} Q^*(x,a).$$

Proof. We recall the definition of the optimal Bellman operator $\mathcal{T}$ as

$$\mathcal{T}W(x,a) = r(x,a) + \sum_y p(y|x,a)\max_{b\in A} W(y,b),$$

which has Q^* has the unique fixed point. Since all the policies are proper, then there exist a vector $\mu \in \mathbb{R}^N$ and a scalar $\beta < 1$ such that $\sum_y p(y|x,a)\mu(y) \leq \beta\mu(x)$, which implies that $\mathcal{T}$ is a contraction in the $L_{\mu,\infty}$-norm.

We can rewrite the Q-learning algorithm as an explicit stochastic approximation of the optimal Bellman operator using noisy observations of the operator, that is

$$Q_{n+1}(x,a) = (1-\eta_n(x,a))Q_n(x,a) + \eta_n[\mathcal{T}Q_n(x,a) + b_n(x,a)],$$

where $b_n(x,a)$ is a zero-mean random variable such that$\mathbb{E}[b_n^2(x,a)] \leq c(1+\max_{y,b} Q_n^2(y,b))$ (where c is a constant). Then we can directly apply Proposition 8 and obtain the almost surely convergence result. □

Q-learning in discounted reward MDPs. The previous algorithm can be simply extended to the discounted case as

$$Q_{n+1}(x,a) = Q_n(x,a) + \eta_n(x,a)\big[r + \gamma\max_{b\in A} Q_n(y,b) - Q_n(x,a)\big],$$

which converges to Q^* exactly under the same conditions as before.

A Concentration Inequalities

Definition 3. *Let X be a random variable and $\{X_n\}_{n\in\mathbb{N}}$ a sequence of random variables. Then*

*(a) $\{X_n\}$ converges to X **almost surely**, $X_n \xrightarrow{a.s.} X$, if*

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1,$$

*(b) $\{X_n\}$ converges to X **in probability**, $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}[|X_n - X| > \epsilon] = 0,$$

*(c) $\{X_n\}$ converges to X **in law** (or in distribution), $X_n \xrightarrow{D} X$, if for any bounded continuous function f*

$$\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)].$$

Remark: given the previous definitions we have $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{P} X \implies X_n \xrightarrow{D} X$.

Proposition 3 (Markov Inequality). Let X be a **positive** random variable. Then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}.$$

Proof. We have that $\mathbb{P}(X \geq a) = \mathbb{E}[\mathbb{I}\{X \geq a\}] = \mathbb{E}[\mathbb{I}\{X/a \geq 1\}] \leq \mathbb{E}[X/a]$. □

Proposition 4 (Hoeffding Inequality). Let X be a **centered** random variable bounded in $[a, b]$. Then for any $s \in \mathbb{R}$,

$$\mathbb{E}[e^{sX}] \leq e^{s^2(b-a)^2/8}.$$

Proof. From convexity of the exponential function, we have that for any $a \leq x \leq b$,

$$e^{sx} \leq \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa}.$$

Let $p = -a/(b-a)$ then (recall that $\mathbb{E}[X] = 0$)

$$\begin{aligned} \mathbb{E}e^{sx} &\leq \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= (1 - p + pe^{s(b-a)})e^{-ps(b-a)} = e^{\phi(u)} \end{aligned}$$

with $u = s(b-a)$ et $\phi(u) = -pu + \log(1 - p + pe^u)$. The derivative of ϕ is $\phi'(u) = -p + \frac{p}{p+(1-p)e^{-u}}$. Furthermore $\phi(0) = \phi'(0) = 0$ and $\phi''(u) = \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \leq 1/4$.

Thus from Taylor's theorem, the exists a $\theta \in [0, u]$ such that

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

□

Proposition 5 (Chernoff-Hoeffding Inequality). Let $X_i \in [a_i, b_i]$ be n independent random variables with mean $\mu_i = \mathbb{E}X_i$. Then

$$\mathbb{P}\big(|\sum_{i=1}^{n} X_i - \mu_i| \geq \epsilon\big) \leq 2\exp\Big(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big). \tag{5}$$

Proof. We have

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n} X_i - \mu_i \geq \epsilon) &= \mathbb{P}(e^{s\sum_{i=1}^{n} X_i - \mu_i} \geq e^{s\epsilon}) \\
&\leq e^{-s\epsilon}\mathbb{E}[e^{s\sum_{i=1}^{n} X_i - \mu_i}], \quad \text{Markov inequality} \\
&= e^{-s\epsilon}\prod_{i=1}^{n}\mathbb{E}[e^{s(X_i-\mu_i)}], \quad \text{independent random variables} \\
&\leq e^{-s\epsilon}\prod_{i=1}^{n} e^{s^2(b_i-a_i)^2/8}, \quad \text{Hoeffding inequality} \\
&= e^{-s\epsilon + s^2\sum_{i=1}^{n}(b_i-a_i)^2/8}
\end{aligned}$$

If we choose $s = 4\epsilon/\sum_{i=1}^{n}(b_i - a_i)^2$, then $\mathbb{P}\big(\sum_{i=1}^{n} X_i - \mu_i \geq \epsilon\big) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$. Similar computation for $\mathbb{P}\big(\sum_{i=1}^{n} X_i - \mu_i \leq -\epsilon\big)$ leads to the result in eq. (5). □

B Basics of Stochastic Approximation

B.1 Monte-Carlo Approximation of a Mean

Definition 4. *Let X be a random variable with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{V}[X]$ and $x_n \sim X$ be n i.i.d. realizations of X. The empirical mean built on n i.i.d. realizations is defined as*

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Then $\mathbb{E}[\mu_n] = \mu$, $\mathbb{V}[\mu_n] = \frac{\mathbb{V}[X]}{n}$ and

- **Weak law of large numbers:** $\mu_n \xrightarrow{P} \mu$.
- **Strong law of large numbers**: $\mu_n \xrightarrow{a.s.} \mu$.
- **Central limit theorem (CLT)**: $\sqrt{n}(\mu_n - \mu) \xrightarrow{D} \mathcal{N}(0, \mathbb{V}[X])$.

B.2 Stochastic Approximation of a Mean

Definition 5. *Let X a random variable bounded in $[0, 1]$ with mean $\mu = \mathbb{E}[X]$ and $x_n \sim X$ be n i.i.d. realizations of X. Let the estimator μ_n be defined as $\mu_1 = x_1$, and $n > 1$,*

$$\mu_n = (1 - \eta_n)\mu_{n-1} + \eta_n x_n \tag{6}$$

where (η_n) *is a sequence of* learning steps.

Before stating the main result for the previous estimator, we report a useful lemma.

Proposition 6 (Borel-Cantelli). Let $(E_n)_{n\geq 1}$ be a sequence of events such that $\sum_{n\geq 1}\mathbb{P}(E_n) < \infty$, then the probability of the intersection of an infinite of those elements is 0. More formally,

$$\mathbb{P}\Big(\limsup_{n\to\infty} E_n\Big) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty} E_k\Big) = 0.$$

Proof. Let assume that the probability that an infinite number of event $\mathbb{E}_n$ occur is $\geq a > 0$, i.e. $\mathbb{P}\Big(\limsup_{n\to\infty} E_n\Big) \geq a$ and let denote this set as $\mathcal{E}$. Then the probability of each of the events $E_i \in \mathcal{E}$ is at least a, which leads to $\sum_{n>1}\mathbb{P}(E_n) \geq \sum_{E_n\in\mathcal{E}} a = \infty$, which contradicts the assumption. □

Proposition 7. If for any n, $\eta_n \geq 0$ and are such that

$$\sum_{n\geq 0}\eta_n = \infty, \tag{7}$$

$$\sum_{n\geq 0}\eta_n^2 < \infty, \tag{8}$$

then $\mu_n \xrightarrow{a.s.} \mu$ and we say that μ_n is a **consistent** estimator.

Remark: The learning steps $\eta_n = \frac{1}{n}$ satisfies the previous conditions. Thus the previous proposition corresponds to the **strong law of large numbers** for the empirical mean $\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Proof. We focus on the case $\eta_n = n^{-\alpha}$. The general proof can be found in *On Stochastic Approximation*, Dvoretzky, 1956.

In order to satisfy the two conditions we need $1/2 < \alpha \leq 1$. In fact, as an example we have that

$$\alpha = 2 \Rightarrow \sum_{n\geq 0}\frac{1}{n^2} = \frac{\pi^2}{6} < \infty \text{ (see the Basel problem for the proof)}$$

$$\alpha = 1/2 \Rightarrow \sum_{n\geq 0}\Big(\frac{1}{\sqrt{n}}\Big)^2 = \sum_{n\geq 0}\frac{1}{n} = \infty \text{ (harmonic series).}$$

Case $\alpha = 1$ In this case μ_n is the empirical mean of x_i and thus we can directly apply the Chernoff-Hoeffding inequality in Proposition 5, and we have that for any **fixed** n

$$\mathbb{P}\big(\big|\mu_n - \mu\big| \geq \epsilon\big) \leq 2e^{-2n\epsilon^2}. \tag{9}$$

Although as n increases the probability to have a deviation of ϵ between the empirical mean and the true expectation reduces to zero, this is not enough to prove an *almost surely* convergence. In particular, let $(\epsilon_k)_k$ an arbitrary sequence converging to 0 as $k \to \infty$, then the almost surely convergence can be rewritten as

$$\mathbb{P}\Big(\lim_{n\to\infty}\mu_n = \mu\Big) = \mathbb{P}(\forall k, \exists n_k, \forall n \geq n_k, \big|\mu_n - \mu\big| \leq \epsilon_k) = 1.$$

We define the sequence of events $E_n = \{|\mu_n - \mu| \geq \epsilon\}$. From eq.(9) we have that $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$ which allows us to apply the Borel-Cantelli lemma and obtain that with probability 1 there exist only a **finite** number of n values such that $|\mu_n - \mu| \geq \epsilon$. Given a sequence $(\epsilon_k)_k$ such that $\epsilon_k \to 0$, we have that for any ϵ_k there exist only a finite number of instants were $|\mu_n - \mu| \geq \epsilon_k$, which corresponds to have that there exists a value n_k such that $\mathbb{P}(\forall n \geq n_k, |\mu_n - \mu| \leq \epsilon_k) = 1$. We can repeat the same reasoning for all ϵ_k in the sequence and obtain that

$$\mathbb{P}(\forall k, \exists n_k, \forall n \geq n_k, |\mu_n - \mu| \leq \epsilon_k) = 1,$$

which is exactly the definition of convergence almost surely.

Case $1/2 < \alpha < 1$**.** We first notice that the stochastic approximation μ_n can be written as

$$\begin{aligned}
\mu_1 &= x_1 \\
\mu_2 &= (1-\eta_2)\mu_1 + \eta_2 x_2 = (1-\eta_2)x_1 + \eta_2 x_2 \\
\mu_3 &= (1-\eta_3)\mu_2 + \eta_3 x_3 = (1-\eta_2)(1-\eta_3)x_1 + \eta_2(1-\eta_3)x_2 + \eta_3 x_3 \\
&\dots \\
\mu_n &= \sum_{i=1}^{n} \lambda_i x_i, \qquad (10)
\end{aligned}$$

with $\lambda_i = \eta_i \prod_{j=i+1}^{n}(1-\eta_j)$ such that $\sum_{i=1}^{n} \lambda_i = 1$. The previous expression highlights the fact that μ_n can be interpreted as a weighted sum of the samples x_i, which allows us to apply again the Chernoff-Hoeffding inequality as

$$\mathbb{P}\big(\big|\sum_{i=1}^{n} \lambda_i x_i - \sum_{i=1}^{n} \lambda_i \mathbb{E}[x_i]\big| \geq \epsilon\big) = \mathbb{P}\big(|\mu_n - \mu| \geq \epsilon\big) \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} \lambda_i^2}}.$$

We now need to provide a bound on the sum of the squared coefficients λ_i. From the definition of the coefficients λ_i we have that

$$\log \lambda_i = \log \eta_i + \sum_{j=i+1}^{n} \log(1-\eta_j) \leq \log \eta_i - \sum_{j=i+1}^{n} \eta_j$$

where the inequality follows from $\log(1-x) < -x$. Thus we obtain $\lambda_i \leq \eta_i e^{-\sum_{j=i+1}^{n} \eta_j}$. Furthermore, for any $1 \leq m \leq n$,

$$\begin{aligned}
\sum_{i=1}^{n} \lambda_i^2 &\leq \sum_{i=1}^{n} \eta_i^2 e^{-2\sum_{j=i+1}^{n} \eta_j} \\
&\overset{(a)}{\leq} \sum_{i=1}^{m} e^{-2\sum_{j=i+1}^{n} \eta_j} + \sum_{i=m+1}^{n} \eta_i^2 \\
&\overset{(b)}{\leq} m e^{-2(n-m)\eta_n} + (n-m)\eta_m^2 \\
&\overset{(c)}{=} m e^{-2(n-m)n^{-\alpha}} + (n-m)m^{-2\alpha},
\end{aligned}$$

where (a) follows from $\eta_i \leq 1$, (b) from taking the elements which maximize the summations, and (c) from the definition of $\eta_m = m^{-\alpha}$. Let $m = n^\beta$ with $\beta = (1+\alpha/2)/2$ (i.e. $1 - 2\alpha\beta = 1/2 - \alpha$):

$$\sum_{i=1}^{n} \lambda_i^2 \leq n e^{-2(1-n^{-1/4})n^{1-\alpha}} + n^{1/2-\alpha} \leq 2n^{1/2-\alpha}$$

for n *big enough*, which leads to

$$\mathbb{P}\big(\big|\mu_n - \mu\big| \geq \epsilon\big) \leq e^{-\frac{\epsilon^2}{n^{1/2-\alpha}}}.$$

From this point we follow the same steps as for $\alpha = 1$ (application of the Borel-Cantelli lemma) and obtain the convergence result for μ_n. □

B.3 Stochastic Approximation of a Fixed Point

Let $\mathcal{T} : \mathbb{R}^N \to \mathbb{R}^N$ a contraction in the weighted $L_{\mu,\infty}$, that is, $\exists \beta < 1$, $\exists \mu \in \mathbb{R}^N_{+,*}$, $\forall V_1, V_2 \in \mathbb{R}^N$, $||\mathcal{T}V_1 - \mathcal{T}V_2||_{\mu,\infty} \leq \beta ||V_1 - V_2||_{\mu,\infty}$. We denote by V the fixed point of $\mathcal{T}$. We assume that noisy observations of the unknown operator $\mathcal{T}$ are available, i.e., $\hat{\mathcal{T}}V = \mathcal{T}V + b$, with b a zero-mean noise (i.e., $\mathbb{E}[b] = 0$). Given a series of noisy observations, we apply stochastic approximation to estimate the operator $\mathcal{T}$. For any $x \in X = \{1, \ldots, N\}$, we defined the stochastic approximation

$$V_{n+1}(x) = (1 - \eta_n(x))V_n(x) + \eta_n(x)(\hat{\mathcal{T}}V_n(x)) = (1 - \eta_n(x))V_n(x) + \eta_n(x)(\mathcal{T}V_n(x) + b_n),$$

where η_n is a learning step.

Let $\mathcal{F}_n = \{V_0, \ldots, V_n, b_0, \ldots, b_{n-1}, \eta_0, \ldots, \eta_n\}$ be the history of the algorithm. Then the following proposition guarantees the convergence of the stochastic approximation.

Proposition 8. Let be the noise b have zero mean conditioned on the filtration $\mathcal{F}_n$, i.e., $\mathbb{E}[b_n(x)|\mathcal{F}_n] = 0$ and its variance be bounded as $\mathbb{E}[b_n^2(x)|\mathcal{F}_n] \leq c(1 + ||V_n||^2)$ for a constant c. For any x, the learning rates $\eta_(x)$ are positive and satisfy the stochastic approximation conditions

$$\begin{aligned} \sum_{n \geq 0} \eta_n &= \infty, \\ \sum_{n \geq 0} \eta_n^2 &< \infty, \end{aligned}$$

then for any $x \in X$

$$V_n(x) \xrightarrow{a.s.} V(x).$$

Proof. Although the proof follows similar arguments as before, in this case the arguments are most sophisticated because of the operator $\mathcal{T}$ and different approaches are possible. In *Neuro Dynamic Programming de Bertsekas et Tsitsiklis*, 1996, the proof uses the notion of Lyapunov function. A more general proof is based on a continuous time approximation using ODE (ordinary differential equations), see e.g., Kushner et Yin *Stochastic Approximation and Recursive Algorithms and Applications*, 2003. Finally, a first proof has been simultaneously produced in Jaakola, Jordan et Singh, *On the convergence of Stochastic Iterative Dynamic Programming Algorithms*, 1994, and Tsitsiklis, *Asynchronous Stochastic Approximation and Q-Learning*, 1994. □

B.4 Other Stochastic Approximation Algorithms

Robbins-Monro (1951) algorithm. We want to find the zero of a noisy function f, i.e., solve the equation $f(x) = 0$. Since f is noise, in each x_n we observe $y_n = f(x_n) + b_n$ where b_n is a zero-mean independent noise. Let

$$x_{n+1} = x_n - \eta_n y_n.$$

If f is an increasing function and x^* is the solution, then under the same assumptions on the learning step, we have that $x_n \xrightarrow{a.s.} x^*$.

Kiefer-Wolfowitz (1952) algorithm. We want to find the local minimum of a function f such that its gradient is noisy, that is for any x_n we can compute $g_n = \nabla f(x_n) + b_n$. Then by using the stochastic approximation

$$x_{n+1} = x_n - \eta_n g_n.$$

we obtain that under the same assumptions on the learning steps (η_n) (and an additional assumption on the fact that the Hessian $\nabla^2 f$ is positive), we have that $x_n \xrightarrow{a.s.} x^*$ where x^* is the minimum of f. This is often referred to as the **stochastic gradient** algorithm.