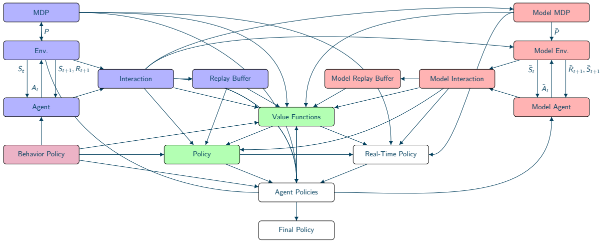
Reinforcement Learning
Reinforcement Learning: Prediction and Planning in the Tabular Setting

Erwan Le Pennec
Erwan.Le-Pennec@polytechnique.edu

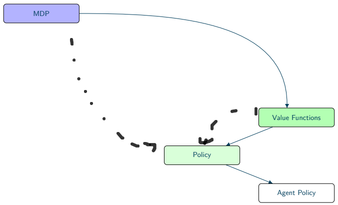
RL: What Are We Going To See?



Outline

- Operations Research and MDP.
- Reinforcement learning and interactions.
- More tabular reinforcement learning.
- Reinforcement and approximation of value functions.
- Actor/Critic: a Policy Point of View

Operations Research and MDP

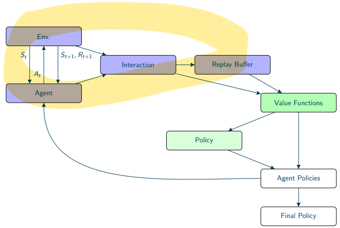


How to find the best policy knowing the MDP?

- Is there an optimal policy?
- How to estimate it numerically?

- Finite states/actions space assumption (tabular setting).
- Focus on interative methods using value functions (dynamic programming).
- Policy deduced by a statewise optimization of the value function over the actions.
- Focus on the discounted setting.

Reinforcement Learning and Interactions



How to find the best policy not knowing the MDP?

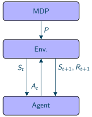
- How to interact with the environment to learn a good policy?
- Can we use a Monte Carlo strategy outside the episodic setting?
- How to update value functions after each interaction?

- Focus on stochastic methods using tabular value functions (Q learning, SARSA…)
- Policy deduced by a statewise optimization of the value function over the actions.

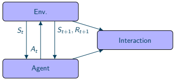
Outline

Reinforcement Learning

From



to



From Probability to Statistics?

- What to do if one has no knowledge of the underlying MDP?
- Only information through interactions!
- Prediction? Planning?
- Focus on the discounted setting

Outline

Monte Carlo, i.e. Just Play!

T time steps
N episodes

- Most simple way to evaluate a policy.

Just Play Following Policy Π

- Play N episodes following the policy.
- During each episode, compute the (discounted) gain.
- Compute the average gain.

- What is computed?

Average Gain or Value Function

$$\mathbb{E}[G_0] \quad \text{vs} \quad v_{t,\Pi}(s) = \mathbb{E}[G_t | S_t = s]$$

Prediction as Value Function Evaluation

- Not the same goal.
- By construction,

$$\mathbb{E}[G_0] = \sum_s \mu_0(s) v_{0,\Pi}(s)$$

- Much easier to compute the average gain than the value function (even if we use a stationary policy)

- Average gain is nevertheless the most classical way to evaluate a policy (with a single number).
- Implicit episodic setting if we do not want to use approximated gain.

Average Gain Estimation

Episodic: Evaluation by MC

input: MDP environment, initial state distribution μ_0, policy Π and discount factor γ
parameter: Number of episodes N
init: $V = 0$,$n = 0$
repeat
- $n \leftarrow n + 1$
- $t \leftarrow 0$
- $G \leftarrow 0$
- Pick initial state S_0 following μ_0
- **repeat**
  - Pick action A_t according to $\pi(\cdot|S_t)$
  - $G \rightarrow G + \gamma^t R_{t+1}$
  - $t \leftarrow t + 1$
- **until** *episod ends at time* T
- $V \leftarrow V + G$

until $n == N$
$V \leftarrow V/N$
output: Average gain V

Monte Carlo Prediction

- How to estimate $v_{t,\Pi}$?

Just Play Following Policy Π

- Play N episodes following the policy.
- During episode, record S_t and R_t.
- After each episode, compute recursively for each time t the gain G_t.
- Estimate $v_{t,\Pi}(s)$ by the average G_t over all trajectories such that $S_t = s$

- **May require a lot of game to have a non empty set for each state s at each time t**

Monte Carlo Prediction

- How to estimate v_Π for a stationary policy?

Just Play Following Policy Π

- Play N episodes following the policy.
- During each episode, record S_t and R_t.
- After each episode, compute recursively for each time t the gain G_t.
- Estimate $v_\Pi(s)$ by the average over all trajectories of all G_t such that $S_t = s$, whatever t.

- The same state may be reached several time during a single episode...
- First-visit variant: Use only the first visit of s for each episode.

Monte Carlo Prediction

Episodic: Prediction by MC

input: MDP environment, initial state distribution μ_0, policy Π and discount factor γ
parameter: Number of episodes N
init: $\forall s, V(s), n = 0, N(s) = 0$
repeat
- $n \leftarrow n + 1$
- $t \leftarrow 0$
- Pick initial state S_0 following μ_0
- **repeat**
  - (If First-visit) $N(S_t) \leftarrow N(S_t) + 1$
  - Pick action A_t according to $\pi(\cdot|S_t)$
  - Record R_{t+1}, S_{t+1}
  - $t \leftarrow t + 1$
- **until** *episod ends at time* T
- $G_{T+1} = 0$
- $t \rightarrow T + 1$
- **repeat**
  - $t \leftarrow t - 1$
  - Compute $G_t = R_{t+1} + \gamma G_{t+1}$
  - (If First-visit) $V(S_t) = V(S_t) + G_t$
- **until** $t = 0$

until $n == N$
for $s \in \mathcal{S}$ **do**
- $V(s) \leftarrow V(s)/N(s)$

end
output: Value function V

Monte Carlo Prediction Analysis

First-Visit Variant Analysis

- Straightforward analysis as all the used values for a given state s are independent.
- Variance of order $1/N(s)$ where $N(s)$ is the number of episods where s is visited.
- Convergence if the number of visit goes to ∞.
- Strong assumption is practice as some states may not be visited by a given policy (if we cannot play on the initial state).

- Every-visit works. . . but not necessarily better!

Outline

Monte Carlo Planning

- Can we use a MC approach to find a good policy?

A First Attempt

- Estimate $v_\pi(s)$ by $V_\pi(s)$ using MC.
- Compute $Q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_\pi(s)$
- Enhance the current policy by setting $\pi(s) = \operatorname{argmax}_a Q_\pi(s,a)$

- Inspired by the Operations Research results. . .
- But unusable as r and p are unknown!

Monte Carlo Planning

A Second Attempt

- Estimate $q_\pi(s, a)$ by $Q_\pi(s, a)$ using MC.
- Enhance the current policy by setting $\pi(s) = \operatorname{argmax}_a Q_\pi(s, a)$

- Requires that $N(s, a)$ the number of times that an episode contains the state s followed by action a goes to ∞.
- Impossible with a deterministic policy!

Monte Carlo Planning

Classical Exploratory Policies...

- Stochastic policies ensuring that any action can occurs at any state.
- ϵ-exploratory policy: use a determistic policy and replace it with a random action with probability ϵ.
- Gibbs policy: use a policy where $\pi(a|s) \propto e^{Q(a,s)} > 0$.

A Final Attempt

- Start from an exploratory policy.
- Estimate $q_\pi(s,a)$ by $Q_\pi(s,a)$ using MC.
- Enhance the current policy while remaining a exploratory policy.

- Last step is not straightforward...
- except for ϵ-deterministic policy for which the ϵ-exploratory policy with base policy $\pi(s) = \operatorname{argmax}_a Q_\pi(s,a)$ works.
- No convergence proof.

Outline

Advanced Implementation of Monte Carlo Prediction

$$X_1 \cdots X_n \rightarrow \overline{X}_n = \frac{1}{n}\sum X_i = \overline{X}_{n-1} + \frac{1}{n}\left(X_n - \overline{X}_{n-1}\right) = \frac{1}{n}X_n + \left(1-\frac{1}{n}\right)\overline{X}_{n-1} = \frac{1}{n}X_n + \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n-1} X_i = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$V_\pi(S_t) \leftarrow V_\pi(S_t) + \alpha(N(S_t))(G_t - V_\pi(S_t))$$

On-Line Monte Carlo

- Average for a given state can be updated each time we have the gain G_t for a state S_t.
- Just use $\alpha(N) = 1/N$ and increment $N(S_t)$.
- No need to record the values between episodes. . .

- We still need to wait until the end of each episode to compute G_t.
- Can we do better?

Advanced MC Prediction

Episodic: Prediction by MC

input: MDP environment, initial state distribution μ_0, policy Π and discount factor γ
parameter: Number of episodes N
init: $\forall s, V(s), n = 0, N(s) = 0$
repeat
  $n \leftarrow n + 1$
  $t \leftarrow 0$
  Pick initial state S_0 following μ_0
  repeat
    (If First-visit) $N(S_t) \leftarrow N(S_t) + 1$
    Pick action A_t according to $\pi(\cdot|S_t)$
    Record R_{t+1}, S_{t+1}
    $t \leftarrow t + 1$
  until *episod ends at time* T
  $G_{T+1} = 0$
  $t \rightarrow T + 1$
  repeat
    $t \leftarrow t - 1$
    Compute $G_t = R_{t+1} + \gamma G_{t+1}$
    (If First-visit) $V(S_t) = V(S_t) + \frac{1}{N(S_t)}(G_t - V(S_t))$
  until $t = 0$
until $n == N$
output: Value function V

- We still need to wait until the end of each episode to compute G_t.
- Can we do better?

Prediction with Temporal Differencies

From $V_\pi(S_t) \leftarrow V_\pi(S_t) + \alpha(N(S_t))(G_t - V_\pi(S_t))$

to $V_\pi(S_t) \leftarrow V_\pi(S_t) + \alpha(N(S_t)) \underbrace{(R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t))}_{\delta_t}$

Bootstrap Strategy

- Replace G_t by an instantaneous estimate $R_{t+1} + \gamma V_\pi(S_{t+1})$.
- Amounts to replace $\gamma R_{t+2} + \gamma^2 R_{t+3}$ by an approximation of its expectation given S_{t+1}: $v_\pi(S_{t+1})$.
- Bootstrap as we use the current estimate $V_\pi(S_{t+1})$ instead of the true value.
- $\delta_t = R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)$ is called a temporal difference.

- No need to wait until the end of the episodes!
- Can be used in the discounted setting.

Discounted / Episodic

TD Prediction

Discounted: Prediction by TD

input: MDP environment, initial state distribution μ_0, policy Π and discount factor γ
parameter: Number of step T
init: $\forall s, V(s), n = 0, N(s) = 0,\ t' = 0$
repeat
- $t \leftarrow 0$
- Pick initial state S_0 following μ_0
- **repeat**
  - $N(S_t) \leftarrow N(S_t) + 1$
  - Pick action A_t according to $\pi(\cdot|S_t)$
  - $V(S_t) \leftarrow V(S_t) + \alpha(N(S_t))\,(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$
  - $t \leftarrow t + 1$
- **until** *episod ends at time* T' *or* $t' = T$

until $t' = T$
output: Value function V

- **But does this work?**

Prediction with Temporal Differencies

$$\mathbb{E}[\delta_t|S_t] = \mathbb{E}[R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)|S_t] = (\mathcal{T}^\pi - \mathrm{Id})\, V_\pi(S_t)$$

TD and Bellman Operator

- TD as an approximate Policy Iteration:
$$\mathbb{E}[V_\pi](S_t) \leftarrow V_\pi + \alpha(N(S_t))\,(\mathcal{T}^\pi - \mathrm{Id})\, V_\pi(S_t)$$
- Proof of convergence of this algorithm to a zero of $\mathcal{T}^\pi - \mathrm{Id}$, i.e. the fixed point of $\mathcal{T}^\pi$!
- Proof requires a mild assumption of α (satisfied by $\alpha(N) = 1/N$) and the strong assumption that $N(s)$ goes to ∞.

- MC could be interpreted in a similar way (stochastic approximation) by noticing that $\mathbb{E}[G_t - V_\pi(S_t)|S_t] = v_\pi(S_t) - V_\pi(S_t)$.
- Often use with a constant α

MC vs TD

$$
\begin{aligned}
& V_\pi(S_t) \leftarrow V_\pi(S_t) + \alpha(N(S_t))(G_t - V_\pi(S_t)) \\
\text{or} \quad & V_\pi(S_t) \leftarrow V_\pi(S_t) + \alpha(N(S_t)) \underbrace{(R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t))}_{\delta_t}
\end{aligned}
$$

MC vs TD

- Both are based on stochastic approximation.
- Both converges (under similar assumptions) to the correct value function.
- TD does not require to wait until the end of the episode.
- No theorical difference in the speed of convergence but often TD is better...
- Solve different approximate problems when used with a finite set of episodes:
  - MC compute the empirical gain from any state.
  - TD compute the value function of the empirical Bellman operator (the one obtained by using the empirical transition probabilities)

- If V_π is kept constant during an episode

$$G_t - V_\pi(S_t) = \sum \gamma^{t'-t} \delta_t$$

Outline

Stochastic Approximation

$$\theta_{k+1} = \theta_k + \alpha_k h_k(\theta_k) \quad \text{with} \quad h_k(\theta) = H(\theta) + \epsilon_k + \eta_k$$
$$\implies \theta_k \to \{\theta, H(\theta) = 0\}$$

Stochastic Approximation

- Family of sequential stochastic algorithm converging to a zero of a function.
- Classical assumptions:
  - $\mathbb{E}[\epsilon_k] = 0$, $\mathbb{V}\text{ar}\,[\epsilon_k] < \sigma^2$, and $\mathbb{E}[\|\eta_k\|] \to 0$,
  - $\sum_k \alpha_k \to \infty$ and $\sum_k \alpha_k^2 < \infty$,
  - the algorithm converges if we replace h_k by H.

- Convergence toward a neighborhood if α is kept constant (as often in practice).
- Most famous example are probably Robbins-Monro and SGD.
- Proof quite technical in general.
- The convergence with H is easy to obtain for a contraction.

Stochastic Approximation and ODE

From $\theta_{k+1} = \theta_k + \alpha_k h_k(\theta_k)$ with $h_k(\theta) = H(\theta) + \epsilon_k + \eta_k$

to $\frac{d\tilde{\theta}}{dt} = H(\tilde{\theta})$

TD = $h = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$

$H = (\mathcal{T}^\pi - Id) V$

ODE Approach

- General proof showing that the algorithm converges provided the ODE converges.
- Rely on the rewriting the equation

$$\frac{\theta_{k+1} - \theta_k}{\alpha_k} = h_k(\theta_k) = H(\theta_k) + \epsilon_k + \eta_k$$

$\frac{dV}{dt} = H(V) = -(I - \gamma P_\pi)V + r_\pi$

- α_k can be interpreted as a time difference allowing to define a time $t_k = \sum_{t' \leq t} \alpha_k$.
- $\theta(t)$ is piecewise affine and defined through its derivative at time $t \in (t_k, t_{k+1})$.
- This piecewise function remains close to any solution of the ODE starting from θ_k for an arbitrary amount of time provided k is large enough.

- More general proofs based on martingale.

Asynchronous Update

$$\text{From} \quad \theta_{k+1} = \theta_k + \alpha_k h_k(\theta_k) \quad \text{with} \quad h_k(\theta) = H(\theta) + \epsilon_k + \eta_k$$
$$\text{to} \quad \forall i, \theta_{k+1}(i) = \theta_k(i) + \alpha_k(i) h_k(\theta_k)(i)$$

Asynchronous Update

- Componentwise action on θ.
- Not necessarily the same stepsize $\alpha_k(i)$ for all components.
- $\alpha_k(i) = 0$ is permitted!
- Previous results hold provided for every component i, $\sum_k \alpha_k(i) \to \infty$ and $\sum_k \alpha_k^2(i) < \infty$,

- Exact setting of TD approximation!

Outline

Planning with Temporal Differencies

A State Value Function Attempt

- $V_\star$ is the fixed point of $\mathcal{T}^\star$.
- Approximate it as the zero of $\mathcal{T}^\star - \mathrm{Id}$.
- By construction

$$\mathcal{T}^\star v(S_t) = \max_a \mathbb{E}[R_{T+1} + \gamma v(S_{t+1}) | S_t, a]$$

- Not an expectation!

A State-Action Value Function Attempt

- $q_\star$ is the fixed point of $\mathcal{T}^\star$.
- Approximate it as the zero of $\mathcal{T}^\star - \mathrm{Id}$.
- By construction

$$\mathcal{T}^\star q(S_t, A_t) = \mathbb{E}\left[R_{t+1} + \gamma \max_a q(S_{t+1}, a) \middle| S_t, A_t\right]$$

- An expectation!

Q Learning

Discounted: Planning by Q-Learning

input: MDP environment, initial state distribution μ_0, policy Π and discount factor γ
parameter: Number of step T
init: $\forall s, a, Q(s,a), N(s,a) = 0$, n=0, $t' = 0$
repeat
- $t \leftarrow 0$
- Pick initial state S_0 following μ_0
- **repeat**
  - $N(S_t) \leftarrow N(S_t) + 1$
  - Pick action A_t according to $\pi(\cdot|S_t)$
  - $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha(N(S_t, A_t))\left(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right)$
  - $t \leftarrow t + 1$
  - $t' \leftarrow t' + 1$
- **until** *episod ends at time* T' *or* $t' = T$

until $t' = T$
output: Deterministic policy $\tilde{\pi}(s) = \text{argmax}_a Q(s,a)$

Planning with Q Learning

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha(N(S_t, A_t)) \left(\underbrace{R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)}_{\delta_t} \right)$$

Q-Learning

- Update is independent of the policy Π.
- Convergence of the Q-value function provided the policy is such that $N(s, a)$ tends to ∞ for any state and any action.
- Implies a convergence of the policy.
- Relies on temporal difference.

- Most classical (tabular) planning algorithm!

Outline

Planning with Policy Improvement

from
$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha(N(S_t, A_t)) \left(\underbrace{R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)}_{\delta_t} \right)$$

to
$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha(N(S_t, A_t)) \left(\underbrace{R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)}_{\delta_t} \right)$$

$$\Pi(S_t) = \operatorname*{argmax}_a Q(S_t, a) (\text{plus exploration})$$

Policy Improvement

- More emphasis on the policy with a link between the policy used to play and the optimized policy.
- Almost equivalent to use the current policy in the Q-Learning algorithm.

Discounted

SARSA

Discounted: Planning by SARSA

input: MDP environment, initial state distribution μ_0, policy Π and discount factor γ
parameter: Number of step T
init: $\forall s, a, Q(s,a), N(s,a) = 0$, n=0, $t' = 0$
repeat
- $t \leftarrow 0$ Pick initial state S_0 following μ_0
- **repeat**
  - $N(S_t) \leftarrow N(S_t) + 1$
  - Pick action A_t according to $\pi(\cdot|S_t)$
  - $Q(S_{t-1}, A_{t-1}) \leftarrow Q(S_{t-1}, A_{t-1}) + \alpha(N(S_{t-1}, A_{t-1}))\,(R_t + \gamma Q(S_t, A_t) - Q(S_{t-1}, A_{t-1}))$
  - $\Pi(S_{t-1}) = \operatorname{argmax}_a Q(S_{t-1}, a)$ (plus exploration)
  - $t \leftarrow t + 1$
  - $t' \leftarrow t' + 1$
- **until** *episod ends at time T' or $t' = T$*

until $t' = T$
output: Deterministic policy $\tilde{\pi}(s) = \operatorname{argmax}_a Q(s,a)$

- Does this work?

SARSA and exploration

$$\Pi(S_t) = \operatorname*{argmax}_a Q(S_t, a)(\text{plus exploration})$$

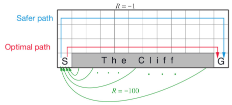
SARSA and Exploration

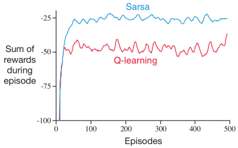
- No hope of convergence if we do not explore all possible actions (and states).
- Impossible if the policy used is deterministic.
- Exploration is required!
- Most classical choice: ϵ-greedy policy with a decaying ϵ.

- Convergence proof is harder than for Q-Learning.
- Relies on the similarity in the limit (when ϵ goes to 0) with the Q-Learning algorithm.

Outline

Q-Learning vs SARSA





How different are they?

- In Q-learning, the exploratory policy used is decoupled from the optimized policy.
- This exploratory policy may yield low rewards on average.
- In SARSA, the two policies are linked with the hope on having higher rewards during the learning step.
- Subtle different behavior even if we modify the exploratory policy in Q-Learning.

Exploration vs Exploitation

Exploration vs Exploitation

- Exploration: explore new policies to be able to discover the best ones.
- Exploitation: use good policies to obtain a good return.
- Exploration is a requirement.

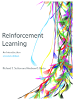
- No tradeoff if we optimize only the final result!
- Tradeoff between the two if we consider that the returns during training matters!
- Q-learning use the first approach and SARSA try to tackle the second.
- Tradeoff if we study a regret:

$$\sum_t \left[\mathbb{E}_{\Pi_\star}[R_t] - \mathbb{E}_{\Pi_t}[R_t] \right]$$

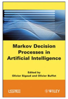
  which forces us to be good as fast as possible.
- No natural definition in the discounted setting.

Outline

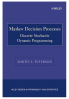
References



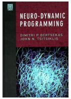
R. Sutton and A. Barto.
Reinforcement Learning, an Introduction (2nd ed.)
MIT Press, 2018



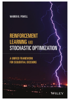
O. Sigaud and O. Buffet.
Markov Decision Processes in Artificial Intelligence.
Wiley, 2010



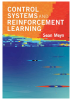
M. Puterman.
Markov Decision Processes. Discrete Stochastic Dynamic Programming.
Wiley, 2005



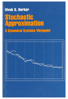
D. Bertsekas and J. Tsitsiklis.
Neuro-Dynamic Programming.
Athena Scientific, 1996



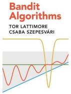
W. Powell.
Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions.
Wiley, 2022



S. Meyn.
Control Systems and Reinforcement Learning.
Cambridge University Press, 2022



V. Borkar.
Stochastic Approximation: A Dynamical Systems Viewpoint.
Springer, 2008



T. Lattimore and Cs. Szepesvári.
Bandit Algorithms.
Cambridge University Press, 2020

Licence and Contributors

Contributors

- Main contributor: E. Le Pennec
- Contributors: S. Boucheron, A. Dieuleveut, A.K. Fermin, S. Gadat, S. Gaiffas, A. Guilloux, Ch. Keribin, E. Matzner, M. Sangnier, E. Scornet.