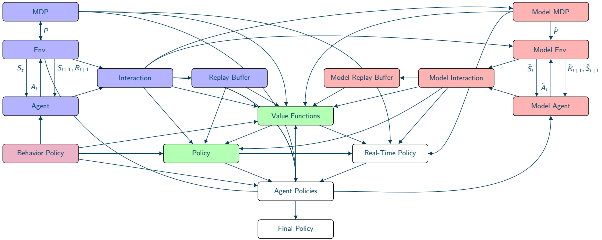
Reinforcement Learning
Extensions

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

M2DS - Reinforcement Learning – Fall 2024

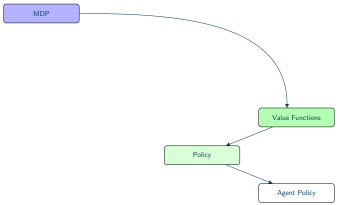
RL: What Are We Going To See?



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
- Extensions

Operations Research and MDP

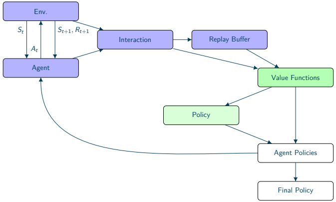


How to find the best policy knowing the MDP?

- 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

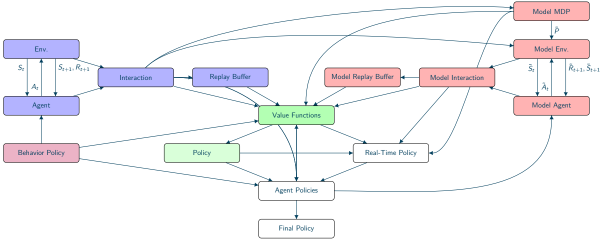


How to find the best policy not knowing the MDP?

- 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.

More Tabular Reinforcement Learning

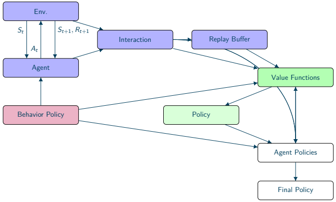


Can We Do Better?

- Is there a gain to wait more than one step before updating?
- Can we interact with a different policy than the one we are estimating?
- Can we use an estimated model to plan?
- Can we plan in real-time instead of having to do it beforehand?

- Finite states/actions space setting (tabular setting).

Reinforcement and Approximation of Value Functions

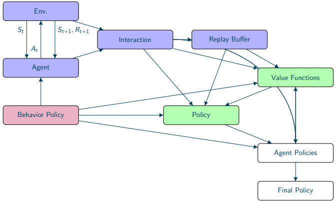


How to Deal with a Large/Infinite states/action space?

- How to approximate value functions?
- How to estimate good approximation of value functions?

- Finite action space setting.
- Stochastic algorithm (Deep Q Learning...).
- Policy deduced by a statewise optimization of the value function over the actions.

Actor/Critic: a Policy Point of View

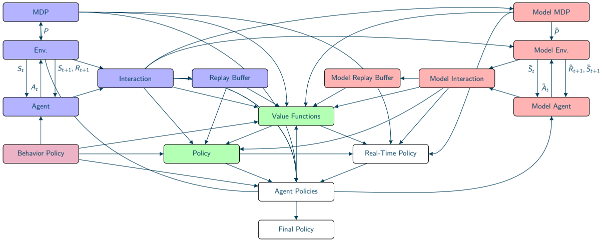


Could We Directly Parameterized the Policy?

- How to parameterize a policy?
- How to optimize this policy?
- Can we combine parametric policy and approximated value function?

- State Of The Art Algorithms (DPG,PPO, SAC...)

Extensions



Can We Do Something Different in This Setting?

- How to deal with the total and average returns?
- How to deal with partial observations?
- How to learn a policy or an implicit reward by observing an actor?

Outline

Outline

Total Reward

$$v_\Pi(s) = \mathbb{E}_\Pi\left[\sum_{t'=1}^{+\infty} R_{t+1} \middle| S_0 = s\right]$$

- Total reward not necessarily well defined!
- Need to **assume** this is the case!

Properness Assumptions - Finite duration of episodes

- H-proper policy: It exists an absorbing state s_{abs} such that $\forall s,\ \mathbb{E}_\Pi[\min_{t, S_t = s_{\text{abs}}} t | S_0 = s] \leq H < +\infty$
- Episodic model: every policy is H-proper $\sim$ discounted setting for a weighted sup-norm.
- Stochastic Shortest Path: there is a proper policy and any non proper policy Π is such that $\exists s, v_\Pi(s) = -\infty$.

- Other models proposed by Puterman (Positive Bounded and Negative Models) have been abandoned by Puterman himself!

Bellman Operator and Optimality Equation

$$\sup_{\Pi} v_{\Pi}(s) = v_\star(s) = \underbrace{\max_a r(s,a) + \sum_{s'} p(s'|s,a) v_\star(s')}_{\mathcal{T}^\star(v_\star)(s)}$$

- Similar to the discounted setting as:
    - We can focus on Markovian policy.
    - The optimal value $v_\star$ satisfies the Bellman optimality equation.

But...

- $\mathcal{T}^\star$ is not a contraction and thus there may be several solutions of the equation.
- If π is such that $\mathcal{T}^\pi v_\star = \mathcal{T}^\star v_\star$, we need to assume that $\limsup (P^\pi)^n v_\star(s) \leq 0$ to prove that $\Pi = (\pi, \pi, \ldots)$ is optimal.
- There may not exist an optimal policy!

- Existence of optimal policies in the finite state-action setting by defining the total reward to the limit of discounted setting when $\gamma \to 1$ and using the finiteness of the policy set...

Stochastic Shortest Path

$$\Pi \ H\text{-proper} \Leftrightarrow \forall s, \ \mathbb{E}_\Pi\left[\min_{t, S_t = s_{\text{abs}}} t \middle| S_0 = s\right] \leq H < +\infty$$

Assumptions

- It exists a proper policy.
- For any improper policy, it exists s such that $v_\Pi(s) = -\infty$.

Properties

- For any proper policy, v_π is the unique solution of $v = \mathcal{T}^\pi v$, and $\mathcal{T}^\pi$ is a contraction.
- $v_\star$ is the unique solution of $v = \mathcal{T}^\star v$.
- Value Iteration and Policy Iteration converge in a stable manner.
- Modified Policy Iteration converges provided $v_0 \leq \mathcal{T}^\star v_0$.

Stochastic Shortest Path and Reinforcement Learning

$$\delta_t = R_t + Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)$$

Prediction

- Convergence of TD-learning algorithms for any proper policy.

$$\delta_t = R_t + \max_Q(S_{t+1}, a) - Q(S_t, A_t)$$

Planning

- Convergence of Q-learning algorithms is the Stochastic Shortest Path setting if the Q estimates remain bounded.

- See *Neuro-Dynamic Programming* from Bertsekas and Tsitsiklis!
- May be very slow in practice!

Stochastic Shortest Path and Policy Gradient

$$\begin{aligned}\nabla v_{\pi_\theta}(s) &= \sum_{t'} \mathbb{E}_{\pi_\theta}[\nabla \log \pi_\theta(A_{t'}|S_{t'}) a_{\pi_\theta}(S_{t'}, A_{t'}) | S_0 = s] \\ &= \sum_s \left(\sum_t \mathbb{P}_{\pi_\theta}(S_t = s | S_0 = s) \right) \left(\sum_a \pi_\theta(a|s) \nabla \log \pi_\theta(a|s) q_{\pi_\theta}(s, a) \right)\end{aligned}$$

Policy Gradient

- Formula valid in the Stochastic Shortest Path Assumption (if the current policy is proper).
- Approximate Policy Improvement Lemma with a H^2 multiplicative constant (instead of $O(H)$).

Actor-Critic

- Valid approach provided all the policies considered remain propers.
- Main difficulty is to maintain a good estimate of $q_{\pi_\theta}\ldots$

Outline

Average Return

$$\overline{v}_\Pi(s) = \lim_{T\to\infty} \frac{1}{T} v_{T,\Pi}(s) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_\Pi\left[\sum_{t=1}^{T} R_t \middle| S_0 = s\right]$$

$$\longrightarrow \overline{v}_{+,\Pi}(s) = \limsup_{T\to\infty} \frac{1}{T} v_{T,\Pi}(s)$$

$$\overline{v}_{-,\Pi}(s) = \liminf_{T\to\infty} \frac{1}{T} v_{T,\Pi}(s)$$

Average Return(s)

- Limit $\overline{v}_\Pi$ may not be defined!
- **Prop:** $\overline{v}_\Pi$ is well defined if Π is stationary and $\frac{1}{T}\sum_{t=1}^{T}(P^\pi)^{t-1}$ tends to a stochastic matrix.
- Limits $\overline{v}_{+,\Pi}$ and $\overline{v}_{-,\Pi}$ always defined!

Average Returns and Optimality

$$\overline{v}_{+,\star}(s) = \sup_{\Pi} \overline{v}_{+,\Pi}(s) \quad \text{and} \quad \overline{v}_{-,\star}(s) = \sup_{\Pi} \overline{v}_{-,\Pi}(s)$$

Optimality of $\Pi_\star$

- Average optimal:

$$\overline{v}_{-,\Pi_\star} \geq \overline{v}_{+,\star}(s)$$

- Lim-sup average optimal (best case analysis):

$$\overline{v}_{+,\Pi_\star} \geq \overline{v}_{+,\star}(s)$$

- Lim-inf average optimal (worst case analysis):

$$\overline{v}_{-,\Pi_\star} \geq \overline{v}_{-,\star}(s)$$

- More complex setting!
- Let's start with Prediction...

Prediction for a Stationary Markov Policy

$$\bar{v}_\Pi(s) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_\pi^{t-1} r_\pi = \left(\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} P_\pi^{t-1} \right) r_\pi = P_\pi^\infty r_\pi$$

Stochastic Matrix P_π^∞

- Measures the average amount of time spend on a state s' starting from state s at $t = 0$ when using policy π.
- Structure linked to the properties of the resulting Markov chain:
  - If aperiodic, $P_\pi^\infty = \lim_T P_\pi^T$ i.e. P_π^∞ is close to the probability of reaching s' from s at any large T.
  - If unichain, then P_π^∞ has identical rows and corresponds to the stationary distribution.
  - If multichhain, then P_π^∞ has a diagonal block structure with rows equal withing each block corresponding to the stationary distribution in each chain.

- Implies that $\bar{v}_\Pi(s) = \bar{v}_\Pi(s')$ in the Markov process is unichain.
- Limit P_π^∞ may be hard to compute...

Average Reward and Relative Value Functions

$$U_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=1}^{\infty}(R_t - \overline{v}_\pi(S_t))\middle| S_0 = s\right] \quad \Leftrightarrow U_\pi = \underbrace{(\mathrm{Id} - P_\pi + P_\pi^\infty)^{-1}(\mathrm{Id} - P_\pi^\infty)}_{H_\pi} r_\pi$$

Link between U_π and $\overline{v}_\pi$

- $(\mathrm{Id} - P_\pi)\overline{v}_\pi = 0$
- $\overline{v}_\pi + (I - P_\pi)U_\pi = r_\pi$

Characterization by a system

- If $(\mathrm{Id} - P_\pi)\overline{v} = 0$ and $\overline{v} + (I - P_\pi)U = r_\pi$ then
  - $\overline{v} = \overline{v}_\pi$,
  - $U = U_\pi + u$ with $(I - P_\pi)u = 0$,
  - If $P_\pi^\infty U = 0$ then $u = 0$.

- Prediction possible by solving this system as we do not need U_π.

Optimality Equations

$$\overline{v}(s) = \max_a \sum_{s'} p(s'|s,a)\overline{v}(s')$$

$$U(s) + \overline{v}(s) = \max_{a \in B_s} r(s,a) + \sum_{s'} p(s'|s,a)U(s) \text{with } B_s = \{a | \sum_{s'} p(s'|s,a)\overline{v}(s') = \overline{v}(s)\}$$

$$\pi_\star(s) \in \operatorname*{argmax}_{a \in B_s} r(s,a) + \sum_{s'} p(s'|s,a)U(s)$$

Existence

- If there is a solution $(\overline{v}, U)$ of the system then $\overline{v} = \overline{v}_\star$ and $\pi_\star$ is an optimal policy.
- There may exist other optimal policies not satisfying the argmax property.
- There may not exist solutions to the system.

- Associated relative value iteration and modified policy iteration can be defined.
- Convergence under strong assumptions. . .

Average Return and Relative Value Functions

$$r(\pi) = \lim_T \mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=0}^{T-1} R_t\right] = \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) r$$

$$G_t = \sum_{t' \geq t} (R_t - r(\pi))$$

$$v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s] \quad \text{and} \quad q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

Connection with Stochastic Shortest Path

- Provided there is a state s that is visited with positive probability in the first m steps for any starting state and any policy.
- $r(\pi)$ is the average cost between a visit s and the next one...

Reinforcement Learning Algorithms

- Simultaneous estimation of q and r...
- Much less theory as there is no contraction!

Algorithm(s)

Average: 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, $t = 0$, $r = 0$
Pick initial state S_0 following μ_0
repeat

$N(S_t) \leftarrow N(S_t) + 1$
Pick action A_t according to $\pi(\cdot|S_t)$
$Q(S_{t-1}, A_{t-1}) \leftarrow Q(S_{t-1}, A_{t-1}) + \alpha(N(S_{t-1}, A_{t-1}))\,(R_t - r_{t-1} + \gamma Q(S_t, A_t) - Q(S_{t-1}, A_{t-1}))$
$r \leftarrow r + \alpha_t(R_t - r)$
$\Pi(S_{t-1}) = \mathrm{argmax}_a\, Q(S_{t-1}, a)$ (plus exploration)
$t \leftarrow t + 1$

until $t == T$
output: Deterministic policy $\tilde{\pi}(s) = \mathrm{argmax}_a\, Q(s, a)$

- Q-learning variant (known as R-learning) and other estimations of r exist.
- No convergence proof.

Policy Gradient

$$\nabla r(\pi) = \lim_{T} \frac{1}{T} \mathbb{E}_{\pi}\left[\sum_{i=1}^{T} \nabla \log \pi(A_t|S_t) q_{\pi}(S_t, A_t)\right]$$

$$\nabla r(\pi) = \lim_{T} \frac{1}{T} \mathbb{E}_{\pi}\left[\sum_{i=1}^{T} \nabla \log \pi(A_t|S_t) a_{\pi}(S_t, A_t)\right]$$

Policy Gradient

- REINFORCE type algorithms, using MC estimate of q and a are possible,
- but q and a are the relative ones, not the classical ones, and are much harder to estimate.

- Actor/Critic algorithms combining parametric estimation of q (or a) and gradient exist.

Outline

To Discount or Not?

$$\text{To Discount:} \quad J(\pi) = \mathbb{E}_\pi\left[\sum_t \rho^t R_t\right] \qquad Q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_t \rho^t R_t \middle| s_0 = s, a_0 = a\right]$$

$$\text{or Not (SSP):} \quad J(\pi) = \mathbb{E}_\pi\left[\sum_t R_t\right] \qquad Q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_t R_t \middle| s_0 = s, a_0 = a\right]$$

To Discount or Not? Open Question!

- Discount is (quite) artificial.
- No discount in the evaluation part most of the time.
- Discount often used in training due to better convergence for value functions. . . toward a (quite) artificial policy target!

- In practice, often hybrid scheme with no discount for the policy gradient part, but discount for the value functions part! No strong justification but often better numerical performance!
- Average reward much less used!

Outline

POMDP

$$o \sim \mathbb{P}(\cdot | s, a)$$

Partially Observed Markov Decision Process

- MDP strongest assumption is that s is observed!
- POMDP replaces this assumption by the observation of o with a known law of $\mathbb{P}(o|s, a)$.
- Can be recasted as a MDP where the state is the probability of being in a state s given the current observation!
- Much higher dimensional setting!

- Policy gradient algorithms remain valid in the POMDP setting when replacing s with o.
- Difficult part is to obtain a good value function estimate.

Outline

Imitation Learning

$$\text{Good } S_t, A_t, (R_{t+1},) S_{t+1}, A_{t+1} \rightarrow \pi$$

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{t} \log \pi_\theta(A_t | S_t)$$

Imitation Learning

- Learn policy from demonstrations (observations).
- Most classical approach: maximum likelihood.
- Need to cover all states (possibly through the approximation)
- Reward is not used.

- DAGGER: Sequential approach to add feedback from trajectory with an estimated policy through the decision that would have been made.

Inverse Reinforcement Learning

$$\text{Good } S_t, A_t, S_{t+1}, A_{t+1} \text{ or } \pi \to R \to \pi^\star$$

Inverse Reinforcement Learning

- **Heuristic:** Learn a reward which **explains** the observed policy and used it to obtain a better policy (or to generalize to different models).
- No clear mathematical formulation:
  - Reward so that the observed policy is optimal (with a margin).
  - Expected return/optimal value function linked to observed policy (trajectories) probability (with entropic regularization)
  - Most generic formulation?

$$\min_{\pi'} \max_{R} \mathbb{E}_\pi[R] - \mathbb{E}_{\pi'}[\cancel{[}R] + K(\pi') - C(R)$$

- Exact problem considered not always clear for a given algorithm (and different from one algorithm to another)!

- Very hard problem!

Learning from Preferences

$$S_t, A_t, S_{t+1}, A_{t+1} \text{ vs } S_t, A'_t, S'_{t+1}, A'_{t+1} \to R \to \pi^\star$$

Learning from Preferences

- Often easier to compare trajectories than to make a demonstration.
- **Reinforcement Learning from Human Feedback**: Learn a reward from the demonstration using a preference model (Bradley-Terry?) and use it to find a policy.
- **Direct Policy Optimization**: shortcut to optimize directly the policy thanks to the explicit preference model used.
- Proximity constrains are often added to avoid moving too fast from a current policy.

- Key to the performances of current LLMs.

LLM: Large Language Model

Outline

More!

- Regrets
- Sample optimality
- Robustness
- Multi-agents (Games. . .)
- LLM and world models. . .

$\mathbb{E}_\pi[G_0]$ gmd

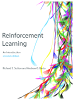
$\sum_{i=1}^{M} \mathbb{E}_{\pi_i}[G_0]$ gmd

$\sum_{i=1}^{M} \left[\mathbb{E}_{\pi_*}(G_0) - \mathbb{E}_{\pi_i}(G_0) \right]$

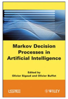
$\min_{P} \mathbb{E}_\pi^P(G_0)$ gmd

Outline

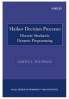
References



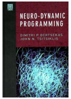
R. Sutton and A. Barto.
Reinforcement Learning, an Introduction (2nd ed.)
MIT Press, 2018



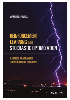
O. Sigaud and O. Buffet.
Markov Decision Processes in Artificial Intelligence.
Wiley, 2010



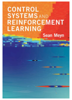
M. Puterman.
Markov Decision Processes. Discrete Stochastic Dynamic Programming.
Wiley, 2005



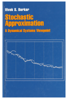
D. Bertsekas and J. Tsitsiklis.
Neuro-Dynamic Programming.
Athena Scientific, 1996



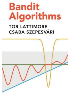
W. Powell.
Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions.
Wiley, 2022



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.