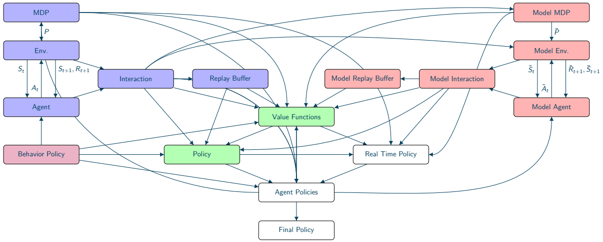
Reinforcement Learning
Operations Research: Prediction and Planning

E. Le Pennec

ÉCOLE POLYTECHNIQUE

M2DS - Reinforcement Learning – Fall 2023

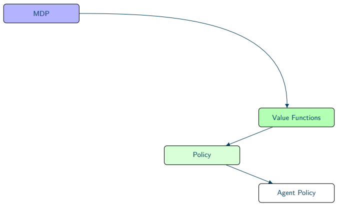
RL: What Are We Going To See?



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



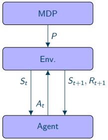
How to find the best policy knowing the MDP?

- 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.

Outline

Markov Decision Process / Operations Research



MDP / OR

- Known MDP model
- Focus on the finite horizon setting

$$G_t^T = \sum_{t'=t+1}^{T} R_{t'}$$

and the discounted setting:

$$G_t^\gamma = \sum_{t'=t+1}^{\infty} \gamma^{t'-(t+1)} R_{t'}$$

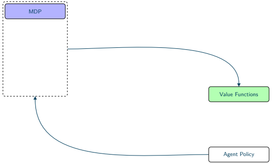
- We will later consider the other settings.

Policy

Policy

- Finite horizon : emphasis on Markovian policies
$$\Pi_t(A_t = a_t) = \pi_t(A_t = a_t | S_t = s_t) = \pi_t(a_t | s_t)$$
- Discounted return: emphasis on stationary Markovian policies
$$\Pi_t(A_t = a_t) = \pi(A_t = a_t | S_t = s_t) = \pi(a_t | s_t)$$

Prediction



Prediction

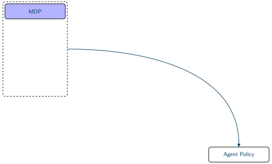
- How to efficently evaluate the quality of a policy

$$v_{t,\Pi}(s) = \mathbb{E}_\Pi\left[\sum_{t'=t+1}^{T} \gamma^{t'-(t+1)} R_{t'} \middle| S_t = s\right]$$

when we can ensure that the sum is finite?

- $v_{t,\Pi}$ independent of t in the discounted setting if the policy is stationary.

Planning



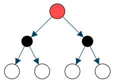
Policy

- How to find a policy π such that
$$\sum_{s,t} \mu(s,t) v_{t,\Pi}(s)$$
is as large as possible?
- Emphasis on $\mu(s,t) = 0$ if $t \neq 0$ and $\mu(s,0) = \mathbb{P}_0(S_0 = s_0)$.

Outline

Bellman Equation

$$
\begin{aligned}
v_{t,\Pi}(s) &= \sum_a \pi_t(a|s) \sum_{s',r} p(s',r|s,a)\left(r + \gamma v_{t+1,\Pi}(s')\right)\\
&= \sum_a \pi_t(a|s) r(s,a) + \gamma \sum_{s'} \sum_a p(s'|s,a) \pi_t(a|s) v_{t+1,\Pi}(s')
\end{aligned}
$$



Bellman Equation

- Link between $v_{t,\Pi}$ and $v_{t+1,\Pi}$.
- Straightforward consequence of

$$
G_t = \sum_{t'=t+1}^{T} \gamma^{t'-(t+1)} R_{t'} = R_{t+1} + \gamma \sum_{t'=t+2}^{T} \gamma^{t'-(t+2)} R_{t'} = R_{t+1} + \gamma G_{t+1}
$$

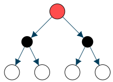
and thus

$$
\mathbb{E}[G_t|S_t = s] = \mathbb{E}[R_{t+1}|S_t = s] + \gamma \mathbb{E}[\mathbb{E}[G_{t+1}|S_{t+1}]|S_t = s]
$$

Bellman Operator

$$\mathcal{T}^{\pi_t} : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$$

$$\mathcal{T}^{\pi_t} v(s) = \underbrace{\sum_a \pi_t(a|s) r(s,a)}_{r_{\pi_t}(s)} + \gamma \sum_{s'} \underbrace{p(s'|s,a) \sum_a \pi_t(a|s)}_{P^{\pi_t}(s,s')} v(s')$$



Bellman Operator

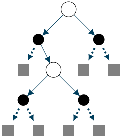
- Affine operator from the space of state value functions to the space of state value functions.
- By construction,
$$v_{t,\Pi} = \mathcal{T}^{\pi_t} v_{t+1,\Pi}$$
- r_{π_t} is the vector of average immediate rewards using policy π_t while P^{π_t} is the one step state transition matrix using policy π_t.

Outline

Finite Horizon: Naive Approach

$$v_{t,\Pi}^{T}(s) = \sum_{a_t, r_{t+1}, s_{t+1}, \cdots, r_T} \left(\sum_{t'=t+1}^{T} r_{t'} \right) \mathbb{P}_{\Pi}(A_t = a_t \ldots, R_T = r_T | S_t = s)$$

$$= \sum_{a_t, r_{t+1}, s_{t+1}, \cdots, r_T} \left(\sum_{t'=t+1}^{T} r_{t'} \right) \pi_t(a_t|s) \times \cdots \times p(s_T, r_T | s_{T-1}, a_{T-1})$$



Finite Horizon: Naive Approach

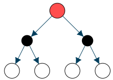
- Exhaustive exploration of the trajectories.
- Complexity of order $(|\mathcal{A}| \times |\mathcal{S}| \times |\mathcal{R}|)^{T-t}$ for the value function at time t.

- Complexity can be reduced to $(|\mathcal{A}| \times |\mathcal{S}|)^{T-t}$ by noticing that

$$v_{t,\Pi}^{T}(s) \quad = \sum_{a_t, s_{t+1}, \cdots, s_{t-1}, a_{t-1}} \left(\sum_{t'=t+1}^{T} r(s_t, a_t) \right) \pi_t(a_t|s) \times \cdots \times p(s_T | s_{T-1}, a_{T-1})$$

Finite Horizon: Recursive Prediction

$$v^T_{T,\Pi} = 0$$
$$v^T_{t-1,\Pi} = \mathcal{T}^{\pi_{t-1}} v^T_{t,\Pi}$$



Finite Horizon: Recursive Prediction

- After time T, the finite horizon return $G^T_t = 0$ hence $v^T_{T,\Pi} = 0$ whatever the policy.
- The Bellman equation yields second equation.
- Equivalent rewriting
$$v^T_{t-1,\Pi}(s) = r_{\pi_{t-1}}(s) + \sum_{s'} P_{\pi_{t-1}}(s,s') v^T_t$$
- Complexity of order only $T \times |\mathcal{S}|^2(|\mathcal{A}| + |\mathcal{S}|)$ to compute all the value functions.

Finite Horizon: Value Iteration

Finite Horizon: Prediction by Value Iteration

input: MDP model $\langle(\mathcal{S},\mathcal{A},\mathcal{R}),P\rangle$ and policy Π
parameter: Horizon T
init: $v_T^T(s)=0\,\forall\, s\in\mathcal{S},\ t=T$
repeat
  $t\leftarrow t-1$
  for $\forall\, s\in\mathcal{S}$ **do**

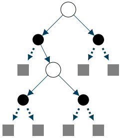
$$v_t^T(s)\leftarrow\sum_{a\in\mathcal{A}}\pi_t(a|s)\left(r(s,a)+\gamma\sum_{s'\in\mathcal{S}}p(s'|s,a)v_{t+1}^T(s')\right)$$

  end
until $t=0$
output: Value functions v_t^T

- Most classical formulation

Discounted: Naive Approach

$$v^{\gamma}_{t,\Pi}(s) = \sum_{t'=t+1}^{\infty} \gamma^{t'-(t+1)} \mathbb{E}_{\Pi}[R_{t'} | S_t = s] \simeq \sum_{t'=t+1}^{T} \gamma^{t} \mathbb{E}_{\Pi}[R_{t'} | S_t = s] = v^{\gamma,T}_{t,\Pi}(s)$$



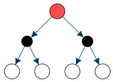
$$v^{\gamma,T}_{t,\Pi}(s) = \sum_{a_t, s_{t+1}, \cdots, s_{t-1}, a_{t-1}} \left(\sum_{t'=t+1}^{T} \gamma^{t'-(t+1)} r(s_t, a_t) \right) \pi_t(a_t | s) \times \cdots$$
$$\times\, p(s_T | s_{t-1}, a_{t-1})$$

Naive approach

- Exhaustive exploration of truncated trajectories.
- Back to the finite horizon setting...
- **Prop:** Control on the error as $\left| v^{\gamma}_{\Pi} - v^{\gamma,T}_{t,\Pi} \right|_{\infty} \leq \frac{\gamma^{T+1-t}}{1-\gamma} \max_{r \in \mathcal{R}} |r|$
- Relation between the error $\epsilon \simeq \gamma^{T-t}$ and the numerical complexity $C = (|\mathcal{A}| \times |\mathcal{S}|)^{T-t}$ of order $C \simeq \epsilon^{-1}$.

Discounted: Recursive Prediction with Naive Initialization

$$v^{\gamma}_{T,\Pi} \simeq v^{\gamma,T'}_{T,\Pi} = \tilde{v}_{T,\Pi}$$
$$v^{\gamma}_{t-1,\Pi} = \mathcal{T}^{\pi_{t-1}} v^{\gamma}_{t,\Pi} \simeq \tilde{v}_{t-1,\Pi} = \mathcal{T}^{\pi_{t-1}} \tilde{v}_{t,\Pi}$$



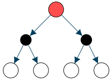
Recursive Prediction

- Requires an initialization at time T with a horizon T'.
- The Bellman equation yields the second equation.
- Complexity of order only $T \times |\mathcal{S}|^2(|\mathcal{A}| + |\mathcal{S}|)$ to compute all the value functions after the initialization of cost $(|\mathcal{A}| \times |\mathcal{S}|)^{T'-T}$.
- **Prop:** If the approximation error between $v^{\gamma}_{T,\Pi}$ and $v^{\gamma,T'}_{T,\Pi}$ is bounded by ϵ then
$$\|v^{\gamma}_{t,\Pi} - \tilde{v}_{t,\Pi}\|_\infty \leq \gamma^{T-t}\epsilon, \quad \forall t \leq T$$

Discounted and stationary: Bellman Equation

$$v_\Pi = \mathcal{T}^\pi v_\Pi$$

$$v_\Pi(s) = \sum_a \pi(a|s) r(s,a) + \gamma \sum_{s'} \sum_a p(s'|s,a) \pi(a|s) v_\Pi(s')$$



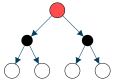
Bellman Equation

- Time independent value function v_Π.
- **Prop:** Unique solution of the linear equation $v_\Pi = \mathcal{T}^\pi v_\Pi$
- Complexity of order $(|A| + |S|) \times |S|^2$ to obtain the solution.

Discounted and stationary: Recursive Implementation

$$v_\Pi = \mathcal{T}^\pi v_\Pi$$

$$v_{k+1} = \mathcal{T}^\pi v_k \quad \text{with arbitrary } v_0$$



Bellman Iteration

- **Prop:** Unique fixed point of the Bellman operator $v \mapsto \mathcal{T}^\pi v$.
- **Prop:** The iterates $v_{k+1} = \mathcal{T}^\pi v_k$ converges toward v_Π and

$$\|v_k - v_\Pi\|_\infty \leq \gamma^k \|v_0 - v_\Pi\|_\infty$$

- Complexity of order $(k + |A|)|S|^2$ to obtain the kth iterate.
- Exponential decay of the error with respect to the complexity.

Bellman Operator and Contraction

$$\|\mathcal{T}^\pi v - \mathcal{T}^\pi v'\|_\infty \leq \gamma \|v - v'\|_\infty$$

Proof

- By definition
$$\|\mathcal{T}^\pi v - \mathcal{T}^\pi v'\|_\infty = \gamma \|P^\pi (v - v')\|_\infty$$
- It suffices then to notice that P^π is a transition matrix, so that
$$\sum_j P^\pi_{i,j} = 1$$
and thus $|\sum_j P^\pi_{i,j} z_j| \leq \max |z_j|$

Consequences

- Unicity of the solution of $\mathcal{T}^\pi v = v$.
- Linear decay γ^k of the error with the iterates.

Bellman Operator and Bellman Equation Solution

$$v_\Pi = \left(\sum_{k=0}^{\infty} \gamma^k \left(P^\pi\right)^k \right) r_\pi$$

A Closed Formula for the State Value Function

- $v_\Pi = \mathcal{T}^\pi v_\Pi \Leftrightarrow \left(I - \gamma P^\pi\right) v_\Pi = r_\pi$
- As P^π is a transition matrix, its eigenvalues are smaller than 1 and thus $\left(I - \gamma P^\pi\right)$ is invertible of inverse

$$\left(I - \gamma P^\pi\right)^{-1} = \sum_{k=0}^{\infty} \gamma^k \left(P^\pi\right)^k$$

- Could have been obtained without the Bellman equation as the $\left(\left(P^\pi\right)^k\right)_{s,s'}$ is, by construction, the probability of being at state s' at time k starting from s at time 0 and following Π.

Discounted and stationary: Value Iteration

Discounted: Prediction by Value Iteration

input: MDP model $\langle(\mathcal{S}, \mathcal{A}, \mathcal{R}), P\rangle$, discount factor γ, and stationary policy π
init: $\tilde{v}(s) \,\forall\, s \in \mathcal{S}$
repeat
  $\tilde{v}_{\text{prev}} \leftarrow \tilde{v}$
  for $s \in \mathcal{S}$ **do**

$$\tilde{v}(s) \leftarrow \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \tilde{v}_{\text{prev}}(s') \right)$$

  end
output: Value function $\tilde{v}$

- When to stop?

Discounted and stationary: Value Iteration

Discounted: Prediction by Value Iteration

input: MDP model $\langle(\mathcal{S},\mathcal{A},\mathcal{R}),P\rangle$, discount factor γ, and stationary policy π
parameter: $\delta > 0$ as accuracy termination threshold
init: $\tilde{v}(s)\,\forall\, s \in \mathcal{S}$
repeat
  $\tilde{v}_{\text{prev}} \leftarrow \tilde{v}$
  $\Delta \leftarrow 0$
  for $s \in \mathcal{S}$ **do**

$$\tilde{v}(s) \leftarrow \sum_{a\in\mathcal{A}} \pi(a|s)\left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\tilde{v}_{\text{prev}}(s')\right)$$

$$\Delta \leftarrow \max\left(\Delta, |\tilde{v}(s) - \tilde{v}_{\text{prev}}(s)|\right)$$

  end
until $\Delta < \delta$
output: Value function $\tilde{v}$

- **Prop:** when the algorithms stops

$$\|\tilde{v} - v_{\Pi}\|_\infty \leq \frac{2\delta}{1-\gamma}$$

Discounted and stationary: Value Iteration

Discounted: Prediction by Value Iteration - Gauss-Seidel Version

input: MDP model $\langle(\mathcal{S}, \mathcal{A}, \mathcal{R}), P\rangle$, discount factor γ, and stationary policy π
parameter: $\delta > 0$ as accuracy termination threshold
init: $\tilde{v}(s) \forall s \in \mathcal{S}$
repeat
- $\Delta \leftarrow 0$
- **for** $s \in \mathcal{S}$ **do**
  - $\tilde{v}_{\text{prev}} \leftarrow \tilde{v}(s)$
  - $$\tilde{v}(s) \leftarrow \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \tilde{v}(s') \right)$$
  - $\Delta \leftarrow \max(\Delta, |\tilde{v}(s) - \tilde{v}_{\text{prev}}|)$
- **end**

until $\Delta < \delta$
output: Value function $\tilde{v}$

- Gauss-Seidel variation mostly used in practice.
- No need to store the previous value function.

Outline

Optimal Policy

Optimal Policy

- An optimal policy $\Pi_\star$ should be better than any other policies:
$$\forall s, \forall t, v_{t,\Pi_\star}(s) = \sup_{\pi} v_{t,\Pi}(s)$$

Several Questions

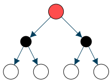
- Do this policy exists?
- Is it unique?
- How to characterize it?
- How to obtain it?

- Even the sup above could be an issue if it is not attained!

Explicit Recursive Solution

- After horizon T, any policy leads to a 0 return.
- At time $T-1$,
  - the total return G_T is the immediate return at time T and thus
$$v_{T,\Pi^\star}(s) = \sup_{\pi(a|s)} \sum_a \pi(a|s) r(a,s) = \sup_a r(a,s)$$
  - the optimal policy $\pi^\star_{T-1}$ exists and is determistic.
- By recursion,
  - the total return at time $t-1$ is the immediate return at time t plus the total return at time $t-1$ and thus
$$\begin{aligned} v_{t-1,\Pi^\star}(s) &= \sup_{\pi(a|s)} \sum_a \pi(a|s) \left(r(a,s) + \sum_{s'} p(s'|s,a) v_{t,\Pi^\star} \right) \\ &= \sup_a \left(r(a,s) + \sum_{s'} p(s'|s,a) v_{t,\Pi^\star} \right) \end{aligned}$$
  - the optimal policy $\pi^\star_{t-1}$ exists and is determistic.

Discounted Setting and Optimal Stationary Policy



Heuristic

- Optimal policy: $v^{\Pi^\star}(s) = \sup_\pi v_\Pi(s)$
- Stationary solution:

$$
\begin{aligned}
v_{\Pi^\star}(s) &= \sup_\pi \left(\mathcal{T}^\pi v_{\Pi^\star}\right)(s) \\
&= \sup_{\pi_t(\cdots|s)} \sum_a \pi(a|s) \left(r(a,s) + \gamma \sum_{s'} p(s'|s,a) v_{\Pi^\star}(s') \right) \\
&= \sup_a \left(r(a,s) + \gamma \sum_{s'} p(s'|s,a) v_{\Pi^\star}(s') \right)
\end{aligned}
$$

- Optimal deterministic policy: $\pi^\star(s) \in \operatorname{argmax}\left(r(a,s) + \gamma \sum_{s'} p(s'|s,a) v_{\Pi^\star}(s')\right)$.

- Is everything well defined? Yes but one has to be more cautious!

Optimal Value Function and Bellman Operator

Optimal Value Function

- Optimal value function: $v_\star(s) = \sup_\Pi v_\Pi(s)$
- Defined state by state so that it is not necessarily attained by a single $\Pi^\star$

Optimal Bellman operator

- Similar to the Bellman operator but do not depend on a policy:

$$\mathcal{T}^\star v(s) = \sup_a \left(r(a,s) + \gamma \sum_{s'} p(s'|s,a) v(s') \right)$$

Link between the two

- $v \geq \mathcal{T}^\star v$ implies $v \geq v_\star$.
- $v \leq \mathcal{T}^\star v$ implies $v \leq v_\star$.

Optimal Value Function and Bellman Operator

Bellman Operator and Fixed Point

- **Prop:** $\mathcal{T}^\star$ is a γ-contraction for the sup-norm and thus it exists a unique $v_{\star\star}$ such that $v_{\star\star} = \mathcal{T}^\star v_{\star\star}$.

Fixed Point and Optimal Value Function

- **Prop:** : $v_\star = v_{\star\star}$ and is thus the unique fixed point of $\mathcal{T}^\star$.
- **Proof:** $v_{\star\star} = \mathcal{T}^\star v_{\star\star}$ and thus $v_{\star\star} = v_\star$ according the link between the optimal value function and the Bellman operator.

- Does this mean something about policies?

Optimal Policy and Bellman Operator

Bellman Operator and Policy

- **Prop:** For any v, any policy π_v satisfying
$$\pi_v(s) \in \operatorname*{argmax}_a \left(r(a,s) + \gamma \sum_{s'} p(s'|s,a) v(s') \right)$$
is such that $\mathcal{T}^\star v(s) = \sup_\pi \mathcal{T}^\pi v(s) = \mathcal{T}^{\pi_v} v(s)$

Bellman Operator and Optimal Policy

- **Prop:** Any stationary policy $\pi_\star$ satisfying
$$\pi_\star(s) \in \operatorname*{argmax}_a \left(r(a,s) + \gamma \sum_{s'} p(s'|s,a) v^\star(s') \right)$$
is optimal.
- **Proof:** Indeed by construction, $\mathcal{T}^\star v_\star = \mathcal{T}^{\pi_\star} v_\star$ and thus, as $\mathcal{T}^\star v_\star = v_\star$, $v_{\pi_\star} = v_\star$.

Optimal Policy and Bellman Operator

Summary

- It exists a unique $v_\star$ such that $\mathcal{T}^\star v_\star = v_\star$
- $\forall s, v_\star(s) = \sup_\pi v_\pi(s)$
- Any policy $\pi_\star$ satisfying:

$$\forall s, \pi_\star(s) \in \operatorname*{argmax}_a \left(r(a,s) + \gamma \sum_{s'} p(s'|s,a) v^\star(s') \right)$$

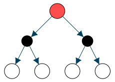
is optimal as $\forall s, v_{\pi_\star}(s) = v_\star(s) = \sup_\pi v_\pi(s)$

- Existence result but not (yet) a constructive algorithm!

Outline

Linear System and Linear Programming

$$v_\pi = \mathcal{T}^\pi v_\pi \qquad v_\star = \mathcal{T}^\star v_\star$$



Explicit Resolution of the Equations?

- Prediction:
  - Simple linear system for v_π.
  - Already mentionned before...
  - Complexity of order $(|A| + |S|)|S|^2$.
- Planning:
  - More complex linear programming system for $v_\star$ due to the max operator.
  - Optimal policy easily deduced from $v_\star$.
  - Complexity of order $(|A||S|)^3$.

Linear Programming

$$\text{From } \forall s, v(s) = \sup_a r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')$$

$$\text{to } \min_v \sum_s \mu(s) v(s)$$

$$\text{such that } \forall (s,a), v(s) \geq r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')$$

Different formulations but same solution

- Using $v \geq \mathcal{T}^\star v \Leftrightarrow v \geq v_\star$, the condition implies $v \geq v_\star$
- Now for any μ satisfying $\mu(s) > 0$, $\sum_s \mu(s) v(s) \geq \sum_s \mu(s) v_\star(s)$ as soon as the condition is satisfied, hence $v_\star$ is a solution.
- If for any state $v(s) > v_\star(s)$ then $\sum_s \mu(s) v(s) > \sum_s \mu(s) v_\star(s)$ and thus $v_\star$ is the unique minimizer.

Primal Problem

$$\text{Primal: } \min_v \sum_s \mu(s) v(s)$$
$$\text{such that } \forall (s,a), v(s) \geq r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')$$

Some properties

- Can be solved with a linear programming solver.
- Unicity of solution (and thus independence with respect to μ) can be proved without using $v_\star$.
  - **Proof:** let v_1 a solution for μ_1 and v_2 a solution for μ_2 then $\min(v_1, v_2)$ satifies the constraints. Furthermore if exists $v_2(s) < v_1(s)$ then $\min(v_1, v_2)$ is a strictly better solution for μ_2 which is impossible.

Dual Problem

Primal: $$\min_v \sum_s \mu(s) v(s)$$

$$\text{such that } \forall (s,a), v(s) \geq r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')$$

Dual: $$\max_{\lambda(s,a)\geq 0} \sum_{s,a} \lambda(s,a) r(s,a)$$

$$\text{such that } \forall s, \sum_a \lambda(s,a) = \mu(s) + \gamma \sum_{s',a} p(s|s',a) \lambda(s',a)$$

Derivation

- Usual derivation through the Lagrangian:

$$\mathcal{L}(v,\lambda) = \sum_s \mu(s) v(s) + \sum_{s,a} \lambda(s,a) \left(r(s,a) + \gamma \sum_{s',a} p(s|s',a) v(s') - v(s) \right)$$

- Strong duality as Slater condition holds when $\gamma < 1$ with $v = \frac{1+\epsilon}{1-\gamma} \max_{s,a} r(s,a)$.

Dual and Interpretation

$$\text{Dual:} \quad \max_{\lambda(s,a)\geq 0} \sum_{s,a} \lambda(s,a) r(s,a)$$

$$\text{such that } \forall s, \sum_{a} \lambda(s,a) = \mu(s) + \gamma \sum_{s',a} p(s|s',a)\lambda(s',a)$$

$$\text{Interpretation}: \max_{\pi} \sum_{k=0}^{\infty} \gamma^k \sum_{s,a} \mathbb{P}(S_t = a, A_t = a | S_0 \sim \mu, \pi)\, r(s,a)$$

Interpretation in terms of policy

- For any feasible λ, define $u(s) = \sum_a \lambda(s,a)$ and the policy $\pi(a|s) = \lambda(s,a)/u(s)$.
- **Prop:** $u = (\mathrm{Id} - \gamma P^{\pi})\mu = \sum_{k=0}^{\infty} \gamma^k \left(P^{\pi}\right)^k \mu$.
- **Prop:** $\lambda(s,a) = \pi(a|s)u(s) = \sum_{k=0}^{\infty} \gamma^k \mathbb{P}(S_t = a, A_t = a | S_0 \sim \mu, \pi)$
- Conversely for any π they is a feasible λ.
- Any optimal $\lambda_\star$ (and thus policy) satisfies $\lambda_\star(s,a) = 0$ if $v_\star(s) > r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\star(s')$ (optimal policy support)

Outline

Finite Horizon

Finite Horizon: Planning by Value Iteration

input: MDP model $\langle(\mathcal{S}, \mathcal{A}, \mathcal{R}), P\rangle$
parameter: Horizon T
init: $v_T^T(s) = 0 \, \forall \, s \in \mathcal{S}, \; t = T$
repeat

  $t \leftarrow t - 1$
  for $s \in \mathcal{S}$ **do**

$$v_t^T(s) \leftarrow \max_{a \in \mathcal{A}} \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v_{t+1}^T(s') \right)$$

  end

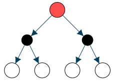
until $t = 0$

output: Deterministic policy $\pi_t(s) \in \operatorname{argmax}_{a \in \mathcal{A}} \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) v_{t+1}^T(s') \right)$

- Algorithm used to prove the existence of an optimal policy.
- No necessarily unique as argmax may not be unique.

Optimal Value Function, Fixed Point and Contraction

$$v_\star = \mathcal{T}^\star v_\star \quad \text{and} \quad \|\mathcal{T}^\star v - \mathcal{T}^\star v'\|_\infty \leq \gamma \|v - v'\|_\infty$$
$$\implies v_{k+1} = \mathcal{T}^\star v_k \to v_\star$$



Bellman Operator

- Properties of Optimal Bellman Operator:
  - $v_\star$ is a fixed point of $\mathcal{T}^\star$.
  - $\mathcal{T}^\star$ is a γ-contraction for the $\|\cdot\|_\infty$ norm.
- Classical fixed point theorem setting.
- Practical algorithm to approximate $v_\star$.

Value Iteration Algorithm

Discounted: Value Iteration Planning

input: MDP model $\langle(\mathcal{S}, \mathcal{A}, \mathcal{R}), P\rangle$, and discount factor γ
parameter: $\delta > 0$ as accuracy termination threshold
init: $\tilde{v}(s) \, \forall \, s \in \mathcal{S}$
repeat
  $\tilde{v}_{\text{prev}} \leftarrow \tilde{v}$
  $\Delta \leftarrow 0$
  for $s \in \mathcal{S}$ **do**
    $$\tilde{v}(s) \leftarrow \max_{a \in \mathcal{A}} r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \tilde{v}_{\text{prev}}(s')$$
    $\Delta \leftarrow \max\left(\Delta, |\tilde{v}(s) - \tilde{v}_{\text{prev}}(s)|\right)$
  end
until $\Delta < \delta$
output: Value function $\tilde{v}$

- Same convergence criterion (and similar proof) than in the planning case.
- Which policy?

Value Iteration Algorithm

Discounted: Value Iteration Planning

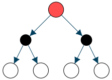
input: MDP model $\langle(\mathcal{S},\mathcal{A},\mathcal{R}),P\rangle$, and discount factor γ
parameter: $\delta > 0$ as accuracy termination threshold
init: $\tilde{v}(s)\,\forall\, s\in\mathcal{S}$
repeat
  $\tilde{v}_{\text{prev}} \leftarrow \tilde{v}$
  $\Delta \leftarrow 0$
  for $s\in\mathcal{S}$ **do**
    $$\tilde{v}(s) \leftarrow \max_{a\in\mathcal{A}} r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\tilde{v}_{\text{prev}}(s')$$
    $$\Delta \leftarrow \max\left(\Delta, |\tilde{v}(s) - \tilde{v}_{\text{prev}}(s)|\right)$$
  end
until $\Delta < \delta$
output: Deterministic policy $\tilde{\pi}(s) \in \operatorname{argmax}_a r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\tilde{v}(s')$

- Natural idea: define a policy using the argmax of the existence proof.
- Do we have a convergence guarantee on the resulting policy?

Value and argmax Policy

$$\tilde{\pi}(s) \in \operatorname*{argmax}_a r(s,a) + \gamma \sum_{s'} p(s'|s,a)\tilde{v}(s')$$

$$\implies \|v_{\tilde{\pi}} - v_\star\|_\infty \leq \frac{2\gamma}{1-\gamma}\|\tilde{v} - v_\star\|_\infty$$



Value and argmax Policy

- Bound on the loss of the final policy!
- Rely on the fact that, by construction, $\mathcal{T}^{\tilde{\pi}}\tilde{v} = \mathcal{T}^\star\tilde{v}$
- **Proof:**

$$\begin{aligned}
\|v_{\tilde{\pi}} - v_\star\|_\infty &= \|\mathcal{T}^{\tilde{\pi}}v_{\tilde{\pi}} - \mathcal{T}^{\tilde{\pi}}\tilde{v} + \mathcal{T}^\star\tilde{v} - \mathcal{T}^\star v_\star\|_\infty \\
&\leq \|\mathcal{T}^{\tilde{\pi}}v_{\tilde{\pi}} - \mathcal{T}^{\tilde{\pi}}\tilde{v}\|_\infty + \|\mathcal{T}^\star\tilde{v} - \mathcal{T}^\star v_\star\|_\infty \\
&\leq \gamma\|v_{\tilde{\pi}} - \tilde{v}\|_\infty + \gamma\|\tilde{v} - v_\star\|_\infty \\
&\leq \gamma\|v_{\tilde{\pi}} - v_\star\|_\infty + 2\gamma\|\tilde{v} - v_\star\|_\infty
\end{aligned}$$

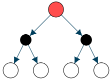
Value Iteration Algorithm

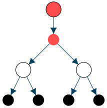
Discounted: Value Iteration Planning

input: MDP model $\langle(\mathcal{S}, \mathcal{A}, \mathcal{R}), P\rangle$, and discount factor γ
parameter: $\delta > 0$ as accuracy termination threshold
init: $\tilde{v}(s) \, \forall \, s \in \mathcal{S}$
repeat
  $\tilde{v}_{\text{prev}} \leftarrow \tilde{v}$
  $\Delta \leftarrow 0$
  for $s \in \mathcal{S}$ **do**
    $$\tilde{v}(s) \leftarrow \max_{a \in \mathcal{A}} r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \tilde{v}_{\text{prev}}(s')$$
    $$\Delta \leftarrow \max\left(\Delta, |\tilde{v}(s) - \tilde{v}_{\text{prev}}(s)|\right)$$
  end
until $\Delta < \delta$
output: Deterministic policy $\tilde{\pi}(s) \in \operatorname{argmax}_a r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) \tilde{v}(s')$

- **Prop:** $\|v_{\tilde{\pi}} - v_\star\|_\infty \leq \dfrac{4\gamma\delta}{1-\gamma}$

From State Value to State-Action Value Functions





$$v_\pi(s) = \mathbb{E}_\pi\left[\sum_k \gamma^k R_t | S_0 = s\right] \qquad q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_k \gamma^k R_t | S_0 = s, A_0 = a\right]$$

$$\mathcal{T}^\pi v(s) = \sum_a \pi(a|s)\left(r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s')\right) \qquad \mathcal{T}^\pi q(s,a) = r(s,a) + \sum_{s'} p(s'|s,a) \sum_a \pi(a|s') q(s',a)$$

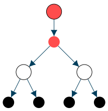
$$\mathcal{T}^\star v(s) = \max_a r(s,a) + \gamma \sum_{s'} p(s'|s,a) v(s') \qquad \mathcal{T}^\star q(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_a q(s',a)$$

Two equivalent point of view?

- Everything could have been defined using the state-action point of view.
- Knowing v_π is equivalent to knowing q_π as

$$v_\pi(s) = \sum_a \pi(s|a) q_\pi(s,a) \quad \text{and} \quad q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_\pi(s').$$

State-Action Bellman Operators



$$\mathcal{T}^\pi q(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_a \pi(a|s') q(s',a)$$

$$\mathcal{T}^\star q(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_a q(s',a)$$

Properties

- **Prop:** $\mathcal{T}^\pi$ and $\mathcal{T}^\star$ are γ contractions for the $\|\cdot\|_\infty$ norm.
- **Prop:** q_π is the unique solution of $\mathcal{T}^\pi q = q$
- **Prop:** $q_\star$ defined $q_\star(s,a) = \sup_\pi q_\pi(s,a)$ is the unique solution of $q = \mathcal{T}^\star q$ and is attained for any policy $\pi_\star$ satisfying $\pi_\star(s) \in \operatorname{argmax} q_\star(s,a)$.
- **Prop:** Any such policy satisfies: $v_{\pi_\star}(s) = q_{\pi_\star}(s, \pi_\star(s)) = v_\star(s)$.

State-Action Value Iteration Algorithm

Discounted: Planning by State-Action Value Iteration

input: MDP model $\langle(\mathcal{S},\mathcal{A},\mathcal{R}),P\rangle$, and discount factor γ
parameter: $\delta > 0$ as accuracy termination threshold
init: $\tilde{q}(s,a)\,\forall\,(s,a)\in\mathcal{S}\times\mathcal{A}$
repeat
  $\tilde{q}_{\text{prev}} \leftarrow \tilde{q}$
  $\Delta \leftarrow 0$
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**

$$\tilde{q}(s,a) \leftarrow \left(r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\max_{a'}\tilde{q}_{\text{prev}}(s',a')\right)$$

      $\Delta \leftarrow \max\left(\Delta, |\tilde{q}(s,a) - \tilde{q}_{\text{prev}}(s,a)|\right)$
    end
  end
until $\Delta < \delta$
output: Deterministic policy $\tilde{\pi}(s) \in \operatorname{argmax}_a \tilde{q}(s,a)$

- Same complexity but more storage than with state value function...
- but will be useful later!

Outline

Value Fonction vs Policy Point of View

$$v, q \longrightarrow \Pi \quad \text{or} \quad \Pi \longrightarrow v, q?$$

Planning

- Focus so far on value-fonction point of view!
- Heuristic: find a good approximation of the optimal value function and deduce a good policy.
- Can we work directly on the policy itself?

- For prediction, only the policy point of view makes sense!

Toward Policy Improvement

$$\forall s, \pi_+(s) \in \operatorname*{argmax}_a q_\pi(s, a) \implies \forall v_{\pi_+}(s) \geq v_\pi(s)$$

Classical Policy Improvement Lemma

- **Prop:** Given a policy π and its q value-function, one can obtain a better policy with the argmax operator.
- **Prop:** If no improvement is possible, it means that π is already optimal.
- **Proof:** Use $\mathcal{T}^{\pi_+} v_\pi = \mathcal{T}^\star v_\pi \geq \mathcal{T}^\pi v_\pi = v_\pi$ to prove $\left(\mathcal{T}^{\pi_+}\right)^k v_\pi \geq v_\pi$ which implies the result by letting k goes to $+\infty$.

- Leads to a sequential improvement algorith...

Policy Improvement Lemma

$$\mathbb{E}[v_{\pi'}(S_0)] - \mathbb{E}[v_\pi(S_0)] = \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{\pi'}\left[\sum_a \pi'(a|S_t)\left(q_\pi(S_t,a) - v_\pi(S_t)\right)\right]$$
$$= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}_{\pi'}\left[\sum_a \left(\pi'(a|S_t) - \pi(a|S_t)\right) q_\pi(S_t,a)\right]$$

A Generic Improvement Lemma

- No assumptions on π and π'!
- Easy proof.
- Imply the previous lemma as $\max_a Q_\pi(s,a) - v_\pi(s) \geq 0$.
- Show that improvement choices are possible.

- Will prove to be useful later...

Policy Iteration

Discounted: Planning by Policy Iteration

input: MDP model $\langle(\mathcal{S},\mathcal{A},\mathcal{R}),P\rangle$, and discount factor γ
parameter: Initial policy $\tilde{\pi}$
repeat
  Compute $q_{\tilde{\pi}}$.
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $\tilde{pol}(s) \leftarrow \mathrm{argmax}\, q_{\tilde{\pi}}(s,a)$
    end
  end
output: Deterministic policy $\tilde{\pi}$.

Some issues

- How to obtain q_π?
- When to stop?

Policy Iteration

Discounted: Planning by Policy Iteration

input: MDP model $\langle(\mathcal{S}, \mathcal{A}, \mathcal{R}), P\rangle$, and discount factor γ
parameter: Initial policy $\tilde{\pi}$
repeat
    $stable \leftarrow 0$
    Compute $q_{\tilde{\pi}}$.
    for $s \in \mathcal{S}$ **do**
        $old - action \leftarrow \tilde{\pi}(s)$
        $\tilde{\pi}(s) \leftarrow \text{argmax}\, q_{\tilde{\pi}}(s, a)$
        if $\tilde{\pi}(s) \neq old - action$ **then**
            $stable \leftarrow 0$
        end
    end
until $stable == 1$
output: Deterministic policy $\tilde{\pi}$.

Finite Setting

- Finite set of action-states implies a finite set of policy.
- Convergence of the algorithm in finite time!

Policy Iteration

Convergence Rate

- Crude analysis:
  - Bound after k steps of the algorithm
$$\|v_{\pi_k} - v_\star\|_\infty \leq \gamma \|v_{\pi_{k-1}} - v_\star\|_\infty \leq \gamma^k \|v_{\pi_0} - v_\star\|_\infty$$
$$\|v_{\pi_k} - v_\star\|_\infty \leq \frac{\gamma}{1-\gamma} \|v_{\pi_k} - v_{\pi_{k-1}}\|_\infty$$
  - Not much better than value iteration but much higher complexity as q_{π_k} is obtained by solving the Bellman equation!
- Much faster in practice...
- Clever analysis (Putterman):
  - Under some mild assumptions and provided $\|P^{\pi_k} - P^\star\| \leq K \|v_{\pi_k} - v_\star\|_\infty$ then
$$\|v_{\pi_k} - v_\star\|_\infty \leq \frac{K\gamma}{1-\gamma} \|v_{\pi_{k-1}} - v_\star\|_\infty^2$$
  - May explain the better convergence in practice!

Outline

Value Iteration: (Relaxed) First Order Method

Value Iteration

- Iteration:

$$\begin{aligned} v_k &= \mathcal{T}^\star v_{k-1} \\ &= v_{k-1} + (\mathcal{T}^\star - \mathrm{Id})\, v_{k-1} \end{aligned}$$

- Relaxation

$$v_k = v_{k-1} - \alpha \left(\mathrm{Id} - \mathcal{T}^\star\right) v_{k-1}$$

  can be proved to converge for any $\alpha < \frac{2}{1+\gamma}$.
- Can be interpreted as a first order method with pseudo-gradient $(\mathcal{T}^\star - \mathrm{Id})\, v_{k-1}$.
- No function corresponding to this gradient!

- Is there a better choice for α than $\alpha = 1$?
- No as the resulting operator is a contraction of constant

$$|1 - \alpha| + \alpha\gamma \geq \gamma$$

Policy Iteration: Newton-Raphson Method

Policy Iteration

- Explicit iteration:

$$\text{Solve } v_{\pi_{k-1}} = \mathcal{T}^{\pi_k} v_{\pi_{k-1}}$$
$$\text{Let } \pi_k \text{ such that } \mathcal{T}^{\pi_k} v_{\pi_{k-1}} = \mathcal{T}^\star v_{\pi_{k-1}}$$

- Implicit iteration on v_{π_k}:

$$\begin{aligned} v_{\pi_k} &= (\mathrm{Id} - \gamma P^{\pi_k})^{-1} r_{\pi_k} \\ &= (\mathrm{Id} - \gamma P^{\pi_k})^{-1} \left(r_{\pi_k} + (\gamma P^{\pi_k} - \mathrm{Id}) v_{\pi_{k-1}} + (\mathrm{Id} - \gamma P^{\pi_k}) v_{\pi_{k-1}} \right) \\ &= v_{\pi_{k-1}} - (\mathrm{Id} - \gamma P^{\pi_k})^{-1} (\mathrm{Id} - \mathcal{T}^{\pi_k}) v_{\pi_{k-1}} \end{aligned}$$

- Can be interpreted as a second order method with pseudo-gradient $(\mathrm{Id} - \mathcal{T}^{\pi_k}) v_{\pi_{k-1}} = (\mathrm{Id} - \mathcal{T}^\star) v_{\pi_{k-1}}$ and pseudo-Hessian $(\mathrm{Id} - \gamma P^{\pi_k})$.

- Not a formal analysis but give a good insight on the better convergence of policy iteration.

Outline

Stability of Value and Policy Iteration

Ideal Value and Policy Iteration?

- Iterative algorithms.
- Convergence proofs assume perfect computation.
- What happens if we make a (small) error at each step?

- Particularly important for Policy Iteration in which one resolves a linear system at each step!

Value Iteration Stability

$$v_k = \mathcal{T}^\star v_{k-1} + \epsilon_{k-1}$$

$$\implies \|v_k - v_\star\|_\infty \leq \gamma^k \|v_0 - v_\star\|_\infty + \frac{\max_{0 \leq k' < k} \|\epsilon_{k'}\|_\infty}{1-\gamma}$$

$$\implies \|v_{\pi_k} - v_\star\|_\infty \leq \frac{2\gamma^{k+1}}{1-\gamma} \|v_0 - v_\star\|_\infty + \frac{2\gamma \max_{0 \leq k' < k} \|\epsilon_{k'}\|_\infty}{(1-\gamma)^2}$$

Stability with respect to the error

- Proof relies on the contraction property of $\mathcal{T}^\star$ (hence similar results for $\mathcal{T}^\pi$).
- Error term $\frac{\max_{0 \leq k' < k} \|\epsilon_{k'}\|_\infty}{1-\gamma}$ can be replaced by $\sum_{k'=0}^{k-1} \gamma^{k-k'} \|\epsilon_{k'}\|_\infty$
- Convergence if $\|\epsilon_k\|_\infty$ tends to 0.
- Remains in a neighborhood of the optimal solution if $\|\epsilon_k\|_\infty$ is bounded.

Policy Iteration

$$v_{k-1} = v_{\pi_{k-1}} + \epsilon_{k-1} \quad \text{and} \quad \mathcal{T}^{\pi_k} v_{k-1} = \mathcal{T}^\star v_{k-1}$$

$$\implies \|v_{\pi_k} - v_\star\|_\infty \leq \gamma^k \|v_{\pi_0} - v_\star\|_\infty + \frac{\gamma(2-\gamma) \max_{0\leq k'<k} \|\epsilon_{k'}\|_\infty}{(1-\gamma)^2}$$

Stability with respect to the error

- Quite involved proof but crude results.
- Error term $\frac{\max_{0\leq k'<k} \|\epsilon_{k'}\|_\infty}{1-\gamma}$ can be replaced by $\sum_{k'=0}^{k-1} \gamma^{k-k'} \|\epsilon_{k'}\|_\infty$
- Convergence if $\|\epsilon_k\|_\infty$ tends to 0.
- Remains in a neighborhood of the optimal solution if $\|\epsilon_k\|_\infty$ is bounded.

- Policy Iteration only requires an approximate estimate of $v_{\pi_{k-1}}$, for instance obtained by Bellman iteration...

Outline

Modified Policy Iteration

Discounted: Planning by Generalized Policy Iteration

input: MDP model $\langle(\mathcal{S},\mathcal{A},\mathcal{R}),P\rangle$, and discount factor γ
parameter: Initial q
repeat
  for $s \in \mathcal{S}$ **do**
    $\tilde{\pi}(s) \leftarrow \operatorname{argmax}_a q(s,a)$
  end
  repeat
    $q_{\text{prev}} \rightarrow q$
    for $(s,a) \in \mathcal{S} \times \mathcal{A}$ **do**
      $$q(s,a) \leftarrow r(s,a) + \gamma \sum_{s,a'} p(s'|s,a)\tilde{\pi}(a'|s)q_{\text{prev}}(s,a)$$
    end
output: Deterministic policy $\tilde{\pi}$.

- Algorithm driven by q.
- Flexibility in the number of prediction steps after each policy improvement steps.
- Special cases:
    - Large number: Policy Iteration with (small) error.
    - One: Value Iteration!

MPI Analysis

$$\mathcal{T}^{\pi_k} v_k = \mathcal{T}^\star v_k \quad \text{and} \quad v_{k+1} = \left(\mathcal{T}^{\pi_k}\right)^{m_k} v_k$$

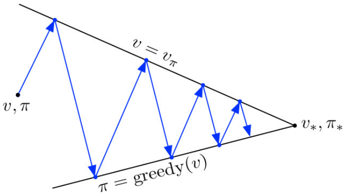
$$\Longrightarrow \|v_{k+1} - v_\star\|_\infty \leq \gamma \left(\frac{1-\gamma^{m_k}}{1-\gamma} \|P^{\pi_k} - P^\star\| + \gamma^{m_k} \right) \|v_k - v_\star\|_\infty$$

Convergence Results

- Quite technical proof.
- Valid only under the mild assumption $\mathcal{T}^\star v_0 \geq v_0$.
- Very fast decay provided $\|P^{\pi_k} - P^\star\|$ is small.

- No stability with arbitrary errors...

Generalized Policy Iteration



Source: Sutton and Barto

General Policy Iteration

- Two simultaneous interacting processes:
  - One forcing the policy to correspond to the current value function (Policy Improvement)
  - One trying to male the current value function coherent with the current policy (Policy Evaluation)
- Several variations possible on the two processes.

- In GPI, the policy is driven by the value function.
- Typically, stabilizes only if one reaches the optimal value/policy pair.

State Update Order

Discounted: Prediction by Value Iteration - State Update Order

input: MDP model $\langle(\mathcal{S}, \mathcal{A}, \mathcal{R}), P\rangle$, discount factor γ, and stationary policy π
init: $\tilde{v}(s) \, \forall \, s \in \mathcal{S}$
repeat

$\tilde{v}_{\text{prev}} \leftarrow \tilde{v}$
for $s \in \mathcal{S}' \subset \mathcal{S}$ **do**

$$\tilde{v}(s) \leftarrow \sum_{a \in \mathcal{A}} \pi(a|s) \left(r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) \tilde{v}_{\text{prev}}(s') \right)$$

end

output: Value function $\tilde{v}$

Classical strategies

- $\mathcal{S}' = \mathcal{S}$: classical iteration
- $\mathcal{S}' = \{s\}$: Gauss-Seidel
- $\mathcal{S}' = \{s, |\mathcal{T}^\pi \tilde{v}(s) - \tilde{v}(s)| > \epsilon\}$: Prioritized sweeping

- Converges provided all states are visited infinitely often...
- Gain in term of storage or focus on most interesting states...

Policy Improvement Variation

$$\text{Greedy}: \pi(s) \in \operatorname*{argmax}_a q(s,a) \iff \pi(\cdot|s) \in \operatorname*{argmax}_{\tilde{\pi}} \sum_a \tilde{\pi}(a) q(s,a)$$

$$\text{Restricted}: \pi(\cdot|s) \in \operatorname*{argmax}_{\tilde{\pi} \in \tilde{\Pi}_\epsilon} \sum_a \tilde{\pi}(a) q(s,a)$$

$$\text{Regularized}: \pi(\cdot|s) \in \operatorname*{argmax}_{\tilde{\pi}} \sum_a \tilde{\pi}(a) q(s,a) + \epsilon P(\tilde{\pi})$$

Classical Variations

- ϵ-greedy: Restrict $\tilde{\pi}$ to the set of policy s.t. $\tilde{\pi}(a) \geq \epsilon$
  - Explicit solution: $\pi(a|s) = \epsilon + (1-\epsilon) \operatorname{argmax} q(s,a)$
  - Policy improvement property if ϵ decreases.
- Soft-max: Regularize by $\epsilon H(\tilde{\pi})$ where H is the entropy.
  - Explicit solution: $\pi(a|s) \propto \exp(q(s,a)/\epsilon)$
  - No classical policy improvement...

- Tends to greedy when ϵ goes to 0.
- Turn out to be interesting later...

Outline

Episodic Setting

$$\mathbb{E}_\pi\left[\min_t\{t, \forall t' \geq t,\ R_{t'} = 0\}\right] < H \Rightarrow \|\mathcal{T}v - \mathcal{T}v'\|_\xi \leq \frac{H-1}{H}\|v - v'\|_\xi$$

Proper Policy

- A policy π is said to be H-proper if $\mathbb{E}_\pi\left[\min_t\{t, \forall t' \geq t,\ R_{t'} = 0\}\right] \leq H < \infty$
- $\Leftrightarrow$ average duration of an episode using this policy less than a finite horizon H!

Bellman operators

- If a policy π is H-proper, the Bellman operator $\mathcal{T}^\pi$ is a $(H-1)/H$- contraction for a weighted sup-norm.
- If all the policies are H-propers, the optimal Bellman operator $\mathcal{T}^\star$ is a $(H-1)/H$-contraction for a weighted sup-norm.

- Under those strong assumptions, episodic setting $\simeq$ discounted setting with $\gamma = (H-1)/H$.
- Some results can be obtained under the much milder assumption that there is one proper policy and that any non-proper policy has at least one state for which $v_\pi(s) = -\infty$.

Episodic = Discounted?

Infinite Setting

- No issue with the rewards, as only the expectation is used.
- All the theory remains valid if the states are countable, but there is an issue in the algorithms, as we need to store/update an infinite number of states.
- The proof of existence of an optimal policy requires the max to be attained, which cannot be ensured in an infinite (even countable setting).

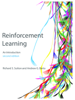
Some results...

- **Thm:** If S is countable, there exists an ϵ-optimal (stationary) policy for any $\epsilon > 0$.
- **Thm:** If S is a Polish space (completely metrizable topological space),
  - there exists a (P, ϵ)-optimal (stationary policy) for any $\epsilon > 0$.
  - if each A_s is countable, there exists an ϵ-optimal (stationary) policy for any $\epsilon > 0$.
  - if each A_s is finite, there exists an optimal (stationary) policy.
  - if each A_s is a compact metric space, $r(s, a)$ is a bounded u.s.c. function on A_s and $p(B|s, a)$ is continuous in a for each Borel subset B and any s, there exists an optimal (stationary) policy.

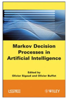
- **Mainly technical difficulties...**

Outline

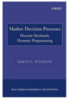
References



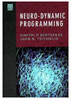
R. Sutton and A. Barto.
Reinforcement Learning, an Introduction (2nd ed.)
MIT Press, 2018



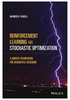
O. Sigaud and O. Buffet.
Markov Decision Processes in Artificial Intelligence.
Wiley, 2010



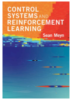
M. Puterman.
Markov Decision Processes. Discrete Stochastic Dynamic Programming.
Wiley, 2005



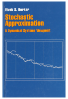
D. Bertsekas and J. Tsitsiklis.
Neuro-Dynamic Programming.
Athena Scientific, 1996



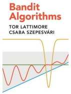
W. Powell.
Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions.
Wiley, 2022



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.