Mathematical Foundations of Reinforcement Learning

Mathematical Foundations of Reinforcement Learning

Richting	Informatica,Wiskunde
Jaar	MWIS, MINF

Examenvragen

2022-2023 : Januari

Oefeningen

1. (5pt) Let $T=50T=50$, and consider the following MDP (see image). Construct a stationary MDP with the same states and at most 2 actions such that all values are preserved.
2. (15pt)
    1. What is the period of the MDP in question 1?
    2. Give an example of a policy such that the resulting Markov chain has at least 2 different stationary distributions.
3. (10pt) Give an example of a stationary MDP such that the value iteration algorithm for discounted rewards converges in finitely many steps and one where it converges in infinitely many steps.
4. (10pt) Give an example of a statonary MDP such that from state ii, no optimal unichain policy exists for the infinite-horizon expected average reward criterion.
5. (5 pt) For the Q-learning algorithm we defined learning rates $\gamma 0=0,\gamma 1,\gamma 2,\ldots \gamma 0=0,\gamma 1,\gamma 2,\ldots$ with $0\leq\gamma k<1,\forall k\in N0\leq\gamma k<1,\forall k\in N$. Let $NiaNia$ be the map that maps nn to the timestep at which the state-action pair $(a,i)(a,i)$ is visited for the nn-th time. We required the following conditions: $\sum+\infty n=1\gamma Nia(n)=+\infty\sum n=1+\infty\gamma Nia(n)=+\infty$ and $\sum+\infty n=1\gamma 2Nia(n)<+\infty\sum n=1+\infty\gamma Nia(n)2<+\infty$ Give an example of learning rates that satisfy all the given conditions.

Theorie

1. (10pt) Let MM be a stationary communicating MDP. Prove that there is an optimal unichain policy from i for the infinite-horizon expected limit-average reward criterion.

2. (10pt)
    1. Consider the first condition of question 5 of the exercises. Argue that this implies that every state-action pair is seen infinitely often with probability 1.
    2. Give an example of a communicating MDP, initial values $Q(0)Q(0)$ such that even under the assumptions, some state-action pair may be visited finitely often with probability 0.
3. (15pt) We will now focus on the two mappings $U:RN \to RNU:RN \to RN$ and $L\pi:RN \to RNL\pi:RN \to RN$ defined for any randomized decision rule $\pi\pi$ as follows. $\{Ux\}i=\max a\in A(i)\{ri(a)+\alpha\sum j\in Spij(a)xj\}\{Ux\}i=\max a\in A(i)\{ri(a)+\alpha\sum j\in Spij(a)xj\}$. And $L\pi x=r(\pi)+\alpha P(\pi)xL\pi x=r(\pi)+\alpha P(\pi)x$. Prove the following result. (You must recall the definition of monotone contraction mapping. Also, you may use results proved during the course as auxiliary lemmas. "The mappings UU and $L\pi L\pi$ for any randomized decision rule $\pi\pi$ are monotone contraction mappings with contraction factor $\alpha\alpha$. Moreover, $v\alpha(\pi\infty)v\alpha(\pi\infty)$ is the unique fixed point of $L\pi x=xL\pi x=x$."
4. (10pt) For the policy iteration algorithm for infinite-horizon discounted rewards, we defined the action set $A(i,f)A(i,f)$ as follows:
$A(i,f):=\{a\in A(i)|ri(a)+\alpha\sum j\in Spij(a)v\alpha j(f\infty)>v\alpha i(f\infty)\}A(i,f):=$
$\{a\in A(i)|ri(a)+\alpha\sum j\in Spij(a)vj\alpha(f\infty)>vi\alpha(f\infty)\}$ Take $i\in Si\in S$ and $f\infty\in C(D)f\infty\in C(D)$. Prove the following statements:
    1. If $A(i,f)=\emptyset A(i,f)=\emptyset$, for every $i\in Si\in S$, then $f\infty f\infty$ is an $\alpha\alpha$-discounted optimal policy.
    2. If $A(i,f)\neq\emptyset A(i,f)\neq\emptyset$, for some $i\in Si\in S$, then $v\alpha(g\infty)>v\alpha(f\infty)v\alpha(g\infty)>v\alpha(f\infty)$ for any $g\infty\in C(D)g\infty\in C(D)$ with $g\neq fg\neq f$ and $g(i)\in A(i,f)g(i)\in A(i,f)$ when $g(i)\neq f(i)g(i)\neq f(i)$.