Reinforcement Learning KU (708.062) WS23

Assignment 1

Iterative Policy Evaluation

Assigned on:	10.11.2023 12:30
Q&A Session:	17.11.2023 12:30
Deadline:	**23.11.2023 23:55**
Submission:	via TeachCenter: `https://tc.tugraz.at/main/course/view.php?id=3110`
Group size:	There are no groups allowed for this task.
Remarks:	Your submission will be graded based on correctness and clarity. Include intermediary steps, and textually explain your reasoning process.

Math Recap

The following concepts have also been discussed in the lecture.

Definition 1 *Let V be a vector space. Then $f : V \mapsto \mathbb{R}_0^+$ is a norm on V provided the following hold:*

1. *$f(v) = 0$ if and only if $v = 0$*
2. *For any $\lambda \in \mathbb{R}, v \in V, f(\lambda v) = |\lambda| f(v)$*
3. *For any $v, u \in V, f(u + v) \leq f(v) + f(u)$*

A vector space together with a norm is called a normed vector space.

According to Definition 1, a norm is a function that assigns a nonnegative number to each vector, which can be interpreted as some notion of "length". The norm of a vector $\mathbf{v}$ is often denoted by $||\mathbf{v}||$. In the n-dimensional Euclidean vector space $V = \{v \,|\, v = (x_1, \ldots, x_n)^T,\ x_1, \ldots, x_n \in \mathbb{R}\}$, a common choice of norm is the max norm $||\mathbf{v}||_\infty = \max_i |v_i|$.

A norm $||\cdot||$ gives rise to a distance measure between two vectors v and u by taking the norm of their difference, i.e. $||v - u||$.

Definition 2 *Let $(v_n; n \geq 0)$ be a sequence of vectors of a normed vector space $V = (V, ||\cdot||)$. Then v_n is called a Cauchy-sequence if $\lim_{n \mapsto \infty} \sup_{m \geq n} ||v_n - v_m|| = 0$, i.e., the elements of the sequence become arbitrarily close as the sequence continues.*

Definition 3 *A normed vector space V is called complete if every Cauchy sequence in V converges to some element in V.*

As an example, every Cauchy sequence in the real numbers $\mathbb{R}$ converges to some real number—hence, the real numbers form a complete vector space. On the other hand, we can construct Cauchy sequences in the rational numbers $\mathbb{Q}$ which converge to some irrational number (e.g. to the number $\pi = 3.141592\ldots$)—hence, the rational numbers are not a complete vector space (they are still a normed vector space though).

Definition 4 *A complete, normed vector space is called a Banach space.*

Definition 5 *Let $V = (V, ||\cdot||)$ be a normed vector space. A mapping $T : V \mapsto V$ is called L-Lipschitz if for any $u, v \in V$,*

$$||T(u) - T(v)|| \leq L||u - v||. \tag{1}$$

T is called a contraction if it is L-Lipschitz with $L < 1$. In this case, L is called the contraction factor of T, and T is called an L-contraction.

Definition 6 *Let $T : V \mapsto V$ be some mapping defined on some vector space V. Any vector $v \in V$ for which $T(v) = v$ is called a fixed point of T.*

The **Banach fixed-point theorem** says that a contraction T in a Banach space always has a unique fixed point, and iterating T will always converge to it:

Theorem 1 *Let V be a Banach space and $T : V \mapsto V$ be a contraction mapping. Then T has a unique fixed point v^*. Furthermore, for any $v_0 \in V$, let $(v_n; n \geq 0)$ be the sequence of vectors defined via $v_{n+1} = T(v_n)$. For any v_0, this sequence converges to v^*.*

1 Iterative Policy Evaluation [5 points]

In the lecture, we learned about computing the value function v_π by solving the Bellman equation via closed-form matrix inversion. However, this approach does not scale well to large-scale MDPs. To this end, we considered an iterative approach based on the Bellman equation, i.e., interpreting the Bellman equation as an update rule:

$$V_{new}(s) \leftarrow r(s) + \gamma \sum_{s'} p(s'|s)\, V_{old}(s'), \tag{2}$$

where γ is the discounting factor, $r(s)$ is the expected reward function and $p(s'|s)$ is the state transition. Iterative Policy Evaluation is detailed in Algorithm 1.

Algorithm 1: Iterative Policy Evaluation, for estimating $V \approx v_\pi$

Input: π, the policy to be evaluated
Data: a small threshold $\theta > 0$ determining accuracy of estimation
Output: $V \approx v_\pi$
initialize $V(s)$ arbitrarily for all $s \in \mathcal{S}$, and $V(\text{terminal})$ to 0;
repeat
  $\Delta \leftarrow 0$;
  foreach $s \in \mathcal{S}$ **do**
    $v_{old} \leftarrow V(s)$;
    $V(s) \leftarrow r(s) + \gamma \sum_{s'} p(s'|s)\, V(s')$;
    $\Delta \leftarrow \max(\Delta, |v_{old} - V(s)|)$;
  end
until $\Delta < \theta$;

Your task: Prove that for $\gamma < 1$, Iterative Policy Evaluation (Algorithm 1) always converges to v_π, for any MDP, any policy π and any initialization of $V(s)$. The key step is to interpret the value function as a $|\mathcal{S}|$-dimensional Euclidean vector and show that the Bellman equation is a contraction for any $\gamma < 1$. When this has been shown, the rest of the proof (which should be provided) will follow via Banach's fixed point theorem.