ECE183DA (Winter 2022)

Design of Robotic Systems I

Prof. Ankur Mehta ⟨mehtank@ucla.edu⟩

Lecture 2 | Planning / control on MDPs

Addendum to lecture videos

Errata

- When I added the policy superscript π to the Bellman backup equation at 8:30 in lec02b, I should have put it on the Bellman operator T. It should read: For any value function V:

$$(T^\pi V)(s) = \underset{s'}{E}\left[r(s, \pi(s), s') + \gamma V(s')\right].$$

  That is, the policy π shows up as an argument to the reward r and the transition probabilities P (inherent in the expectation over s'), both of which are elements of the Bellman operator T^π itself.

- I made a premature simplification in my derivation of the recurrence relationship for the value function under a Bellman backup. I start with the expectation:

$$E\left[r_0 + \sum_{t=1}^{H-1} \gamma^t r_t\right] = E\left[r_0 + \gamma \sum_{\tau=0}^{H-2} \gamma^\tau r_{\tau+1}\right]$$

  In the next few lines, I break this expectation apart:

$$E_{s_1 \dots s_H | s_0}\left[r_0 + \sum_{t=1}^{H-1} \gamma^t r_t\right] = E_{s_1|s_0}\left[r_0 + \gamma E_{s_2 \dots s_H | s_0, s_1}\left[\sum_{\tau=0}^{H-2} \gamma^\tau r_{\tau+1}\right]\right]$$

  Note that I have explictly added the conditional statements to these expectations, which I'd omitted for simplicity in the lecture. Confirm for yourself that these conditionals are necessary and correct.

  And now, the correct equation for the sum inside the expectation becomes:

$$E_{s_2 \dots s_H | s_0, s_1}\left[\sum_{\tau=0}^{H-2} \gamma^\tau r_{\tau+1}\right] = E_{s_2 \dots s_H | s_1}\left[\sum_{\tau=0}^{H-2} \gamma^\tau r_{\tau+1}\right] = E_{s_1 \dots s_{H-1} | s_0}\left[\sum_{\tau=0}^{H-2} \gamma^\tau r_\tau\right]$$

  The first half of that expression comes from the Markov property. The second half is a re-labeling of time indices (which is allowed because of our restriction to time-invariant systems), noting that r_t is shorthand for $r(s_t, a_t, s_{t+1})$. Again, you should confirm both of these steps for yourself.

  With this, the rest of the derivation continues as presented.

Clarification

Some more detail regarding the Bellman backup operator T, and what it means for the operator to map a function onto a function. You could read the equation above as: "The operator T^π takes as input a function V, and outputs a new function. This new function, when evaluated on a state s, returns the quantity $E_{s'}[r(s, \pi(s), s') + \gamma V(s')]$." Another way to look at it is to just note that in discrete spaces, functions are simply other sets – V is a set of N_S real values, one per state in S. The operator T^π can be applied to a set V to yield a different set of N_S real values, each related to the input set V by the relationship above.

Additional References

- This material is covered in Pieter Abbeel's CS287 at UC Berkeley—with slightly different notation—in lectures 2 and 3.
  - `https://people.eecs.berkeley.edu/~pabbeel/cs287-fa19/`
- You can also see a textbook covering this material with significantly different notation, but with many more examples, in chapters 3 and 4 of Sutton and Barto, "Reinforcement Learning"
  - `http://incompleteideas.net/book/the-book.html`