Course 1, Module 5
Dynamic Programming

CMPUT 397
Fall 2019

Weekly Schedule

- **Sunday: Discussion question due, deadline for completing practice quiz**
- Monday: Review of module, Q&A session about content. Finish with class exercise question
- Wednesday: In-class Discussion based on your submitted discussion topics
- **Friday: Graded Assessment (usually python notebook) due**
- Friday: Finish discussion if needed. More in-class exercise questions from worksheet

Are you in the private session?

- **How to find out?**
  - check eclass!!!! Do you have marks for the notebooks you have done?
  - check your email!!! The TAs have been personally emailing people to inform them
- **The hour is late!!!**
  - if you are not in the private session, then do something about it today or risk getting zero on everything to date
- Check Eclass announcements weekly!

Any questions about course admin?

Review of Course 1, Module 5
Dynamic Programming

Video 1: Policy Evaluation vs. Control

- Introduce the two classic problems of RL: prediction and control. Classic assumptions of DP
- Goals:
  - Understand the distinction between policy **evaluation** and **control**
  - Explain the **setting** in which dynamic programming can be **applied**, as well as its limitations

Video 2: Iterative Policy Evaluation

- How to turn Bellman equations into algorithms for **computing** value functions and policies
- Goals:
  - Outline the iterative policy evaluation algorithm for estimating state values for a given policy
  - Apply iterative policy evaluation to compute value functions. Example

Video 3: Policy Improvement

- **Key theoretical result** in RL and DP! How to make the policy better using the value function
- Goals:
  - Understand the **policy improvement theorem**; and how it can be used to construct improved policies
  - And use the value function for a policy to **produce a better policy**

Video 4: Policy Iteration

- Our first control algorithm. Why sequencing evaluation and improvement works!
- Goals:
  - Outline the **policy iteration algorithm** for finding the optimal policy;
  - Understand **“the dance of policy and value”**, how policy iteration reaches the optimal policy by alternating between evaluating a policy and improving it
  - Apply policy iteration to compute optimal policies and optimal value functions

Video 5: Flexibility of the Policy Iteration Framework

- Generalized Policy Iteration: a general framework for control
- **Goals**:
  - Understand the framework of generalized policy iteration
  - Outline value iteration, an important special case of generalized policy iteration
  - *D*ifferentiate synchronous and asynchronous dynamic programming methods

Video 6: Efficiency of Dynamic Programming

- DP is actually pretty good, compared to other approaches! What's the deal with **Bootstrapping**?

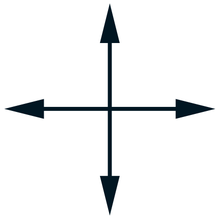
- **Goals**:

  - Describe Monte-Carlo sampling as an **alternative** method for learning a value function

  - Describe brute force search as an **alternative** method for finding an optimal policy; and

  - Understand the advantages of Dynamic programming and "**bootstrapping**" over these alternatives.

Practice Questions



actions

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

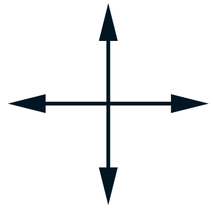
$R_t = -1$
on all transitions

$p(6, -1 | 5, \texttt{right}) =$

$p(10, r | 5, \texttt{right}) =$

$p(7, -1 | 7, \texttt{right}) =$

Practice Questions



Actions

T	1	2	3
4	5	6	7
8	9	10	11
12	13	14	T
	15		

R = -1
on all transitions

T	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	T
	-20.		

Exercise 4.1 In Example 4.1, if π is the equiprobable random policy, what is $q_\pi(11, \texttt{down})$? What is $q_\pi(7, \texttt{down})$? □

Practice Questions

In iterative policy evaluation, we seek to find the value function for a policy π by applying the Bellman equation many times to generate a sequence of value functions v_k that will eventually converge to the true value function v_π. How can we modify the update below to generate a sequence of action value functions q_k?

$$v_{k+1}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_k(s')\right]$$