MIDTERM REVIEW

MIDTERM

WHAT TO BRING

Monday, Feb 12 1:00 pm-1:50 pm in this room

Be early! No entry after 1:30 pm and cannot leave before 1:30 pm

There is a class afterward, so no extra time will be given.

Bring your One Card!

Pens/pencils

We will provide scratch paper, and it will be left in the exam room

The exam is closed note/book/internet/etc

GUIDE TO STUDY FOR THE MIDTERM

SUGGESTIONS

Review the material from this lecture and the rest of the course.

If something is not clear from this review, go find the material from previous lectures and review it.

Review quiz questions

Textbook questions are great practice for the midterm

As with any computing science course, Google is a great place to look for new explanations

BANDITS

METHODS

$\mathscr{A}$ — set of actions

$$q_*(a) \doteq \mathbb{E}[R_t | A_t = a]$$

Exploration strategies: Greedy, ϵ-greedy, UCB

Greedy $\arg\max_a Q_n(a)$

ϵ-greedy: with probability $1 - \epsilon$ select an action from $\arg\max_a Q_n(a)$, otherwise sample action from the uniform distribution on $\mathscr{A}$

UCB — $\arg\max_a Q_t(a) + c\sqrt{\ln(t)/N_t(a)}$

BANDITS

METHODS

$\mathscr{A}$ — set of actions

$q_*(a) \doteq \mathbb{E}[R_t \,|\, A_t = a]$

Exploration strategies: Greedy, ϵ-greedy, UCB

How does each strategy trade off exploration and exploitation?

How do you vary the amount of exploration?

Be able to compare each strategy as $t \to \infty$

Textbook chapter 2 exercises and quiz questions are a good place to review.

BANDITS

ESTIMATING q_*

$$Q_n(a) = \frac{1}{n} \sum_{i=1}^{n} R_i$$

Iterative estimation:

$$Q_{n+1}(a) = Q_n(a) + \alpha_n \left(R_n - Q_n(A)\right)$$

Step size schedules:

$\alpha_n = \alpha$ — Constant

$\alpha_n = 1/t$ — decaying

BANDITS

STEP SIZES

Properties of each step size:

$\alpha_n = 1/t$

- estimate the average of the rewards observed so far
- Not useful for nonstationary rewards

$\alpha_n = \alpha$

- will never converge to q_* but will oscillate around it
- Larger α has a smaller averaging window (tracks most recent rewards)
- Smaller α will more closely track q_* but will take longer to get there

Review book questions and quiz questions on this topic for practice

MDPS

DEFINITIONS

$M = (\mathscr{S}, \mathscr{A}, \mathscr{R}, p, d_0, \gamma)$

$\mathscr{S}$ — set of all states

$\mathscr{A}$ — set of all actions

$\mathscr{R}$ — set of all rewards

S_t, A_t, R_{t+1} — random variables representing the state and action at time t and a random variable representing the reward the agent receives at time $t+1$ after taking action A_t in state S_t

$S_0, A_0, R_1, S_1, A_1, \ldots, R_{T-1}, S_{T-1}$ — an episode of length T (know the differences between episodic and continuing MDPs)

MDPS

DEFINITIONS

p — describes the dynamics of the MDP

$p : \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0,1]$

$p(s', r \mid s, a) \doteq \Pr(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$

$p(s, a, s') \doteq \Pr(S_{t+1} = s' \mid S_t = s, A_t = a)$

$r(s, a, s') \doteq \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s']$

$r(s, a) \doteq \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$

d_0 — describes the start state distribution

$d_0(s) \doteq \Pr(S_0 = s)$

$\gamma \in [0,1]$ — reward discount factor

MARKOV PROPERTY

DEFINITIONS

Markov Property: $\forall t' > t$

$$\Pr(S_{t'} = s' | S_t = s, R_t = r_t, A_{t-1} = a_{t-1}, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0) = \Pr(S_{t'} = s' | S_t = s)$$

$$\Pr(S_{t'} = s' | S_t = s, A_t = a, R_t = r_t, A_{t-1} = a_{t-1}, S_{t-1} = s_{t-1}, \ldots, S_0 = s_0) = \Pr(S_{t'} = s' | S_t = s, A_t = a)$$

The future ($> t$) is conditionally independent of the past ($< t$) given S_t

PROBABILITY

RULES

Chain Rule: $\Pr(A, B) = \Pr(A \mid B)\Pr(B)$

Conditional Probability $\Pr(A \mid B) = \dfrac{\Pr(A, B)}{\Pr(B)}$

Bayes Rule: $\Pr(A \mid B) = \dfrac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}$

Marginalization $\Pr(X = x) = \sum_{y} \Pr(Y = y, X = x)$

Chain Rule for >2 events: $\Pr(A, B, C) = \Pr(A \mid B, C)\Pr(B \mid C)\Pr(C)$

PROBABILITY

RULES

Be able to apply rules and Markov Property to MDPs

See earlier examples in class for practice. Here are two more

$$\Pr(S_2 = s) = \sum_{s_0, a_0, s_1, a_1} d_0(s_0)\pi(a_0 \mid s_0)p(s_0, a_0, s_1)\pi(a_1 \mid s_1)p(s_1, a_1, s)$$

$$\Pr(A_4 = a' \mid S_2 = s, A_3 = a) = \frac{\Pr(A_4 = a', A_3 = a \mid S_2 = s)}{\Pr(A_3 = a \mid S_2 = s)} = ?$$

Hint #1: Ask what terms are missing between the earliest given information and the latest. You will need to add these

Hint #2: Think about where you can apply the Markov property, e.g., given S_t means you can drop a given S_{t-1} term

POLICIES

DEFINITIONS

$\pi : \mathscr{A} \times \mathscr{S} \to [0,1]$

$\pi(a \mid s) \doteq \Pr(A_t = a \mid S_t = s)$ – action probability only depends on the state

Π – set of all policies

$\pi : \mathscr{S} \to \mathscr{A}$, $a = \pi(s)$ is a deterministic policy, i.e., $\Pr(A_t = a \mid S_t = s) = 1$

WHAT IS A DEFINITION

DEFINITIONS

A definition is the base description of the symbol.

An equality is just a mathematical statement that asserts two sides $=$ are equivalent. They are not definitions!

$G_t \doteq \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$ — return or (discounted) cumulative reward

$G_t = R_{t+1} + \gamma G_{t+1}$ — not a definition

VALUE

DEFINITIONS

$$v_\pi(s) \doteq \mathbb{E}\left[G_t \,|\, S_t = s\right]$$

$$q_\pi(s,a) \doteq \mathbb{E}[G_t \,|\, S_t = s, A_t = a]$$

$v_\pi(s) = \sum_a \pi(a \,|\, s) q_\pi(s,a)$ – not a definition

OPTIMAL VALUE FUNCTIONS

DEFINITIONS

$$v_*(s) \doteq \max_{\pi \in \Pi} v_\pi(s)$$

$$q_*(s, a) \doteq \max_{\pi \in \Pi} q_\pi(s, a)$$

OPTIMAL POLICIES

DEFINITIONS

$\pi \geq \pi'$ if $\forall s \in \mathcal{S},\ v_\pi(s) \geq v_{\pi'}(s)$

Definition 1: a policy π is in the set of optimal policies Π_* if no other policy is better in any state

$$\forall \pi' \in \Pi, \pi \geq \pi'.$$

Definition 2: not on the midterm, but useful for the rest of the course

π_* can be any policy $\pi \in \Pi_*$

OPTIMAL POLICIES AND VALUE FUNCTIONS

DEFINITIONS

For every finite MDP, there exists at least one optimal deterministic policy

If, for every state, there is only one optimal action, then $|\Pi_*| = 1$, i.e., only one policy.

If, for some state, there are 2 or more optimal actions, then there are an infinite number of optimal policies.

Let $\mathscr{A}^*(s)$ be the set of optimal actions for a state s.

$\pi \in \Pi_*$ if $\forall s, \quad \sum_{a \in \mathscr{A}^*(s)} \pi(a \mid s) = 1$

There is only nonzero probability on the optimal actions.

OPTIMAL POLICIES AND VALUE FUNCTIONS

DEFINITIONS

For any policies $\pi, \pi' \in \Pi_*$, then $v_*(s) = v_\pi(s) = v_{\pi'}(s)$ and $q_*(s,a) = q_\pi(s,a) = q_{\pi'}(s,a)$

All optimal policies have the same value functions

The optimal value functions are unique

$$v_*(s) = \max_a q_*(s,a)$$

π_* can be deterministic and could take any action with the largest value

REWARD FUNCTIONS AND OPTIMAL POLICIES

PROPERTIES

Be able to reason about the optimal policy given the reward function and γ

$\gamma < 1$ encourages places a preference for having larger rewards sooner

$\gamma = 1$ places no preference on when large rewards occur so long as the total is maximized

How do positive and negative rewards influence optimal behavior?

How might this vary on episodic and continuing problems?

Review MDP examples from class, Coursera, and the textbook.

BELLMAN EQUATION

EXPRESSIONS

$$
\begin{aligned}
v_\pi(s) &= \sum_a \pi(a\,|\,s)\left(r(s,a) + \gamma \sum_{s'} p(s,a,s')v_\pi(s')\right)\\
&= \sum_a \pi(a\,|\,s)\sum_{s'} p(s,a,s')\big(r(s,a,s') + \gamma v_\pi(s')\big)\\
&= \sum_a \pi(a\,|\,s)\sum_{s',r} p(s',r\,|\,s,a)\big(r + \gamma v_\pi(s')\big)\\
&= \sum_a \pi(a\,|\,s)\sum_{s',r} p(s',r\,|\,s,a)\left(r + \gamma \sum_{a'} \pi(a'\,|\,s')q_\pi(s',a')\right)
\end{aligned}
$$

BELLMAN EQUATION

EXPRESSIONS

$$
\begin{aligned}
q_\pi(s,a) = \;& r(s,a) + \gamma \sum_{s'} p(s,a,s') v_\pi(s') \\
= \;& \sum_{s'} p(s,a,s') \big(r(s,a,s') + \gamma v_\pi(s') \big) \\
= \;& \sum_{s',r} p(s',r \,|\, s,a) \big(r + \gamma v_\pi(s') \big) \\
= \;& \sum_{s',r} p(s',r \,|\, s,a) \left(r + \gamma \sum_{a'} \pi(a' \,|\, s') q_\pi(s',a') \right)
\end{aligned}
$$

BELLMAN OPTIMALITY EQUATION

EXPRESSIONS

$$v_*(s) = \max_a r(s,a) + \gamma \sum_{s'} p(s,a,s') v_*(s')$$

$$q_*(s,a) = r(s,a) + \gamma \sum_{s'} p(s,a,s') \max_{a'} q_*(s',a')$$

BELLMAN EQUATION

EXPRESSIONS

Bellman equations are not definitions!

Be able to convert from one form to another

Be able to derive the Bellman Equation from the definition

COMPUTE VALUE FUNCTIONS

EXPRESSIONS

Know the iterative update process for v^k and q^k

$$v_i^{k+1} = \sum_a \pi(a \mid s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^k\right)$$

$$q_{i,a}^{k+1} = r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) \sum_{a'} \pi(a' \mid s_j) q_{j,a}^k$$

In general $\lim_{t \to \infty} v^k \to v_\pi$, but it could reach v_π in finite time (see the example from the class slides)

Be able to work out by hand the update process for simple MDPs

BELLMAN OPERATOR

DEFINITION

$\mathcal{T} : \mathcal{S} \to \mathcal{S}$

$$\mathcal{T}(v)_i = \sum_a \pi(a \,|\, s_i)\left(r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j \right)$$

$\mathcal{T}$ repeated applications of the operator lead to a unique fixed point $v^* = \mathcal{T}(v^*)$

If $v = \mathcal{T}(v)$ then $v = v_\pi$

Because $v = \mathcal{T}(v)$ obeys the Bellman equation for all states

POLICY ITERATION

EXPRESSIONS

Policy improvement theorem: choosing action greedily with respect to q_π improves the policy

Know the general process:

1. Evaluate the policy
2. Be greedy with respect to the value function estimate
3. Stop when the policy stops changing

VALUE ITERATION

EXPRESSIONS

Know the general process

$$v_i^{k+1} = \max_a r(s_i, a) + \gamma \sum_j p(s_i, a, s_j) v_j^k$$

$$\lim_{t \to \infty} v^k \to v_*$$

Can take an infinite amount of time to compute v_*

The optimal policy can often be reached before reaching v_*

Be able to step through value iteration by hand for simple MDPs

QUESTION FROM GOOGLE FORM

FORM: HTTPS://FORMS.GLE/QNAOIFPLZPWAYN2P9

How does UCB exploit:

$$A_t \in \arg\max_a Q_t(a) + c\sqrt{\ln(t)/N_t(a)}$$

The agent is greedy with respect to the value estimate plus an uncertainty term

The impact of $c\sqrt{\ln(t)/N_t(a)}$ on A_t shrinks over time, so the agent eventually becomes very certain of the optimal action.

NEXT CLASS

WHAT YOU SHOULD DO

1. Review / Practice / Study for midterm
2. Make sure you have signed up for the second and third Coursera courses!

Monday: Midterm