CS294-40 Learning for Robotics and Control Lecture 2 - 9/2/2008

Value Iteration

Lecturer: Pieter Abbeel *Scribe: Anand Kulkarni*

1 Lecture outline

- Problem set 1 out this week, due Sept. 23.
- Project reminder - set up appointment to discuss and get approved a final project before Sept. 12.
- Markov decision processes: examples, continued.
- Value iteration: finite horizon, infinite horizon.
- Contractions

2 Markov decision processes (MDPs)

2.1 Examples, continued

2.1.1 Tetris

Recall the popular computer game Tetris. In Tetris, pieces descend vertically one by one to stack on a game board, clearing when a row is fully covered. The game ends when the height of the pieces exceed the top of the game board, and the goal is to play for as long as possible. One piece out of four possibilities becomes available at any given time, and we may choose where to place this piece and in which orientation. Here, the perturbation encapsulates the uncertainty in which piece will become available next.

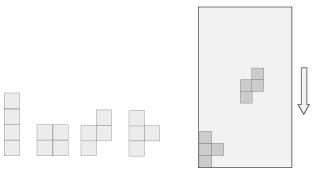

Figure 1: Tetris pieces and gameplay.

$$\begin{aligned} s_t &= \text{(board configuration, shape of current piece)} \\ a_t &= \text{choice of where to place the current piece, orientation and position} \\ w_t &= \text{next falling piece} \end{aligned}$$

A reward function that captures the goal of playing as long as possible is given by:

$$\begin{aligned} R(s_t) = &-1 \cdot \mathbf{1}\{\text{if height of pieces exceeds height of board}\} + 0 \cdot \mathbf{1}\{\text{otherwise}\} \\ &\gamma \in (0,1) \end{aligned}$$

The optimal policy under this reward function will always play for as long as possible to delay the losing condition, regardless of our choice of γ in this range. ($\gamma < 1$ rather than $\gamma = 1$ gives a simple guarantee that the infinite sum in the reward will converge, as well as being more sensible in the context of our model)

Having seen some simple examples of Markov decision processes, we now turn to the question of finding optimal policies for a given Markov decision process.

2.2 Finding optimal policies: value iteration

Our goal in finding an optimal policy is to find a policy that maximizes the expected total of rewards earned over the periods of our decision process.

2.2.1 Finite horizon

We begin with the finite horizon case, where we are concerned with decisions and rewards only up until a given time $t = H$. To avoid having to discuss some measure-theoretic details, we assume that our state space $\mathcal{S}$ and our action space $\mathcal{A}$ are finite.

The value of any policy π is

$$V_\pi(s_0) = E[\sum_{t=0}^{H} \gamma^t R(s_t)|\pi; s_0]$$

and we are interested in finding

$$\max_\pi \quad V_\pi(s_0)$$
$$\pi = \{\mu_0, \mu_1, \ldots, \mu_{H-1}\}, \text{ where } \mu_i : S \to A$$

Since there are $|A|^{|S|}$ possible mappings for each μ_i, there are $(|A|^{|S|})^H$ possible policies π, which is far too many to compute the value of all possible options. Instead, we apply a dynamic programming algorithm known as value iteration to find the optimal policy efficiently. Intuitively, we are applying the notion that given a state, the past and future are independent (the "Markov property").

Define the value function at the kth time-step as

$$V_k(s_k) = \max_{\mu_k, \mu_{k+1}, \ldots, \mu_{H-1}} E[\sum_{t=k}^{H} \gamma^{t-k} R(s_t)|s_k]$$

Our first theorem states that we can find the optimal policy efficiently by working backwards from H.

Theorem 1.

$$V_k(s_k) = \max_{a \in A}[R(s_k) + \gamma \sum_{s'} P(s'|s, a)V_{k+1}(s')], \text{ where } V_H \equiv 0.$$

Proof.

$$\begin{aligned} V_k(s_k) &= \max_{\mu_k, \mu_{k+1}, \ldots, \mu_{H-1}} E[R(s_k) + \sum_{t=k+1}^{H} \gamma^{t-k} R(s_t)|s_k] \\ &= \max_{\mu_k, \mu_{k+1}, \ldots, \mu_{H-1}} E\left[R(s_k) + \sum_{s'} P(s'|s, \mu_k(s)) \cdot (E[R(s_k) + \sum_{t=k+1}^{H} \gamma^{t-k} R(s_t)|s_{k+1} = s'])\middle| s_k\right] \end{aligned}$$

Now, observe that the outer expectation depends only on μ_k, and bring $\mu_{k+1}, \ldots, \mu_H$ inside the expectation.

$$= \max_{\mu_k} E\left[R(s_k) + \sum_{s'} P(s'|s, \mu_k(s)) \cdot \max_{\mu_{k+1}, \mu_{k+1}, \ldots, \mu_{H-1}} E[R(s_k) + \sum_{t=k+1}^{H} \gamma^{t-k} R(s_t)|s_{k+1} = s']\middle| s_k\right]$$

$$= \max_{\mu_k} E[R(s_k) + \sum_{s'} P(s'|s, \mu_k(s)) \cdot \gamma \cdot V_{k+1}(s')], \text{ by definition.}$$

□

Further, we note that a policy $\pi^* = (\mu_0^*, \ldots, \mu_{H-1}^*)$ is optimal if and only if

$$\mu_k^*(s) \in \arg\max_{a \in A}[R(s_k) + \gamma \sum_{s'} P(s'|s, a) V'_{k+1}(s)].$$

Value iteration provides an efficient (i.e., polynomial-time) algorithm to find the optimal policy. Each step backwards from $V_{k+1} \to V_k$ requires $O(|S|^2|A|)$ time, for a total runtime of $O(H * |S|^2|A|)$ to find the optimal policy. Of course, efficient is a relative concept; we will see in later lectures that $|S|$ is often exponential in the number of variables describing the states.

2.2.2 Infinite horizon optimal policies

Now we are interested in finding an optimal policy when our horizon is infinite.

The value of a policy is now

$$V_\pi(s_0) = E[\sum_{t=0}^{\infty} \gamma^t R(s_t)|s_0, \pi]$$

and we are interested in finding $V^*(s) = \max_\pi V_\pi(s)$ and $\pi^* = \arg\max_\pi V_\pi(s)$

Despite the fact that the expectation is taken over an infinite sum, it is guaranteed to converge. This follows because R is a function over a finite state space, and is therefore bounded, and since $\gamma \in (0, 1)$.

How can we go about finding an optimal policy? Intuitively, as rewards further into the future are discounted exponentially, for a large enough finite horizon the value function for the finite horizon process will be close to the value function for the infinite horizon process. We formalize this intuition in Theorem 3. First we introduce the Bellman backup operator, also referred to as the Dynamic Programming operator, denoted T.

Definition 2 (Bellman Backup Operator)**.** *Let $V : S \to R$ be our value function and let $V_\pi(s) = E[\sum_{t=0}^{\infty} \gamma^t R(s_t)|s_0, \pi]$. Define the operator $T : V \to TV$ as*

$$\boxed{(TV)(s) = \max_{a \in A}[R(s) + \gamma \sum_{s'} P(s'|s, a) V(s')]}$$

We now apply this operator to prove that the optimal infinite horizon policy is the limit of the optimal finite horizon policy as the horizon H goes to infinity.

Theorem 3.

$$V^* = \lim_{H \to \infty} T^H V, \quad \forall\ V \in \mathbb{R}^{|S|}$$

Proof.

$$V_\pi(s) = E[\sum_{t=0}^{H-1} \gamma^t R(s_t)|s_0 = s; \pi] + E[\sum_{t=H}^{\infty} \gamma^t R(s_t)|s_0 = s; \pi]$$

Recall that $\|x\|_\infty = \max_i |x_i|$. Thus, $R \leq \|R\|_\infty$, so the second expectation is bounded above by the geometric sum

$$E[\sum_{t=H}^{\infty} \gamma^t R(s_t)|s_0 = s; \pi] \quad \leq \quad \frac{\gamma^H}{1-\gamma}\|R\|_\infty.$$

So,

$$E[\sum_{t=H}^{\infty} \gamma^t R(s_t) + \frac{\gamma^H}{1-\gamma}\|R\|_\infty|s_0 = s; \pi] \quad \geq \quad 0$$

Thus, we can bound $(T^H V)$ above, by

$$\begin{aligned}
(T^H V)(s) &= \max_\pi E[\gamma^H V(S_H) + \sum_{t=0}^{H-1} \gamma^t R(S_t)|\pi, s_0 = s] \\
&\leq \max_\pi E[\gamma^H V(S_H) + \sum_{t=0}^{H-1} \gamma^t R(s_t) + \sum_{t=H}^{\infty} \gamma^t R(s_t) + \frac{\gamma^H}{1-\gamma}\|R\|_\infty] \\
&\leq V^*(s) + \gamma^H\|V\|_\infty + \frac{\gamma^H}{1-\gamma}\|R\|_\infty
\end{aligned}$$

and, identically, below, with

$$\begin{aligned}
(T^H V)(s) &= \max_\pi E[\gamma^H V(S_H) + \sum_{t=0}^{H-1} \gamma^t R(S_t)|\pi, s_0 = s] \\
&\geq \max_\pi E[\gamma^H V(S_H) + \sum_{t=0}^{H-1} \gamma^t R(s_t) + \sum_{t=H}^{\infty} \gamma^t R(s_t) - \frac{\gamma^H}{1-\gamma}\|R\|_\infty] \\
&\geq V^*(s) - \gamma^H\|V\|_\infty - \frac{\gamma^H}{1-\gamma}\|R\|_\infty
\end{aligned}$$

Together, we have that

$$V^*(s) - \gamma^H\|V\|_\infty - \frac{\gamma^H}{1-\gamma}\|R\|_\infty \leq (T^H V)(s) \leq V^*(s) + \gamma^H\|V\|_\infty + \frac{\gamma^H}{1-\gamma}\|R\|_\infty.$$

Taking the limit as H goes to infinity on both sides gives

$$V^*(s) \leq \lim_{H\to\infty} (T^H V)(s) \leq V^*(s)$$

□

From the proof we observe that we can bound the suboptimality of any incomplete run of the value iteration algorithm as a function of the number of backsteps used.

3 Contractions

Theorem 4. *T is a max-norm γ-contraction, i.e.,*

$$\|TV - T\overline{V}\|_\infty \leq \gamma \|V - \overline{V}\|_\infty$$

Proof.

$$\begin{aligned}
|(TV)(s) - (T\overline{V})(s)| &= \left| \left(\max_a R(s) + \gamma \sum_{s'} P(s'|s,a) V(s') \right) - \left(\max_a R(s) + \gamma \sum_{s'} P(s'|s,a) \overline{V}(s') \right) \right| \\
&\leq \max_a \left| \left(R(s) + \gamma \sum_{s'} P(s'|s,a) V(s') \right) - \left(R(s) + \gamma \sum_{s'} P(s'|s,a) \overline{V}(s') \right) \right| \\
&= \max_a \gamma \sum_{s'} P(s'|s,a) |V(s') - \overline{V}(s')| \\
&\leq \max_a \gamma \sum_{s'} P(s'|s,a) \|V - \overline{V}\|_\infty \\
&= \gamma \|V - \overline{V}\|_\infty \max_a \sum_{s'} P(s'|s,a) \\
&= \gamma \|V - \overline{V}\|_\infty
\end{aligned}$$

□