Lecture 3

Policy Iteration

Value iteration is a powerful tool for solving optimal control problems. However, value iteration can be slow to converge, especially in the case of cyclic graphs and negative costs/weights. One observation is that the policy usually converges more rapidly than the value function. Policy iteration is an algorithm which optimizes the policy rather than the value function. It alternates between policy evaluation and policy improvement.

3.1. Policy Evaluation

We can write the optimal cost-to-go as

$$J^*(s_0) = \min_{a_k \in \mathbb{A} \ \forall k} \sum_{k=0}^{\infty} \gamma^k g(s_k, a_k). \tag{3.1}$$

This can be written recursively as

$$J^*(s) = \min_a g(s, a) + \gamma J^*(s'). \tag{3.2}$$

where we drop the dependence on time k.

Now it may be the case that we want to be able to evaluate the cost-to-go of a non-optimal policy, $a_k = \pi(s_k)$. The cost-to-go for policy π can be written as

$$J^\pi(s_0) = \sum_{n=0}^{\infty} \gamma^k g(s_k, \pi(s_k)), \tag{3.3}$$

which can be written recursively as

$$J^\pi(s) = g(s, \pi(s)) + \gamma J^\pi(f(s, \pi(s))). \tag{3.4}$$

3.2. Policy Improvement

To improve a given policy J^π, we use

$$\pi^{k+1}(s) = \arg\min_{a \in \mathbb{A}} g(s, a) + \gamma J^\pi(f(s, a)). \tag{3.5}$$

3.3. Policy Iteration

Initialize $\pi^0(s)$, $J^{(\pi^0,\ 0)}(s)$.
Iterate for $k = 0, 1, 2, ...$ until the policy converges:

1. Policy Evaluation (iterate for $i = 0, 1, 2, ...$ until the cost-to-go converges)

$$J^{(\pi^k,\ i+1)}(s) = g(s, \pi^k(s)) + \gamma J^{(\pi^k,\ i)}(f(s, \pi^k(s))) \tag{3.6}$$

2. Policy Improvement

$$\pi^{k+1}(s) = \arg\min_{a \in \mathbb{A}} g(s, a) + \gamma J^{\pi}(f(s, a)) \tag{3.7}$$

3. Initialize value function for next iteration

$$J^{(\pi^{k+1},\ 0)}(s) = J^{(\pi^k,\ i)}(s) \tag{3.8}$$

3.4. Modified Policy Iteration

Modified policy iteration is commonly implemented in practice. Instead of iterating until convergence, the policy will be evaluated m times.

3.5. Policy Evaluation for Markov Decision Processes

Remember

$$J^*(s) = \min_{a \in \mathbb{A}} \left[g(s, a) + \gamma \sum_{s'} p(s'|s, a) J^*(s') \right]. \tag{3.9}$$

Therefore we can write the policy evaluation as

$$J^{\pi}(s) = g(s, a) + \gamma \sum_{s'} p(s'|s, a) J^{\pi}(s'). \tag{3.10}$$

Note, $s \in \mathbb{S}$, where $\mathbb{S} = \{s0, s1, s2, ..., sN\}$. Thus, the cost-to-go for a particular state, $s0$, is

$$J^{\pi}(s0) = g(s0, \pi(s0)) + \gamma \begin{bmatrix} p(s0|s0, \pi(s0)) & p(s1|s0, \pi(s0)) \dots & p(sN|s0, \pi(s0)) \end{bmatrix} \begin{bmatrix} J^{\pi}(s0) \\ J^{\pi}(s1) \\ \vdots \\ J^{\pi}(sN) \end{bmatrix}. \tag{3.11}$$

We can often write J^{π} as a column vector $\vec{J}^{\pi}$. We then can write

$$\vec{J}^{\pi} = \vec{g} + \gamma T \vec{J}^{\pi} \tag{3.12}$$

$$\vec{J}^{\pi} = (I - \gamma T)^{-1} \vec{g} \tag{3.13}$$

where $\vec{g}$ is a column vector of one-step costs and $T \in \mathbb{P}^{n_S \times n_S}$, where $\mathbb{P} = [0, 1]$.

3.6. The Q-Function

Oftentimes, the "quality" function or Q-function will be used in lieu of a value function. The optimal Q-function is defined as

$$Q^*(s, a) = g(s, a) + \gamma J^*(f(s, a)). \tag{3.14}$$

Policy evaluation for a policy π is given as

$$Q^{\pi}(s, a) = g(s, a) + \gamma J^{\pi}(f(s, a)). \tag{3.15}$$

The policy update step is then quite simple and given as

$$\pi^{k+1}(s) = \arg\min_{a \in \mathbb{A}} Q^{\pi}(s,a). \tag{3.16}$$

Like before extending this to the stochastic case is straight-forward, and the policy evaluation step can be written as

$$Q^{\pi}(s,a) = g(s,a) + \gamma \sum_{s'} p(s'|s,a) J^{\pi}(f(s,a)). \tag{3.17}$$

3.7. Mathematical Analysis

3.7.1 Proof the Policy Evaluation Operator is a Contraction Mapping

Define the policy evaluation operator

$$H(\vec{J}) \triangleq \vec{g} + \gamma \vec{T} \vec{J} \tag{3.18}$$

H is a contraction mapping if

$$||H(\vec{J_1}) - H(\vec{J_2})||_{\infty} \leq \gamma ||\vec{J_1} - \vec{J_2}||_{\infty} \tag{3.19}$$

Proof:

$$\begin{aligned}
||H(\vec{J_1}) - H(\vec{J_2})||_{\infty} &= ||\vec{g} + \gamma T \vec{J_1} - \vec{g} - \gamma T \vec{J_2}||_{\infty} && (3.20)\\
&= ||\gamma T(\vec{J_1} - \vec{J_2})||_{\infty} && (3.21)\\
&\leq \gamma ||T||_{\infty} ||(\vec{J_1} - \vec{J_2})||_{\infty} \qquad \text{(Cauchy-Schwartz inequality: } ||AB|| \leq ||A||||B||) && (3.22)\\
&= \gamma ||(\vec{J_1} - \vec{J_2})||_{\infty} \qquad \left(\max_s \sum_{s'} T(s,s') = 1 \right) && (3.23)
\end{aligned}$$

3.7.2 Proof of Policy Evaluation Convergence

$$\lim_{n \to \infty} H^n(\vec{J}) = \vec{J^{\pi}} \quad \forall J \tag{3.24}$$

Proof:
We know that there is a fixed-point $\vec{J^{\pi}} = H^{\infty}(\vec{J^{\pi}})$.
Since the Bellman operation is a contraction mapping, we know

$$||H^n(\vec{J_1}) - H^n(\vec{J_2})||_{\infty} \leq \gamma^n ||\vec{J_1} - \vec{J_2}||_{\infty}. \tag{3.25}$$

When $n \to \infty$ and $\vec{J_2} = \vec{J^{\pi}}$, we have

$$||H^n(\vec{J_1}) - H^n(\vec{J^{\pi}})||_{\infty} \leq \gamma^n ||\vec{J_1} - \vec{J^{\pi}}||_{\infty} \to 0. \tag{3.26}$$

This leads to

$$||H^{\infty}(\vec{J_1}) - H^{\infty}(\vec{J^{\pi}})||_{\infty} = 0 \tag{3.27}$$

and results in

$$H^{\infty}(\vec{J_1}) = \vec{J^{\pi}}. \tag{3.28}$$

3.7.3 Proof the Value Iteration Operator is a Contraction Mapping

$$H^*(J) \triangleq \min_a \{g(s,a) + \gamma \sum_{s'} Pr(s'|s,a) J(s')\} \tag{3.29}$$

If H^* is a contraction mapping

$$||H^*(J_1) - H^*(J_2)||_\infty \leq \gamma ||J_1 - J_2||_\infty. \tag{3.30}$$

Proof:
Let $H^*(J_1)(s) \geq H^*(J_2)(s)$ and let

$$\pi^*(s) = \arg\min_a g(s,a) + \gamma \sum_{s'} Pr(s'|s,a) J_2(s'). \tag{3.31}$$

$$\begin{aligned}
0 &\leq H^*(J_1(s)) - H^*(J_2(s)) && (3.32)\\
&\leq g(s,\pi^*(s)) + \gamma \sum_{s'} Pr(s'|s,\pi^*(s)) J_1(s') - g(s,\pi^*(s)) - \gamma \sum_{s'} Pr(s'|s,\pi^*(s)) J_2(s') && (3.33)\\
&= \gamma \sum_{s'} Pr(s'|s,\pi^*(s))(J_1(s') - J_2(s')) && (3.34)\\
&\leq \gamma \sum_{s'} Pr(s'|s,\pi^*(s)) ||J_1 - J_2||_\infty \qquad \big(\text{max norm over all } s\big) && (3.35)\\
&\leq \gamma ||J_1 - J_2||_\infty \qquad \big(\sum_{s'} Pr(s'|s,\pi^*(s)) = 1\big) && (3.36)
\end{aligned}$$

Repeat the proof for $H^*(J_1)(s) \leq H^*(J_2)(s)$, for all s.

3.7.4 Proof of Value Iteration Convergence

Show that

$$\lim_{n\to\infty} H^{*(n)}(J) = J^* \quad \forall J \tag{3.37}$$

Proof:
Since H^* is a contraction mapping, we have

$$||H^{*(n)}(J_1) - H^{*(n)}(J_2)||_\infty \leq \gamma^n ||J_1 - J_2|| \tag{3.38}$$

We know $J^* = H^{*(\infty)}(J^*)$. When $n \to \infty$ and $J_2 = J^*$, we have

$$||H^{*(n)}(J_1) - H^{*(n)}(J^*)||_\infty \leq \gamma^n ||J_1 - J^*||_\infty \to 0 \tag{3.39}$$

which leads to

$$||H^{*(\infty)}(J_1) - H^{*(\infty)}(J^*)||_\infty = 0 \tag{3.40}$$

and results in

$$H^\infty(J_1) = J^*. \tag{3.41}$$

3.7.5 Proof of Monotonic Improvement for Policy Iteration

If $\vec{J}_n$ and $\vec{J}_{n+1}$ are sequential value functions in policy iteration, then $\vec{J}_{n+1} \leq \vec{J}_n$.
Proof:
We know $H^*(\vec{J}_n) \leq H^{\pi_n}(\vec{J}_n)$
Let $\pi_{n+1} = \arg\min_a \vec{g} + \gamma T \vec{J}_n$
Then $H^*(\vec{J}_n) = \vec{g}^{\pi_{n+1}} + \gamma T^{\pi_{n+1}} \vec{J}_n \leq \vec{J}_n$
and
$\vec{J}_{n+1} = (I - \gamma T^{\pi_{n+1}})^{-1} \vec{g}^{\pi_{n+1}} \leq \vec{J}_n$

3.7.6 Proof of Policy Iteration Convergence

We know that $J_{n+1} \leq J_n$.
$\mathbb{A}$ and $\mathbb{S}$ are finite, so the algorithm will terminate in a finite number of iterations.
At termination $\pi_{n+1} = \pi_n$, and therefore

$$J_n = J_{n+1} = \min_a \vec{g} + \gamma T \vec{J}_n. \tag{3.42}$$

Bibliography

[1] Russ Tedrake. *Underactuated Robotics*. 2023.

[2] Dimitri Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific, 2019.

[3] Drew Bagnell, Byron Boots, and Sanjiban Choudhury. *Modern Adaptive Control and Reinforcement Learning*. 2022.