Stat 260/CS 294-102. Learning in Sequential Decision Problems.

Peter Bartlett

1. Recall: MDPs.
2. Value iteration.
3. Policy iteration.
4. Linear programming formulation.
5. Q: state-action utility function.

Recall: Markov Decision Processes

Definition: A *Markov Decision Process* (MDP) consists of

1. A state space $\mathcal{X}$,
2. An action space $\mathcal{A}$,
3. A set of Markov chains, $\mathcal{M} = (\mathcal{X}, P_a)$, one for each $a \in \mathcal{A}$,
4. A reward distribution $R : \mathcal{X} \times \mathcal{A} \to \Delta(\mathbb{R})$.

A policy is a sequence of functions $\pi_t : \mathcal{X} \to \Delta(\mathcal{A})$, one for each time t. (A stationary policy is constant with t.)

Recall: Value iteration

Definition: Define the operator $T : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ by

$$(TJ)(x) = \max_{a \in \mathcal{A}} \mathbb{E}\left[r_0 + \alpha J(x_1) \middle| \, x_0 = x, a_0 = a \right].$$

Theorem: For any $\alpha < 1$, there is a vector $J^* \in \mathbb{R}^{\mathcal{X}}$ such that

1. For all $J \in \mathbb{R}^{\mathcal{X}}$, $J^* = \lim_{k \to \infty} T^k J$.
2. J^* is the unique solution to $J = TJ$.
3. $J^* = \max_\pi J^\pi$, where the max is over stationary (or non-stationary) policies π.
4. $J^* = J^{\pi^*}$, where

$$\pi^*(x) = \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[r_0 + \alpha J^*(x_1) \middle| \, x_0 = x, a_0 = a \right].$$

Greedy policy

Notice that π^* is the *greedy choice* with respect to the value function J^*.

Definition: For a value function estimate $\hat{J} \in \mathbb{R}^{\mathcal{X}}$, the corresponding greedy policy is $\pi = G\hat{J}$, where we define the greedy operator $G : \mathbb{R}^{\mathcal{X}} \to \mathcal{A}^{\mathcal{X}}$:

$$(G\hat{J})(x) := \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[r_0 + \alpha \hat{J}(x_1) \,\middle|\, x_0 = x, a_0 = a \right].$$

It's easy to show:

Lemma: For a value function estimate $\hat{J} \in \mathbb{R}^{\mathcal{X}}$, if $\pi = G\hat{J}$,

$$\|J^* - J^\pi\|_\infty \le \frac{2\alpha}{1-\alpha} \|J^* - \hat{J}\|_\infty .$$

Value iteration and (generalized) policy iteration

Value iteration:

$$\hat{J}_{k+1} := T\hat{J}_k, \qquad \pi_{k+1} := G\hat{J}_{k+1}.$$

Policy iteration:

$$\pi_{k+1} := GJ^{\pi_k}.$$

Generalized policy iteration:

$$J_{k+1} := T^l_{\pi_k} J_k, \qquad \pi_{k+1} := GJ_{k+1}.$$

(Generalized) policy iteration

Theorem:

Policy iteration generates a sequence of policies with distinct, increasing values, terminating after a finite number of iterations with an optimal policy, that is, for some k,

$$J^{\pi_0} \leq J^{\pi_1} \leq \cdots \leq J^{\pi_k} = J^*.$$

Generalized policy iteration generates a sequence of policies with $J_k \to J^*$.

Linear program

Bellman equations:

$$J = TJ.$$

Linear programming formulation:

Fix a probability distribution p with support $\mathcal{X}$.

$$\begin{aligned} \min_J \quad & p^T J \\ \text{s.t.} \quad & J \geq TJ. \end{aligned}$$

Linear program

Proof. Uses monotonicity: $J \geq J'$ implies $TJ \geq TJ'$. So $J \geq TJ$ implies $J \geq T^k J \to J^*$. Minimizing $\mu^T J$ sets $J = J^*$. $\square$

Dual linear program

$$
\begin{aligned}
\max_{\mu} \quad & \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \mu(x,a) \mathbb{E}\left[r_0 | x_0 = x, a_0 = a\right] \\
\text{s.t.} \quad & \forall x' \in \mathcal{X},\ \sum_{a \in \mathcal{A}} \mu(x',a) = p(x) \\
& \qquad + \alpha \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \mu(x,a) P[x_1 = x' | x_0 = x, a_0 = a].
\end{aligned}
$$

View λ as discounted expected number of state-action visits, starting from the distribution p. So criterion is expected discounted reward.

Primal-dual are related via optimal policy: $\pi^*(x) = \arg\max_{a \in \mathcal{A}} \lambda(x,a)$.

Q values

Analogous to J^*:

$$Q^*(x,a) := \mathbb{E}\left[r_0 + \alpha \max_{a' \in \mathcal{A}} Q^*(x', a') \middle| x_0 = x, a_0 = a \right],$$

$$\pi^*(x) := \arg\max_{a \in \mathcal{A}} Q^*(x,a).$$

Q iteration:

$$\hat{Q}_{k+1}(x,a) := \mathbb{E}\left[r_0 + \alpha \max_{a' \in \mathcal{A}} \hat{Q}_k(x', a') \middle| x_0 = x, a_0 = a \right],$$

$$\pi_{k+1}(x) := \arg\max_{a \in \mathcal{A}} \hat{Q}_{k+1}(x,a).$$