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

Peter Bartlett

1. Markov decision processes
   and partially observable Markov decision processes.
2. Value functions, Q functions.
3. Finite horizon: dynamic programming.
4. Bellman operator.

Markov Decision Processes

We can think of bandit problems as the simplest example of sequential decision problems, which involve an exploitation/exploration tradeoff. Contextual bandit problems also involve a notion of *state*: the best choice of action depends on the context. But the evolution of the state is out of the control of the strategy. In MDPs, the strategy's actions also influence the state, in a probabilistic way.

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.)

Markov Decision Processes

Examples:

- Inventory control.
- Backgammon.

- Digital marketing.

Partially Observable Markov Decision Processes

Definition: A *Partially Observable Markov Decision Process* (POMDP) consists of

1. An MDP $(\mathcal{X}, \mathcal{A}, P, R)$, and
2. An observation process $\nu : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$, where $\Delta(\mathcal{Y})$ is the set of probability distributions on the observation space $\mathcal{Y}$.

A policy is a function $\pi : \mathcal{Y}^* \rightarrow \Delta(\mathcal{A})$ that maps from observation histories to distributions over actions.

Some Objectives

What is the aim? (Here, $r_t \sim R(x_t, a_t)$.)

1. Maximize total expected reward,

$$J_n(x_0) = \mathbb{E}\left[\sum_{t=0}^{n-1} r_t \,\middle|\, x_0\right].$$

2. Maximize discounted reward,

$$J_\alpha(x_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t r_t \,\middle|\, x_0\right].$$

3. Maximize average reward,

$$J(x_0) = \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{t=0}^{n} r_t \,\middle|\, x_0\right].$$

Finite horizon dynamic programming

Consider a policy $\pi = (\pi_0, \ldots, \pi_{n-1})$.

$$x_0^\pi, a_0^\pi \sim \pi_0(x_0^\pi), r_0^\pi \sim R(x_0^\pi, a_0^\pi), x_1^\pi \sim P_{a_0^\pi}(x_0^\pi),$$
$$\ldots x_t^\pi, a_t^\pi \sim \pi_t(x_t^\pi), r_t^\pi \sim R(x_t^\pi, a_t^\pi), x_{t+1}^\pi \sim P_{a_t^\pi}(x_t^\pi), \ldots$$

Expected total reward of π, starting at x_0:

$$J_n^\pi(x_0) = \mathbb{E}\left[\sum_{t=0}^{n-1} r_t^\pi \,\middle|\, x_0 \right].$$

Optimal reward/policy from x_0:

$$J_n^*(x_0) = \max_\pi J_n^\pi(x_0), \pi^* = \arg\max_\pi J_n^\pi(x_0).$$

Finite horizon dynamic programming

The value to go from x_i, under $\pi = (\pi_i, \ldots, \pi_{n-1})$:

$$J^{\pi}_{i,n}(x_i) = \mathbb{E}\left[\sum_{t=i}^{n-1} r^{\pi}_t \,\middle|\, x_i\right].$$

Bellman's Principle of optimality: For the optimal policy $\pi^* = (\pi^*_0, \ldots, \pi^*_{n-1})$, and for any x_i, however it was reached, the *tail policy* $(\pi^*_i, \ldots, \pi^*_{n-1})$ optimizes the value to go from x_i.

This motivates *dynamic programming*, a backwards induction: find π^*_{n-1}, then π^*_{n-2}, etc.

Finite horizon dynamic programming

First choose π^*_{n-1}:

$$J^*_{n-1,n}(x_{n-1}) = \max_{a_{n-1}\in\mathcal{A}} \mathbb{E}\left[r_{n-1} | x_{n-1},\, a_{n-1}\right].$$

Then choose π^*_{n-2}:

$$J^*_{n-2,n}(x_{n-2}) = \max_{a_{n-2}\in\mathcal{A}} \mathbb{E}\left[r_{n-2} + J^*_{n-1,n}(x_{n-1}) \Big| \, x_{n-2}, a_{n-2}\right].$$

Finite horizon dynamic programming: T

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

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

Then the optimal value is given by $J_n^* = J_{0,n}^*$ where

$$\begin{aligned} J_{n,n}^*(x) &= 0, \\ J_{t,n}^* &= T J_{t+1,n}^*. \end{aligned}$$

Finite horizon policy evaluation: T_π

Similarly, to compute J_n^π: (e.g.: π stationary)

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

$$(T_\pi J)(x) = \mathbb{E}\left[r_0 + J(x_1) \middle| x_0 = x, a_0 = \pi(x_0) \right].$$

Then the value under π is given by $J_n^\pi = J_{0,n}^\pi$ where

$$J_{n,n}^\pi(x) = 0,$$
$$J_{t,n}^\pi = T_\pi J_{t+1,n}^\pi.$$

Infinite horizon discounted reward

$$J(x_0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \alpha^t r_t \middle| x_0\right].$$

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

$$(T_\pi J)(x) = \mathbb{E}\left[r_0 + \alpha J(x_1) \middle| x_0 = x, a_0 = \pi(x_0)\right].$$

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

1. For all $J \in \mathbb{R}^{\mathcal{X}}$, $J^\pi = \lim_{k\to\infty} T_\pi^k J$.
2. J^π is the unique solution to $J = T_\pi J$.

Infinite horizon optimal policy: 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

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

Infinite horizon discounted reward

Lemma: T and T_π are contractions:

$$\|TJ - TJ'\|_\infty \leq \alpha \|J - J'\|_\infty ,$$
$$\|T_\pi J - T_\pi J'\|_\infty \leq \alpha \|J - J'\|_\infty .$$

This follows from:

1. $J \leq J'$ implies $TJ \leq TJ'$ and $T_\pi J \leq T_\pi J'$.
2. For all $c \in \mathbb{R}$, $T^k(J + c\mathbb{1}) \leq TJ + \alpha^k c\mathbb{1}$ and $T_\pi^k(J + c\mathbb{1}) \leq T_\pi J + \alpha^k c\mathbb{1}$.