1. Lecture outline

- Review.
- Policy iteration.
- Function approximation.

2. Review

Value of a policy, $\pi$,

$$V_\pi(s) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R(s_t) | s_0 = s; \pi \right]$$

$$V^*(s) = \max_{\pi} V_\pi(s)$$

Definition of the Bellman backup operator, $T$,

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

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

$\exists \pi^* = (\mu^*, \mu^*, \ldots)$ s.t. $\pi^* \in \arg\max_{\pi} (V_\pi(s))$ and $V_{\pi^*} = V^*$

where $\mu^*(s) \in \arg\max_{a \in A} \left[ R(s) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \right]$

For a fixed stationary policy $\pi = (\mu, \mu, \ldots)$,

$$(T_\mu V)(s) = R(s) + \gamma \sum_{s'} P(s'|s,\mu(s)) V(s')$$

$$\lim_{H \to \infty} (T^H_\mu V) = V_\pi$$

$T$ is a $\gamma$-contraction with respect to the $\infty$-norm, i.e.,

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

$T$ is a contraction $\Rightarrow T^H V$ converges to a unique fixed point from any starting point $V$. 
3 Policy iteration

Pick a policy, \( \pi(0) = (\mu(0), \mu(0), \ldots) \), for \( i = 0, 1, \ldots \)

\[
V_{\pi(i)} = \lim_{H \to \infty} \left( T_{\mu(i)}^H V \right)
\]

\[
\mu^{(i+1)}(s) \in \arg \max_{a \in A} \left[ R(s) + \gamma \sum_{s'} P(s'|s,a) V_{\pi(i)}(s') \right]
\]

From the above definition of \( \mu^{(i+1)} \) we have:

\[
V_{\pi(i)} \geq V_{\pi(i)}
\]

However, can we say something to the extent:

\[
V_{\mu^{(i+1)}, \mu^{(i+1)}, \ldots} \geq V_{\mu^{(i)}, \mu^{(i)}, \ldots}
\]

Let’s assume for the time being that the above is true, i.e.,

\[
V_{\pi(i+1)} \geq V_{\pi(i)} \geq V_{\pi(i-1)} \geq \ldots \geq V_{\pi(0)}
\]

There are two cases. It can either be a strictly better policy or an equally good policy. For a policy to be strictly better, it has to be different from all previous policies. Hence this can happen only \(|A|^{|S|}\) times.

In the other case, i.e.,

\[
V_{\pi(i+1)} = V_{\pi(i)}
\]

we have the following: First note that, by definition of \( \mu^{(i+1)} \) we have

\[
T_{\mu^{(i+1)}} V_{\pi(i)} = TV_{\pi(i)}
\]

Combining Eqn. (3) and (4) gives us:

\[
T_{\mu^{(i+1)}} V_{\pi(i+1)} = TV_{\pi(i+1)}
\]

Taking into account that \( V_{\pi(i+1)} \) is the fixed point of \( T_{\mu^{(i+1)}} \), i.e., \( T_{\mu^{(i+1)}} V_{\pi(i+1)} = V_{\pi(i+1)} \) we get that:

\[
V_{\pi(i+1)} = TV_{\pi(i+1)}
\]

Hence \( V_{\pi(i+1)} \) is the unique fixed point of \( T \).

So we have shown that policy iteration will converge to an optimal policy after at most \(|A|^{|S|}\) iterations in the assumption that Eqn. (1) holds true.

**Proposition 1 Monotonicity**

\[
V_1 \geq V_2 \rightarrow TV_1 \geq TV_2;
\]

also, as a special case: \( \forall \mu : V_1 \geq V_2 \rightarrow T_{\mu} V_1 \geq T_{\mu} V_2 \).

**Proof.** The Bellman operator satisfies the monotonicity property:

\[
(TV_1)(s) = \max_{a \in A} \left[ R(s) + \gamma \sum_{s'} P(s'|s,a) V_1(s') \right] \geq \max_{a \in A} \left[ R(s) + \gamma \sum_{s'} P(s'|s,a) V_2(s') \right] = (TV_2)(s)
\]

where the inequality appears from the fact that \( V_1 \geq V_2 \). [qed].
We’ll now apply monotonicity to show that Eqn. (1) holds:
\[
T_{\mu(i+1)}V_{\pi(i)} \geq T_{\mu(i)} V_{\pi(i)}
\]
\[
\Rightarrow T_{\mu(i+1)}V_{\pi(i)} \geq V_{\pi(i)}
\]
\[
\Rightarrow T_{\mu(i+1)}T_{\mu(i+1)}V_{\pi(i)} \geq T_{\mu(i+1)}V_{\pi(i)}
\]
\[
\Rightarrow T_{\mu(i+1)}T_{\mu(i+1)}T_{\mu(i+1)}V_{\pi(i)} \geq T_{\mu(i+1)}T_{\mu(i+1)}V_{\pi(i)}
\]
\[
\Rightarrow V_{\pi(i+1)} \geq \ldots \geq T_{\mu(i+1)}V_{\pi(i)} \geq \ldots \geq V_{\pi(i)}
\]

Policy iteration works remarkably well in practice. However, a complete satisfactory explanation is still an open problem. For example, the best bound known for the iterations needed is exponential in the number of states, \(|s|\), but the worst case examples known involve a fairly small number of iterations, on the order of \(|s|\).

4 Function approximation

![Figure 1: Linear data fit.](image)

Given \((x(0), y(0)), (x(1), y(1)), \ldots, \) and the following setup:
\[
\begin{pmatrix}
y(0) \\
y(1) \\
y(2) \\
y(3) \\
y(4) \\
y(5)
\end{pmatrix} =
\begin{pmatrix}
1 & x(0) \\
1 & x(1) \\
1 & x(2) \\
1 & x(3) \\
1 & x(4) \\
1 & x(5)
\end{pmatrix} \cdot \begin{pmatrix}
\theta_0 \\
\theta_1
\end{pmatrix} \rightarrow y = X\theta
\]

one can find \(\theta\) using least squares.

\[
\min_\theta \|y - X\theta\|^2 = \min_\theta (y - X\theta)^T (y - X\theta)
\]
\[
= \min_\theta (y^T y - 2y^T X\theta + \theta^T X^T X\theta)
\]
Noting that $\nabla \theta = -2X^Ty + 2X^TX\theta = 0$, $X^Ty = X^TX\theta \Rightarrow \theta = (X^TX)^{-1}X^Ty$

One could also use weighted least squares:

$$\min_\theta (y - X\theta)^TW(y - X\theta) \text{ where } W \succeq 0 \Rightarrow \theta = (X^TWX)^{-1}X^TWy$$

**Figure 2**: Least squares as a projection onto basis.

We could extend this concept to value iteration, $V_{k+1} = TV_k$, using this slightly different notation:

$$V_{k+1} = \Phi\theta_{k+1} = \Pi T\Phi\theta_k$$

where $\Pi$ symbolizes the projection onto the basis (see figure 2):

$$\Pi f = \arg\min_{\Phi\theta} \|f - \Phi\theta\|_2$$

We don’t have a guarantee that this projection is an $\infty$-norm contraction, i.e., we have no guarantees of the following form,

$$\|\Pi V - \Pi V\|_\infty \leq \|V - V\|_\infty$$

which, in other words, means that we don’t know if it converges.

In fact, merely considering the fact that $T$ is a $\gamma$-contraction with respect to the infinity norm, and the fact that $\Pi$ is a (weighted) orthogonal projection we could cook up the diverging scenario pictured in Figure 3.

Can this really happen to a Markov decision process or did we ignore some properties of a Markov decision process that prevent this from happening?

It turns out this particular scenario can happen: it happens for the following MDP: Consider the autonomous (one action only) Markov chain depicted in Figure 4. We set $\gamma \in (0, 1)$ and all rewards to zero, i.e., $R(1) = R(2) = 0$, hence $V^* = (0, 0)^T$. Let $\Phi = (1, 2)^T$ form the basis for our approximations, so all approximations of the value function take the form $\Phi\theta$. An update then yields,

$$\Phi\theta_{k+1} = \Pi T\Phi\theta_k$$

$$\Pi V = \arg\min_{\Phi\theta} \|V - \Phi\theta\|_2$$

$$(TV)(i) = \gamma\epsilon V(1) + \gamma(1 - c)V(2) \text{ for } i = 1, 2$$
Figure 3: An example of repeated application of the Bellman backup followed by a least-squares projection.

Figure 4: Markov chain diagram.
Hence

\[(T\Phi\theta_k)(i) = \gamma(2 - \epsilon)\theta_k\]

\[\theta_{k+1} = \arg \min_{\theta} \left( (\theta - (T\Phi\theta_k)(1))^2 + (2\theta - (T\Phi\theta_k)(2))^2 \right) = \frac{3}{5}\gamma(2 - \epsilon)\theta_k\]

If \(\epsilon \approx 0\) and \(\gamma \approx 1\) then \(\theta_k\) grows unbounded.