FoSML: Lecture #23

- In a realistic setting, an agent in a RL framework does not have a priori knowledge of the environment.
- We can consider a simplified setting, however, in which both the rules of transition and reward system are known to the agent.
- More precisely, we may suppose that the following are known: $\mathcal{S}$
  - (a.) $\mathbb{P}(s' \mid s,a)$, i.e. the probability of transitioning to state s', given that the agent's present state is $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ is taken.
  - (b.) $\mathbb{E}(r(s,a))$, i.e. the expected reward from taking action a in state s.
- Since we don't have to estimate these environmental parameters, the RL framework reduces to a planning problem in this setting.
- We consider three methods for solving this planning problem:
  - (a.) Value iteration → Use the Bellman equations and iteratively determine $V^*(s), \forall s \in \mathcal{S}$
  - (b.) Policy iteration → Use policy evaluation and our matrix inversion result from last lecture
  - (c.) Linear programming → Formulate the problem as maximizing a linear objective subject to linear constraints, then solve with, e.g., Simplex method.

Value Iteration

- Let $V \in \mathbb{R}^{|\mathcal{S}|}$ and let $\Phi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ be defined as in the Bellman equations:

$$\forall s \in \mathcal{S}, \quad \Phi(V)(s) = \max_{a \in \mathcal{A}} \left(\mathbb{E}(r(s,a)) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s,a) \cdot V(s') \right).$$

- This implicitly defines a policy, and we can thus re-write

$$\Phi(V) = \max_{\pi} \left(R_\pi + \gamma P_\pi V \right), \quad \text{(★)}$$

where $P_\pi \in \mathbb{R}^{|S| \times |S|}$ is defined as $(P_\pi)_{ss'} = \mathbb{P}(s' \mid s, \pi(s))$, for all $s, s' \in S$, and where $R_\pi \in \mathbb{R}^{|S|}$ is defined as $(R_\pi)_s = \mathbb{E}(r(s, \pi(s)))$, for all $s \in S$.

- The main idea of value iteration is to iteratively apply Φ to an arbitrary starting point until some convergence criterion is achieved. That is, let $V_0 \in \mathbb{R}^{|S|}$ be arbitrary, and set $V_{i+1} = \Phi(V_i)$, $i \geq 0$, until the stopping condition $\|V_i - \Phi(V_i)\| < \frac{(1-\gamma)}{\gamma} \cdot \varepsilon$ is reached for some $\varepsilon > 0$.

- This converges (though perhaps slowly) to the optimal value:

Theorem (convergence of value iteration): Let V^* be the value function associated to an optimal π^*. Then for any initial guess V_0, the sequence $V_{i+1} = \Phi(V_i)$, $i \geq 0$ converges to V^*.

Proof: Note that it is sufficient to show Φ is contracting, i.e. has a Lipschitz constant < 1 wrt the $\|\cdot\|_\infty$ topology. In fact, we will show Φ is $\gamma (< 1)$ Lipschitz. Indeed, if we show this, then

$$\begin{aligned} &\|V^* - V_{n+1}\|_\infty \\ &= \|\Phi(V^*) - V_{n+1}\|_\infty, \quad \text{since } \Phi(V^*) = V^* \text{ by the Bellman conditions} \\ &= \|\Phi(V^*) - \Phi(V_n)\|_\infty \end{aligned}$$

$\leq \gamma \| V^* - V_n \|_\infty$.

Inducting, we see $\| V^* - V_{n+1} \|_\infty \leq \gamma^{n+1} \cdot \| V^* - V_0 \|_\infty$; taking $n \to \infty$ gives convergence.

It remains to show Φ is γ-Lipschitz. Well, for any $s \in S$, and any $V \in \mathbb{R}^{|S|}$, let $a^*(s)$ be the action maximizing $R_{\pi} + \gamma P_{\pi} V$, over which defines $\Phi(V)(s)$. Then $\forall s \in S$ and $\forall U \in \mathbb{R}^{|S|}$, we have

$$\Phi(V)(s) - \Phi(U)(s) \leq \Phi(V)(s) - \left(\mathbb{E}(r(s, a^*(s)) + \gamma \sum_{s' \in S} \mathbb{P}(s' \mid s, a^*(s)) \cdot U(s') \right)$$

$$= \gamma \sum_{s' \in S} \mathbb{P}(s' \mid s, a^*(s)) \cdot \left(V(s') - U(s') \right)$$

$$\leq \gamma \sum_{s' \in S} \mathbb{P}(s' \mid s, a^*(s)) \cdot \| V - U \|_\infty$$

$$= \gamma \cdot \| V - U \|_\infty \cdot \sum_{s' \in S} \mathbb{P}(s' \mid s, a^*(s))$$

$$= \gamma \cdot \| V - U \|_\infty,$$

as desired. ∎

↑ gets ε-close to optimality, and

Note that the proposed algorithm converges in $O(\log \frac{1}{\varepsilon})$ iterations:

$$\| V^* - V_{n+1} \|_\infty \leq \| V^* - \Phi(V_{n+1}) \|_\infty + \| \Phi(V_{n+1}) - V_{n+1} \|_\infty$$

$$= \| \Phi(V^*) - \Phi(V_{n+1}) \|_\infty + \| \Phi(V_{n+1}) - \Phi(V_n) \|_\infty$$

γ-Lipschitz $\longrightarrow$

$$\leq \gamma \cdot \| V^* - V_{n+1} \|_\infty + \gamma \cdot \| V_{n+1} - V_n \|_\infty$$

$$\Rightarrow (1 - \gamma) \cdot \| V^* - V_{n+1} \|_\infty \leq \gamma \cdot \| V_{n+1} - V_n \|_\infty$$

$$\Rightarrow \|V^* - V_{n+1}\|_\infty \le \frac{\gamma}{1-\gamma}\|V_{n+1} - V_n\|_\infty$$

$$\le \frac{\gamma}{1-\gamma}\cdot\frac{1-\gamma}{\gamma}\cdot\varepsilon \qquad \text{by our stopping condition}$$

$$\le \varepsilon.$$

To see convergence, note that if n is largest integer s.t. $\frac{(1-\gamma)\varepsilon}{\gamma} \le \|V_{n+1} - V_n\|_\infty$, i.e. if the algorithm converges on the $(n+1)^{st}$ step, then

$$\frac{(1-\gamma)\varepsilon}{\gamma} \le \|V_{n+1} - V_n\|_\infty \le \gamma^n \cdot \|\Phi(V_0) - V_0\|_\infty$$

$$\Rightarrow \frac{(1-\gamma)\varepsilon}{\gamma}\cdot\|\Phi(V_0) - V_0\|_\infty \le \gamma^n$$

$$\Rightarrow \log\left(\frac{1-\gamma}{\gamma}\right) + \log(\varepsilon) + \log\left(\|\Phi(V_0) - V_0\|_\infty\right) \le n\cdot\log(\gamma)$$

$$O\left(\log\left(\tfrac{1}{\varepsilon}\right)\right) \ge n.$$

$\Rightarrow$

Policy Iteration

- Instead of iterating on V, we can iterate on π. That is, we can pick an arbitrary initial $\pi = \pi_0$, then solve $(I - \gamma P_\pi)V = R_\pi$ to get V, then set $\pi = \arg\max_\pi (R_\pi + \gamma P_\pi V)$. Iterating until some kind of criterion for stopping is met yields an approximately optimal π.
- Indeed, let $\{V_n\}_{n=0}^\infty$ be the V sequence induced by the above scheme.

Thm (Convergence of Policy Iteration): Let $\{V_n\}_{n=1}^{\infty}$ be the sequence of policy values computed above. Then $\forall n$, $V_n \leq V_{n+1} \leq V^*$.

Proof: Let $\hat{\pi}_{n+1}$ be the policy improvement at the n^{th} iteration of the algorithm. We claim that if $(Y-X) \geq 0$, for any $Y, X \in \mathbb{R}^{|S|}$, then

$$(I - \gamma P_{\hat{\pi}_{n+1}})^{-1}(Y-X) \geq 0 \quad (\star)$$

also. Well, we know from $P_{\hat{\pi}_{n+1}}$ Markov that $\|\gamma P_{\hat{\pi}_{n+1}}\|_\infty < 1$, so that we may expand in Neumann series:

$$(I - \gamma P_{\hat{\pi}_{n+1}})^{-1} = \sum_{k=0}^{\infty} (\gamma P_{\hat{\pi}_{n+1}})^k$$

noting that $P_{\hat{\pi}_{n+1}}$ is pointwise non-negative yields $(\star)$.

Now, by definition of $\hat{\pi}_{n+1}$, we have

$$\underbrace{R_{\hat{\pi}_{n+1}} + \gamma P_{\hat{\pi}_{n+1}} V_n}_{\text{by optimality of } \hat{\pi}_{n+1}} \geq \underbrace{R_{\hat{\pi}_n} + \gamma P_{\hat{\pi}_n} V_n}_{\text{by construction}} = V_n$$

In particular,

$$R_{\hat{\pi}_{n+1}} \geq (I - \gamma P_{\hat{\pi}_{n+1}}) V_n$$

$$\overset{(\star)}{\Rightarrow} \underbrace{(I - \gamma P_{\hat{\pi}_{n+1}})^{-1} R_{\hat{\pi}_{n+1}}}_{\| \; V_{n+1}} \geq V_n$$

Noting the V^* is maximal by definition gives the result.

- If we stop once $\pi_n = \pi_{n+1}$, we may have to iterate over all $|A|^{|S|}$ policies. In practice, bounds of the form $O\left(\frac{|A|^{|S|}}{|S|}\right)$ are possible.
- Policy & value iteration are closely related.

Theorem: Let $\{U_n\}_{n=0}^{\infty}$ be the values generated from value iteration, $\{V_n\}_{n=0}^{\infty}$ those from policy iteration. Then if $U_0 = V_0$, $U_n \leq V_n \leq V^*$, for all n.

Proof: We first show that the $\Phi(V) = \max_{\pi} (R_\pi + \gamma P_\pi V)$ function is monotonic. Indeed, let U, V be s.t. $U \leq V$ and let π be the policy such that $\Phi(U) = R_\pi + \gamma P_\pi U$. Then:

$$\begin{aligned}
\Phi(U) &= R_{\hat{\pi}} + \gamma P_{\hat{\pi}} U \\
&\leq R_{\hat{\pi}} + \gamma P_{\hat{\pi}} V \\
&\leq \max_{\pi} \left\{ R_\pi + \gamma P_\pi V \right\} \\
&= \Phi(V).
\end{aligned}$$

The main result follows by induction. The base case holds by assumption. Then, suppose $U_n \leq V_n$. By monotonicity of Φ,

$$\begin{aligned}
U_{n+1} &= \Phi(U_n) \\
&\leq \Phi(V_n) \\
&= \max_{\pi} \left\{ R_\pi + \gamma P_\pi V_n \right\}.
\end{aligned}$$

Letting $\tilde{\pi}_{n+1}$ be the policy with $\pi_{n+1} = \arg\max_{\tilde{\pi}} \{R_{\tilde{\pi}} + \gamma P_{\tilde{\pi}} V_n\}$, we have

$$\Phi(V_n) = R_{\tilde{\pi}_{n+1}} + \gamma P_{\tilde{\pi}_{n+1}} V_n$$

$$\leq R_{\tilde{\pi}_{n+1}} + \gamma P_{\tilde{\pi}_{n+1}} V_{n+1}$$

$$= V_{n+1},$$

which gives the result.