CMPT 419/983: Theoretical Foundations of Reinforcement Learning

Lecture 5

Sharan Vaswani

October 6, 2023

Recap

- **Bellman equation for policy** π: $v^\pi(s) = \mathbf{r}_\pi(s) + \gamma \sum_{s'} \mathbf{P}_\pi[s, s']\, v^\pi(s')$
$= \sum_{a \in \mathcal{A}} r(s, a)\, \pi[a|s] + \gamma \sum_{s' \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{P}[s'|s, a]\, \pi[a|s]\, v^\pi(s')$.
- **Bellman Optimality**: $\mathcal{T} : \mathbb{R}^S \to \mathbb{R}^S$ s.t. $(\mathcal{T}u)(s) = \max_a \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a) u(s')\}$.
- **Fundamental Theorem**: For policy $\pi^* \in \Pi_{\mathsf{SD}}$, $v^{\pi^*}(s) = \max_{\pi \in \Pi_{\mathbf{HR}}} v^\pi(s)$ for all $s \in \mathcal{S}$.
- $v^* = \mathcal{T} v^* = \max_{\pi \in \Pi_{\mathbf{SD}}} \{\mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^*\} = \mathcal{T}_{\pi^*} v^* = \mathbf{r}_{\pi^*} + \gamma \mathbf{P}_{\pi^*} v^*$
- **Value Iteration**: Iterate $v_k = \mathcal{T} v_{k-1}$ for K iterations. $\forall s \in \mathcal{S}$, return the greedy policy w.r.t v_K i.e. $\hat{\pi}(s) = \arg\max_a \{r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a)\, v_K(s')\}$.
- **VI convergence**: After $K \geq \frac{\log(1/\epsilon\,(1-\gamma))}{1-\gamma}$ iterations, VI returns a v_K s.t. $\|v_K - v^*\|_\infty \leq \epsilon$.
- Since $\hat{\pi}$ is the policy returned by VI, we want a bound on $\left\|v^* - v^{\hat{\pi}}\right\|_\infty$.
- Today, we will prove that VI requires $K \geq \frac{\log(2\gamma/\epsilon\,(1-\gamma)^2)}{1-\gamma}$ iterations to ensure $\left\|v^* - v^{\hat{\pi}}\right\|_\infty \leq \epsilon$.

Policy Error Bound

Claim: For an arbitrary $v \in \mathbb{R}^S$ if (i) π is the greedy policy w.r.t v, i.e. $\pi(s) = \arg\max_a \{r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\, v(s')\}$, (ii) v^π is the value function corresponding to policy π i.e. $v^\pi = \mathcal{T}_\pi v^\pi = \mathbf{r}_\pi + \gamma \mathbf{P}_\pi v^\pi$, then,

$$v^\pi \geq v^* - \frac{2\gamma \left\| v - v^* \right\|_\infty}{1-\gamma} \mathbf{1}$$

Proof: For the proof, we need the following properties of the $\mathcal{T}$ and $\mathcal{T}_\pi$ operators.

$$\mathcal{T} v^* = v^* \quad ; \quad \mathcal{T} v = \mathcal{T}_\pi v \quad ; \quad v^\pi = \mathcal{T}_\pi v^\pi$$

We will also need the following properties: for $u, w \in \mathbb{R}^S$ s.t. $u \leq w$ (element-wise) and a constant c,

$$\mathcal{T}(u) \leq \mathcal{T}(w) \quad ; \quad \mathcal{T}_\pi(u) \leq \mathcal{T}_\pi(w) \qquad \text{(Monotonicity)}$$

$$\mathcal{T}(u + c\mathbf{1}) = \mathcal{T}(u) + c\gamma\, \mathbf{1} \quad ; \quad \mathcal{T}_\pi(u + c\mathbf{1}) = \mathcal{T}_\pi(u) + c\gamma\, \mathbf{1} \qquad \text{(Additivity)}$$

Prove in Assignment 2!

Policy Error Bound

Define $\epsilon := \|v^* - v\|_\infty \implies -\epsilon\mathbf{1} \le v^* - v \le \epsilon\mathbf{1}$ and define $\delta := v^* - v^\pi$.

$$
\begin{aligned}
\delta = v^* - v^\pi &= \mathcal{T}v^* - v^\pi = \mathcal{T}v^* - \mathcal{T}_\pi v^\pi && \text{(By definitions of } \mathcal{T}, \mathcal{T}_\pi) \\
&\le \mathcal{T}(v + \epsilon\mathbf{1}) - \mathcal{T}_\pi v^\pi = \mathcal{T}v + \epsilon\gamma\mathbf{1} - \mathcal{T}_\pi v^\pi && \text{(By monotonicity, additivity of } \mathcal{T}) \\
&= \mathcal{T}_\pi v + \epsilon\gamma\mathbf{1} - \mathcal{T}_\pi v^\pi && \text{(Since } \mathcal{T}v = \mathcal{T}_\pi v) \\
&\le \mathcal{T}_\pi(v^* + \epsilon\mathbf{1}) + \epsilon\gamma\mathbf{1} - \mathcal{T}_\pi v^\pi = \mathcal{T}_\pi v^* + \gamma\epsilon\mathbf{1} + \epsilon\gamma\mathbf{1} - \mathcal{T}_\pi v^\pi \\
& && \text{(By monotonicity, additivity of } \mathcal{T}_\pi) \\
&= \mathcal{T}_\pi v^* - \mathcal{T}_\pi v^\pi + 2\gamma\epsilon\mathbf{1} \\
&= [\mathbf{r}_\pi + \gamma\mathbf{P}_\pi v^*] - [\mathbf{r}_\pi + \gamma\mathbf{P}_\pi v^\pi] + 2\gamma\epsilon\mathbf{1} && \text{(By definition of } \mathcal{T}_\pi) \\
&= \gamma\mathbf{P}_\pi(v^* - v^\pi) + 2\gamma\epsilon\mathbf{1} \\
\implies \delta &\le \gamma\mathbf{P}_\pi\delta + 2\gamma\epsilon\mathbf{1} \\
\implies |\delta| &\le \gamma|\mathbf{P}_\pi\delta| + 2\gamma\epsilon\mathbf{1} \\
& && \text{(Taking an element-wise absolute value and using the triangle inequality)}
\end{aligned}
$$

Policy Error Bound

Recall that $\epsilon = \|v^* - v\|_\infty$, $\delta := v^* - v^\pi$ and $|\delta| \leq \gamma |\mathbf{P}_\pi \delta| + 2\gamma\epsilon\mathbf{1}$ Let us simplify $|\mathbf{P}_\pi \delta|$. For an arbitrary s,

$$
\begin{aligned}
|\mathbf{P}_\pi \delta|(s) &= \left| \sum_{s'} \mathbf{P}_\pi(s, s')\delta(s') \right| \leq \sum_{s'} |\mathbf{P}_\pi(s, s')\delta(s')| = \sum_{s'} \mathbf{P}_\pi(s, s')|\delta(s')| \\
&\leq \|\delta\|_\infty \sum_{s'} \mathbf{P}_\pi(s, s') = \|\delta\|_\infty
\end{aligned}
$$

$$
\implies |\mathbf{P}_\pi \delta| \leq \|\delta\|_\infty \mathbf{1} \implies |\delta| \leq \gamma \|\delta\|_\infty \mathbf{1} + 2\gamma\epsilon\mathbf{1}
$$

$$
\implies \|\delta\|_\infty \leq \gamma \|\delta\|_\infty + 2\gamma\epsilon \implies \|\delta\|_\infty \leq \frac{2\gamma\epsilon}{1-\gamma}
$$

(By taking the element-wise maximum on both sides)

$$
\implies \|v^* - v^\pi\|_\infty \leq \frac{2\gamma \|v^* - v\|_\infty}{1-\gamma} \implies v^\pi \geq v^* - \frac{2\gamma \|v - v^*\|_\infty}{1-\gamma}\mathbf{1} \quad \square
$$

Policy Iteration

Policy Iteration

Algorithm Policy Iteration

1: **Input**: MDP $M = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho)$, π_0.
2: **for** $k = 0 \to K$ **do**
3: **Policy Evaluation**: Calculate v^{π_k} as the solution to $(I - \gamma \mathbf{P}_{\pi_k})v = \mathbf{r}_{\pi_k}$.
4: **Policy Improvement**: $\forall s$, $\pi_{k+1}(s) = \arg\max_a \{r(s,a) + \gamma \sum_{s'} \mathcal{P}(s'|s,a)\, v^{\pi_k}(s')\}$
5: **end for**

- Computational Complexity: $O((S^3 + S^2 A)\, K)$
- We will prove that $K = O\left(\frac{SA}{1-\gamma}\right)$ iterations of PI are sufficient to ensure exact convergence to the optimal policy. Hence, PI requires $O\left(\frac{S^4 A + S^3 A^2}{1-\gamma}\right)$ operations.

We will do the proof in two steps:

(i) Show that the sequence of v^{π_k} converges to v^* at a linear rate (similar to VI).
(ii) Relate v^{π_k} to the greedy policy chosen by PI at each iteration.

Policy Iteration

(i) Claim: For PI, $\|v^{\pi_K} - v^*\|_\infty \leq \gamma^K \|v^{\pi_0} - v^*\|_\infty$.

Proof: We will first prove a more general result: for any π, π', if π' is the greedy policy w.r.t v^π, then, $v^\pi \leq \mathcal{T} v^\pi \leq v^{\pi'}$. To see this, note that,

$$\mathcal{T} v^\pi = \mathcal{T}_{\pi'} v^\pi \quad ; \quad v^\pi = \mathcal{T}_\pi v^\pi \leq \mathcal{T} v^\pi \quad \text{(By definition of } \pi' \text{ and by definitions of } \mathcal{T} \text{ and } \mathcal{T}_\pi\text{)}$$

We will use induction to show that $v^\pi \leq \mathcal{T} v^\pi \leq \mathcal{T}^n_{\pi'} v^\pi$ for all n. As $n \to \infty$, $v^\pi \leq \mathcal{T} v^\pi \leq v^{\pi'}$.

Base Case: For $n = 1$, from the above definition, we know that $v^\pi \leq \mathcal{T} v^\pi = \mathcal{T}_{\pi'} v^\pi$.

Inductive Hypothesis: Assume that $v^\pi \leq \mathcal{T} v^\pi \leq \mathcal{T}^{n-1}_{\pi'} v^\pi$. Let us prove it for n,

$$v^\pi \leq \mathcal{T}^{n-1}_{\pi'} v^\pi \implies \mathcal{T}_{\pi'} v^\pi \leq \mathcal{T}^n_{\pi'} v^\pi \implies \mathcal{T} v^\pi \leq \mathcal{T}^n_{\pi'} v^\pi \implies v^\pi \leq \mathcal{T} v^\pi \leq \mathcal{T}^n_{\pi'} v^\pi$$

Using this result for PI, we get that $v^{\pi_k} \leq \mathcal{T} v^{\pi_k} \leq v^{\pi_{k+1}}$. Using this result recursively,

$$\mathcal{T} v^{\pi_0} \leq v^{\pi_1} \implies \mathcal{T}^2 v^{\pi_0} \leq \mathcal{T} v^{\pi_1} \leq v^{\pi_2} \implies \mathcal{T}^K v^{\pi_0} \leq v^{\pi_K}$$

Policy Iteration

Recall we have proved that $\mathcal{T}^K v^{\pi_0} \leq v^{\pi_K}$. Since v^* is the optimal value function,

$$
\begin{aligned}
\mathcal{T}^K v^{\pi_0} \leq v^{\pi_K} \leq v^* &\implies v^* - v^{\pi_K} \leq v^* - \mathcal{T}^K v^{\pi_0} \\
\implies \left\| v^* - v^{\pi_K} \right\|_\infty &\leq \left\| v^* - \mathcal{T}^K v^{\pi_0} \right\|_\infty \\
\implies \left\| v^* - v^{\pi_K} \right\|_\infty &\leq \left\| \mathcal{T}^K v^* - \mathcal{T}^K v^{\pi_0} \right\|_\infty \leq \gamma^K \left\| v^* - v^{\pi_0} \right\|_\infty \quad \square
\end{aligned}
$$

For proving **(ii)**, we will require an intermediate result – the *value difference lemma*.

Claim: For any $\pi, \pi' \in \Pi_{\text{SR}}$, $v^{\pi'} - v^{\pi} = \left(I - \gamma \mathbf{P}_{\pi'}\right)^{-1} g(\pi', \pi)$ where $g(\pi', \pi) := \mathcal{T}_{\pi'} v^{\pi} - v^{\pi}$.
Proof: Recall that $v^{\pi'} = \left(I - \gamma \mathbf{P}_{\pi'}\right)^{-1} \mathbf{r}_{\pi'}$.

$$
\begin{aligned}
v^{\pi'} - v^{\pi} &= \left(I - \gamma \mathbf{P}_{\pi'}\right)^{-1} \mathbf{r}_{\pi'} - v^{\pi} = \left(I - \gamma \mathbf{P}_{\pi'}\right)^{-1} \left[\mathbf{r}_{\pi'} - \left(I - \gamma \mathbf{P}_{\pi'}\right) v^{\pi}\right] \\
&= \left(I - \gamma \mathbf{P}_{\pi'}\right)^{-1} \left[\left(\mathbf{r}_{\pi'} + \gamma \mathbf{P}_{\pi'} v^{\pi}\right) - v^{\pi}\right] = \left(I - \gamma \mathbf{P}_{\pi'}\right)^{-1} \left[\mathcal{T}_{\pi'} v^{\pi} - v^{\pi}\right] \\
&= \left(I - \gamma \mathbf{P}_{\pi'}\right)^{-1} g(\pi', \pi) \quad \square
\end{aligned}
$$

Policy Iteration

Claim: Consider an arbitrary sub-optimal stationary deterministic policy π'_0 and define π'_K to be the policy returned by PI after K iterations starting from policy π'_0. For all $K \geq K^* := \lceil \frac{\log(1/1-\gamma)}{\log(1/\gamma)} \rceil + 1$, there exists a state s' such that $\pi'_K[s'] \neq \pi'_0[s']$. This means that for all $K \geq K^*$, the action corresponding to $\pi'_0[s']$ is *eliminated* for state s'.

We will use this claim multiple times starting from $\pi'_0 = \pi_0$. In particular,

- After $K \geq K^*$ iterations of PI, we know there exists a state s' for which the action corresponding to $\pi_0[s']$ is eliminated.
- If we continue running PI, after a further K^* iterations, another action would be eliminated. Specifically, for $\pi'_0 = \pi_{K^*}$, there exists a state s'' for which the action corresponding to $\pi_{K^*}[s'']$ is eliminated.
- Since we are considering deterministic policies, we need to eliminate at most $SA - S$ actions, and need to run PI for at most $(SA - S)\, K^*$ iterations. Hence, PI will converge to the optimal policy in $O\left(\frac{S\,A\,\log(1/1-\gamma)}{1-\gamma}\right)$ iterations.

Policy Iteration

Proof: We will make use of the value difference lemma to bound $g(\pi, \pi^*)$. Note that $g(\pi, \pi^*) = \mathcal{T}_\pi v^* - v^* < 0$ for all sub-optimal policies π.

$$-g(\pi'_K, \pi^*) = \left(I - \gamma \mathbf{P}_{\pi'_K}\right)[v^* - v^{\pi'_K}] = [v^* - v^{\pi'_K}] - \gamma \mathbf{P}_{\pi'_K} \underbrace{[v^* - v^{\pi'_K}]}_{\text{Non-negative}} \leq [v^* - v^{\pi'_K}]$$

$$\implies \|g(\pi'_K, \pi^*)\|_\infty \leq \left\| v^* - v^{\pi'_K} \right\|_\infty$$

(Taking element-wise absolute value and max over the states)

$$\leq \gamma^K \left\| v^{\pi'_0} - v^* \right\|_\infty \qquad \text{(From the claim in **(i)**)}$$

$$= \gamma^K \left\| \left(I - \gamma \mathbf{P}_{\pi'_0}\right)^{-1} g(\pi'_0, \pi^*) \right\|_\infty \qquad \text{(Value Difference Lemma)}$$

$$\leq \frac{\gamma^K}{1-\gamma} \|g(\pi'_0, \pi^*)\|_\infty \qquad \text{(Using the Neumann series)}$$

$$\implies \|g(\pi'_K, \pi^*)\|_\infty < \|g(\pi'_0, \pi^*)\|_\infty \qquad \left(K \geq K^* = \left\lceil \frac{\log(1/1-\gamma)}{\log(1/\gamma)} \right\rceil + 1\right)$$

Policy Iteration

Recall that $\|g(\pi'_K, \pi^*)\|_\infty < \|g(\pi'_0, \pi^*)\|_\infty$.
If $s' := \arg\max_s |g(\pi'_0, \pi^*)(s)| \implies \|g(\pi'_0, \pi^*)\|_\infty = -g(\pi'_0, \pi^*)(s')$, then,

$$\|g(\pi'_K, \pi^*)\|_\infty < -g(\pi'_0, \pi^*)(s') \implies \max_s |g(\pi'_K, \pi^*)| \leq -g(\pi'_0, \pi^*)(s')$$

$$\implies -g(\pi'_K, \pi^*)(s') < -g(\pi'_0, \pi^*)(s')$$

$$\implies v^*(s') - (\mathcal{T}_{\pi'_K} v^*)(s') < v^*(s') - (\mathcal{T}_{\pi'_0} v^*)(s') \quad \text{(Recall that } -g(\pi', \pi^*) = v^* - \mathcal{T}_{\pi'} v^*\text{)}$$

$$\implies \mathbf{r}_{\pi'_K}(s') + (\mathbf{P}_{\pi'_K} v^*)(s') > \mathbf{r}_{\pi'_0}(s') + (\mathbf{P}_{\pi'_0} v^*)(s') \qquad \text{(Recall that } \mathcal{T}_{\pi'} v^* = \mathbf{r}_{\pi'} + \mathbf{P}_{\pi'} v^*\text{)}$$

$$\implies \pi'_K(s') \neq \pi'_0(s') \quad \square \qquad \text{(Proof by contradiction)}$$

Linear Programming

Linear Programming and MDPs

Finding an optimal policy in an MDP is equivalent to solving a linear program.

Primal LP: For a starting state distribution $\rho \in \Delta_S$

$$v^* = \arg\min_{v \in \mathbb{R}^S} \langle \rho, v \rangle \quad \text{s.t. } \forall (s, a); \quad v(s) \geq r(s, a) + \gamma \sum_{s'} \mathcal{P}(s'|s, a)\, v(s')$$

- *Intuition*: In Lecture 4, while proving the Fundamental Theorem, we saw that if $v \geq \mathcal{T} v$, then $v \geq v^*$. The constraints in the primal LP correspond to $v \geq \mathcal{T} v$, and the objective is to find the smallest v that satisfies these constraints.
- The primal LP is over-determined and has S variables and $S \times A$ constraints.
- For each $s \in \mathcal{S}$, there exists an $a^*(s)$ such that $v^*(s) = r(s, a^*(s)) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}(s'|s, a^*(s)) v^*(s)$ i.e. the constraint is "tight".
- The stationary deterministic policy $\pi^*(s) = a^*(s)$ is an optimal policy and v^*, the solution to the primal LP is the optimal value function.
- For details and proofs, refer to Section 5.8.1 of [PC'23].

Linear Programming and MDPs

Dual LP: Define $r \in \mathbb{R}^{S \times A}$ to be the reward vector, $\mu \in \mathbb{R}^{S \times A}$ to be the *state-action occupancy measure* and $d^\pi \in \mathbb{R}^S$ to be the *state occupancy measure* such that,

$$\mu(s,a) := (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{t=0}^{\infty} \gamma^t \Pr[S_t = s, A_t = A | S_0 = s_0] \quad ; \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}$$

$$d(s) := (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{t=0}^{\infty} \gamma^t \Pr[S_t = s | S_0 = s_0] \quad \forall s \in \mathcal{S}$$

$$\mu^* = \operatorname*{arg\,max}_{\mu \in [0,\infty)^{S \times A}} \frac{\langle \mu, r \rangle}{1-\gamma} \quad \text{s.t.} \quad \forall s' \in \mathcal{S} \quad \gamma \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \mathcal{P}(s'|s,a)\, \mu(s,a) + (1-\gamma)\, \rho(s') = \sum_{a \in \mathcal{A}} \mu(s',a)$$

- *Intuition*: Maximizing the value function is equivalent to aligning μ to the reward vector r while ensuring that μ satisfies the "flow" constraints.
- The dual LP has SA variables and $SA + S$ constraints. μ^* consists of S non-zeros.
- There is a one-one mapping between μ and π, i.e. $\pi(a|s) = \mu(s,a) / \sum_{a'} \mu(s,a')$,
- Need to derive the dual LP from basics and implement it in Assignment 2!

Linear Programming and MDPs

- The primal and dual LPs satisfy *strong duality* i.e. $\langle \rho, v^* \rangle = \frac{\langle \mu^*, r \rangle}{1-\gamma}$.
- π^* is the greedy policy corresponding to v^* such that $\pi^*(s) = \arg\max_a \mu^*(s, a)$.
- The Simplex method for solving these LPs is equivalent to Policy Iteration.
- The resulting LP can be solved by other algorithms such as interior point methods, primal-dual methods and this connection has been recently exploited for proving sample-complexity results and designing algorithms with function approximation.

- We have studied algorithms that use knowledge of the transition probabilities $\mathcal{P}$ and rewards r to compute the optimal policy.
- These quantities are difficult to obtain in practical scenarios, and hence we need methods that can compute the optimal policy without explicitly relying on this information.
- In the next class, we will consider evaluating a fixed policy π without explicit knowledge of $\mathcal{P}$ and r.