B9120-001 Dynamic Programming Winter 2020

Prof. Daniel Russo

Homework Assignment 7: Due Monday April 6

Optimal Stopping. Consider fitted value iteration applied to optimal stopping. Consider an MDP with a finite state space $\mathcal{S} = \{1, \cdots n\} \cup \{\tau\}$ where τ is a costless absorbing state. In any state $s \neq \tau$, there are two actions, Stop and Continue. Stopping in any state $s \in \mathcal{S}$ transitions the state to τ but generates an immediate reward $r(s)$. A choice to continue accrues no immediate reward transitions the reward to a a new state drawn from the transition matrix P. Assume the Markov chain corresponding to P is irreducible and aperiodic and let $\pi = \pi P$ denote its stationary distribution.

Consider a linear approximation to the continuation value $J_\theta = \Phi\theta \in \mathbb{R}^n$, where $J_\theta(s) = \phi(s)^\top \theta$ $(\ldots \phi(i)^\top$ is the 'ith' row of J_θ). We study fitted value iteration

$$J_{\theta_{k+1}} = \Pi_{2,\pi} T J_{\theta_k}$$

where $\Pi_{2,\pi}$ is projection onto the span of the columns of Φ in the euclidean norm $\|\cdot\|_{2,\pi}$ and T is the operator

$$TJ(s) = \max\{r(s), \alpha \sum_{s' \in \mathcal{S}} P(s,s')J(s')\}.$$

(a) Show that T is a contraction with respect to $\|\cdot\|_{2,\pi}$.

(b) Use this to conclude that J_{θ_k} converges to the unique fixed point of $\Pi_{2,\pi}T$.

(c) State a bound on the distance between the solution to this projected fixed point equation and the optimal continuation value $J^* = TJ^*$. (You do not need to prove this bound)

Hint: In part (a), the following fact about real numbers may be helpful:

$$|\max\{x_1, y\} - \max\{x_2, y\}| \leq |x_1 - x_2|$$

Solution (a) Using the hint, we have

$$\begin{aligned}
\|FJ - F\bar{J}\|_{2,\pi}^2 &= \sum_{s \in \mathcal{S}} \pi(s) \left(\max\{r(s), \alpha \sum_{s' \in \mathcal{S}} P(s,s')J(s')\} - \max\{r(s), \alpha \sum_{s' \in \mathcal{S}} P(s,s')\bar{J}(s')\} \right)^2 \\
&\leq \sum_{s \in \mathcal{S}} \pi(s) \left(\alpha \sum_{s' \in \mathcal{S}} P(s,s')J(s')\} - \alpha \sum_{s' \in \mathcal{S}} P(s,s')\bar{J}(s') \right)^2 \\
&\leq \alpha^2 \sum_{s \in \mathcal{S}} \pi(s) \sum_{s' \in \mathcal{S}} P(s,s') \left(J(s') - \bar{J}(s') \right)^2 \\
&= \alpha^2 \sum_{s' \in \mathcal{S}} \pi(s') \left(J(s) - \bar{J}(s) \right)^2 \\
&= \alpha^2 \|J - \bar{J}\|_{2,\pi}^2.
\end{aligned}$$

(b) Since $\Pi_{2,\pi}$ is a non-expansion in $\|\cdot\|_{2,\pi}$ the composition $\Pi_{2,\pi}T$ is a contraction with modulus α.

(c) Let $\hat{J} = \Pi_{2,\pi}T\hat{J}$. We have

$$\|\hat{J} - J^*\|_{2,\pi} \leq \sqrt{\frac{1}{1-\alpha}}\|\Pi_{2,\pi}J^* - J^*\|_{2,\pi}$$

following the exactly same proof shown in class. See Class 8 slide 25.

Monotonicity of projections Many dynamic programming arguments critically use the monotonicity of the Bellman operator. But is the projected Bellman operator monotone? Construct an example showing that linear projection operators are not monotone in general. That is, For $J \preceq J'$ element-wise, we may not have $\Pi J \preceq \Pi J'$.

Solution Consider the projection operator Π from $\mathbb{R}^2$ to the linear subspace $V = \{(-1,1)\theta : \theta \in \mathbb{R}\}$. This projection is with respect to the (un-weighted) Euclidean norm $\|J\| = \sqrt{J(1)^2 + J(2)^2}$. Take $J = (1,0) \geq (0,0)$. Then $(\Pi J)(1) < 0$ whereas $(\Pi\vec{0})(1) = 0$.