CS294-40 Learning for Robotics and Control Lecture 5 - 9/11/2008

Function Approximation

Lecturer: Pieter Abbeel *Scribe: Nimbus Goehausen*

1 Lecture outline

- Review
- Function Approximation
- Alternative route to obtain converging fitted value iteration algorithms

2 Review

The Bellman backup operator T:

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

T is a γ-contraction with respect to the infinity norm, i.e., $\forall\ V, \overline{V}$:

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

This implies $\exists$ unique V^* such that $\forall V$:

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

When using linear function approximation, we represent V as follows:

$$V \approx \Phi\theta$$

One natural way of finding θ for a given V is by weighted linear regression:

$$\theta = \arg\min_\theta (V - \Phi\theta)^T W (V - \Phi\theta). \tag{1}$$

This gives us:

$$\theta = (\Phi^T W \Phi)^{-1} \Phi W V.$$

This particular function approximation representation of V can be represented using the projection operator $\Pi_W = \Phi(\Phi^T W \Phi)^{-1}\Phi^T W$. In particular, we have $\Pi_W V = \Phi\theta$ for θ the (weighted) least squares solution.

Fitted value iteration iterates applying the Bellman backup operator and a function approximation operator. For the weighted regression function approximation, we attain the following fitted value iteration algorithm:

$$V_{k+1} \leftarrow \Pi_W T V_k.$$

Or, written out more explicitly, fitted value iteration proceeds as follows:

$$\begin{aligned}\bar{V}_{k+1} &\leftarrow TV_k \\ \theta_{k+1} &\leftarrow (\Phi^T W \Phi)^{-1} \Phi^T W \bar{V}_{k+1} \\ V_{k+1} &\leftarrow \Phi \theta_{k+1}\end{aligned}$$

Weighted least least squares is a non-expansion with respect to the corresponding weighted norm, i.e., it has the following property:

$$\|\Pi_W V - \Pi_W \overline{V}\|_W \leq \|V - \overline{V}\|_W.$$

3 Function Approximation

A lofty goal:

$$\|TV - T\overline{V}\|_W \leq C\|V - \overline{V}\|_W \text{for } C \in (0,1).$$

This goal is a bit too ambitious so we go for this:

$$\|T_\mu V - T_\mu \overline{V}\|_W \leq C\|V - \overline{V}\|_W \text{for } C \in (0,1).$$

We start by explicitly writing out the left-hand side, where we have P be the transition distribution (matrix) when acting according to the fixed policy $\pi = (\mu, \mu, \ldots)$:

$$\begin{aligned}\|R + \gamma P \cdot V - (R + \gamma P \cdot \overline{V})\|_W^2 &= \gamma^2 \|P(V - \overline{V})\|_W^2 \\ &= \gamma^2 \sum_i W_i (P(V - \overline{V}))_i^2 \\ &= \gamma^2 \sum_i W_i (\sum_j P_{j|i} \cdot (V_j - \overline{V_j}))_i^2\end{aligned}$$

For a convex function f we have Jensen's inequality:

$$E[f(X)] \geq f(E[X]).$$

Using Jensen's inequality we can say that:

$$\gamma^2 \sum_i W_i (\sum_j P_{j|i} \cdot (V_j - \overline{V_j}))_i^2 \leq \gamma^2 \sum_i W_i \sum_j P_{j|i} \cdot (V_j - \overline{V_j})^2 \tag{2}$$

$$(\textit{Picking } W \textit{ such that } \sum_j P_{j|i} \cdot W_i = W_j) \tag{3}$$

$$= \gamma^2 \sum_j W_j (V_j \overline{V_j})^2 \tag{4}$$

$$= \gamma^2 \|V - \overline{V}\|_W^2 \tag{5}$$

Hence we have that $\Pi_W T_\mu$ is a γ-contraction with respect to the $\|\cdot\|_W$ norm when W is a diagonal matrix with entries corresponding to the stationary distribution when acting according to μ. As a consequence, we have that the ΠT_μ-operator has a unique fixed point $\hat{V}_\mu$, i.e.:

$$\lim_{H \to \infty} (\Pi T_\mu)^H V \to \hat{V}_\mu.$$

Now let's see whether this fixed point $\hat{V}_\mu$ is a decent approximation of V_μ.

$$
\begin{aligned}
\|\widehat{V}_\mu - V_\mu\|_W^2 &= \|\widehat{V}_\mu - \Pi_W V_\mu\|_W^2 + \|\Pi_W V_\mu - V_\mu\|_W^2 \quad \text{(Pythagoras)} \\
&\leq \|\Pi_W T_\mu \widehat{V}_\mu - \Pi_W V_\mu\|_W^2 + \|\Pi_W V_\mu - V_\mu\|_W^2 \quad (\hat{V}_\mu \text{ is fixed point of } \Pi_W T_\mu) \\
&\leq \|T_\mu \widehat{V}_\mu - V_\mu\|_W^2 + \|\Pi_W V_\mu - V_\mu\|_W^2 \quad (\Pi_W \text{ is a non-expansion w.r.t. } \|\cdot\|_W) \\
&\leq \|T_\mu \widehat{V}_\mu - T_\mu V_\mu\|_W^2 + \|\Pi_W V_\mu - V_\mu\|_W^2 \quad (V_\mu \text{ is the fixed point of } T_\mu) \\
&\leq \gamma^2 \|\widehat{V}_\mu - V_\mu\|_W^2 + \|\Pi_W V_\mu - V_\mu\|_W^2 \quad (T_\mu \text{ is a } \gamma\text{-contraction w.r.t. } \|\cdot\|_W)
\end{aligned}
$$

Hence, we have:

$$
\|\widehat{V}_\mu - V_\mu\|^2 \leq \frac{\|\Pi_W V_\mu - V_\mu\|_W^2}{1-\gamma^2}. \tag{6}
$$

This result shows us that, if we choose a good set of basis functions Φ, then the fixed point $\hat{V}_\mu$ of $\Pi_W T_\mu$ will be a good approximation of the fixed point V_μ of T_μ.

3.1 Example: How to run this through sampling to deal with a large state space

We fix policy $\pi = (\mu, \mu, \ldots)$ and we run the policy m times and for each run take one state sampled from the stationary distribution. This gives us $s^0, s^1, ..., s^m$).

We iterate through the following steps:

$$
\begin{aligned}
\forall i : V_{k+1}(s^i) &\leftarrow (T_\mu V_k)(s^i) \\
\text{Least squares: } (V_{k+1}(s^0) ... V_{k+1}(s^m))^T &\approx \Phi \cdot \theta
\end{aligned}
$$

Iteration continues until we have convergence and have found the fixed point of $\Pi_W T_\mu$.

4 An alternative approach to fitted value iteration

Recall the contraction property of T:

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

We will limit our choice of projection operations by picking Π such that:

$$
\|\Pi V - \Pi \overline{V}\|_\infty \leq \|V - \overline{V}\|_\infty. \tag{7}
$$

In combination with the contraction property, this directly implies that ΠT is a γ-contraction with respect to the infinity norm, i.e., we have

$$
\|\Pi T V - \Pi T \overline{V}\|_\infty \leq \gamma \|V - \overline{V}\|_\infty.
$$

For any contraction ΠT we have

$$
\exists \text{ unique } \ \widehat{V} : \lim_{H \to \infty} (\Pi T)^H V = \widehat{V}.
$$

Now let's find a choice of projection operator that satisfies the non-expansion property of Eqn. (7).

Let $\overline{S} = (s^0, s^1, ..., s^m)$ be the set of states for which we will do the Bellman backup operation. Let $g : S \to \overline{S}$ be a mapping that returns a reference state inside $\overline{S}$ for every state in S. Consider the following algorithm:

For $k = 0, 1, 2, \ldots$

$$
\forall s \in \overline{S} : \ V_{k+1}(s) = (T\Pi)(V_k)(s) = \max_{a \in A} R(s) + \gamma \sum_{s'} P(s'|s,a) V_k(g(s')) \tag{8}
$$

A very natural choice for $g(\cdot)$ would be to map states to their nearest neighbor in $\overline{S}$, but for the purpose of the current analysis, no such assumptions are needed about $g(\cdot)$.

For any choice of $\overline{S}$ and g, it's easily verified that the projection Π is a non-expansion w.r.t. the infinity norm, and hence ΠT is a γ-contraction w.r.t. the infinity norm.

Let $\widehat{V}^* = \lim_{H\to\infty}(\Pi T)^H V$, namely the unique fixed point of ΠT. Now let's investigate whether $\widehat{V}^*$ is a decent approximation of V^*:

$$\begin{aligned}
\|\widehat{V}^* - V^*\| &= \|\widehat{V}^* - \Pi V^* + \Pi V^* - V^*\|_\infty \\
&\leq \|\widehat{V}^* - \Pi V^*\|_\infty + \|\Pi V^* - V^*\|_\infty \\
&= \|\Pi T\widehat{V}^* - \Pi V^*\|_\infty + \|\Pi V^* - V^*\|_\infty \\
&\leq \|T\widehat{V}^* - V^*\|_\infty + \|\Pi V^* - V^*\|_\infty \\
&\leq \gamma\|\widehat{V}^* - V^*\|_\infty + \|\Pi V^* - V^*\|_\infty
\end{aligned}$$

Hence, we have

$$\|\widehat{V}^* - V^*\| \leq \frac{1}{1-\gamma}\|\Pi V^* - V^*\|_\infty.$$

While the analysis holds true for any choice of $g(\cdot)$, the choice of $g(\cdot)$ will greatly affect $\|\Pi V^* - V^*\|_\infty$ and thereby greatly affect the above performance guarantee.