Class Tutorial 11

1. Projection Operator

Consider the projection operator Π that projects a vector $V \in \mathbb{R}^n$ to a linear subspace S that is the span of k features $\phi_1(s),\ldots,\phi_k(s)$ w.r.t. the d-weighted Euclidean norm:

$$\|X\|_d^2 = \sum_{j=1}^{n} d(j)X(j)^2$$

a. Write down Π explicitly.

b. Show that a projection is **non-expansive**:

$$\|\Pi V - \Pi\tilde{V}\|_d \le \|V - \tilde{V}\|_d$$

Solution

a. Let $\Phi \in \mathbb{R}^{n\times k}$ denote a matrix with $\phi(s)$ in its rows. Note that every vector in S can be written as Φw for some w.

For $V \in \mathbb{R}^n$, by definition, we have

$$\Pi V = \Phi w^*$$
$$w^* = \arg\min_{w\in\mathbb{R}^k} \|\Phi w - V\|_d^2 = \arg\min_{w\in\mathbb{R}^k} (\Phi w - V)^\top D(\Phi w - V)$$

where $D = diag(d)$.

Taking a gradient and setting to zero gives:

$$\frac{\partial}{\partial w}(\Phi w - V)^\top D(\Phi w - V) = 2\Phi^\top D\Phi w - 2\Phi^\top DV$$
$$w^* = (\Phi^\top D\Phi)^{-1}\Phi^\top DV$$

and

$$\Pi = \Phi(\Phi^\top D\Phi)^{-1}\Phi^\top D.$$

b. The error $V - \Pi V$ is orthogonal to ΠV (property of a projection to linear subspaces, see Random Signals course), therefore we have the Pythagorean Theorem:

$$\|X\|_d^2 = \|\Pi X\|_d^2 + \|X - \Pi X\|_d^2$$

Now, using the linearity of Π :

$$\| \Pi V - \Pi \tilde{V} \|_d = \| \Pi(V - \tilde{V}) \|_d \leq \| \Pi(V - \tilde{V}) \|_d + \| (I - \Pi)(V - \tilde{V}) \|_d = \| V - \tilde{V} \|_d$$

2. Projected Bellman Operator

Consider an MDP and some policy π and let denote P the state transition probability matrix under π. Assume that P has a unique **stationary distribution** d, i.e., $\sum_{i=1}^{n} d_i P_{ij} = d_j$.

Also, let $T^{\pi}(v) \doteq r + \gamma P v$ denote the Bellman operator for policy π.

a. Show that $\| Pz \|_d^2 \leq \| z \|_d^2$.

b. Show that ΠT^{π} is a contraction in the $\| \cdot \|_d$ norm.

Solution

a.

$$\begin{aligned}
\| Pz \|_d^2 &= \sum_{i=1}^{n} d_i \left(\sum_{j=1}^{n} P_{ij} z_j \right)^2 \\
&\overset{Jensen}{\leq} \sum_{i=1}^{n} d_i \sum_{j=1}^{n} P_{ij} z_j^2 \\
&= \sum_{j=1}^{n} z_j^2 \sum_{i=1}^{n} d_i P_{ij} \\
&= \sum_{j=1}^{n} d_j z_j^2 \\
&= \| z \|_d^2
\end{aligned}$$

b. We have that T^{π} is a contraction in the $\| \cdot \|_d$ norm:

$$\| T^{\pi} V - T^{\pi} \tilde{V} \|_d = \| \gamma P (V - \tilde{V}) \|_d \leq \gamma \| V - \tilde{V} \|_d$$

Since Π is non-expansive in the $\| \cdot \|_d$ norm, ΠT^{π} is a contraction.

3. TD(0)

Consider the following problem: given a sequence of states $s_1, \ldots s_n$ and rewards $r_1, \ldots, r_n$ generated by an MDP with policy π, our goal is to estimate the value function $V^{\pi}(s)$. We approximate $V^{\pi}(s)$ using linear function approximation, i.e.,

$$\tilde{V}^{\pi}(s) = \phi(s)^{\top} w$$

where $\phi(s)$ denotes the features of state s and w are the approximation weights. Thus, our goal is to find w.

The **TD(0) algorithm** computes w iteratively as follows:

$$\begin{aligned} w_{t+1} &= w_t + \alpha_t \delta_t \phi(s_t) \\ \delta_t &\triangleq r(s_t) + \gamma \phi(s_{t+1})^\top w_t - \phi(s_t)^\top w_t \end{aligned} \qquad (1)$$

We will now prove that (a simplified version of) TD(0) converges. Our approach is to represent TD(0) as a stochastic approximation of the form

$$w_{t+1} = w_t + \alpha_t \left(h(w_t) + m_t \right)$$

where h is a deterministic function and m_t are independent with $\mathbb{E}\left[m_t \mid w_t \right] = 0$.

Assume that the MDP under policy π has a unique **stationary distribution** d.

a. Let $P_n(s) \triangleq \Pr(s_n = s)$ and $P_n(s,s') \triangleq \Pr(s_n = s, s_{n+1} = s')$. Calculate $P_\infty(s) \triangleq \lim_{n\to\infty} P_n(s)$ and $P_\infty(s,s') \triangleq \lim_{n\to\infty} P_n(s,s')$.

We now make the following simplifying assumption: in the TD(0) algorithm (1) each pair s_t, s_{t+1} is drawn i.i.d. from $P_\infty(s,s')$.

b. Calculate $\mathbb{E}\left[\delta_t \phi(s_t) \mid w_t \right]$, and show that it may be written as a linear function $Aw_t + b$.

c. Write the TD(0) algorithm as a stochastic approximation. What is the resulting ODE?

d. Show that the equilibrium point of the ODE corresponds to the solution of the projected fixed point equation $V = \Pi T^\pi V$. Conclude that the solution is unique.

e. Show that $z^\top \gamma D P z \le z^\top D z \qquad \forall z \ne 0$. Use the result to show that $D\left(\gamma P - I\right)$ is negative definite in the sense that $z^\top D\left(\gamma P - I\right) z < 0 \qquad \forall z \ne 0$. Show that if Φ is full rank then A is also negative definite.

It can be shown that the negative definiteness of A means that the eigenvalues of A have a negative real part.

f. Using the stochastic approximation theorem learned in class, state a convergence theorem for TD(0).

g. Why did we need to assume that each pair s_t, s_{t+1} is drawn i.i.d. from $P_\infty(s,s')$?

Solution

a. We have $\lim_{n\to\infty} P_n(s) = d(s)$ and $\lim_{n\to\infty} P_n(s,s') = d(s) P_{s,s'}$.

b. We have

$$\begin{aligned}\mathbb{E}\left[\delta_t\phi(s_t)\,|\,w_t\right] &= \mathbb{E}\left[\left(r(s_t)+\gamma\phi(s_{t+1})^\top w_t-\phi(s_t)^\top w_t\right)\phi(s_t)\,|\,w_t\right]\\ &= \mathbb{E}\left[r(s_t)\phi(s_t)\,|\,w_t\right]+\gamma\mathbb{E}\left[\phi(s_t)\phi(s_{t+1})^\top w_t\,|\,w_t\right]-\mathbb{E}\left[\phi(s_t)\phi(s_t)^\top w_t\,|\,w_t\right]\\ &= \mathbb{E}\left[r(s_t)\phi(s_t)\right]+\gamma w_t\mathbb{E}\left[\phi(s_t)\phi(s_{t+1})^\top\right]-w_t\mathbb{E}\left[\phi(s_t)\phi(s_t)^\top\right]\end{aligned}$$

Observe that

$$\mathbb{E}\left[r(s_t)\phi(s_t)\right]=\sum_s d(s)r(s)\phi(s)=\Phi^\top Dr$$

where Φ has $\phi(s)$ in its rows, and $D=diag(d)$. Similarly, we have

$$\mathbb{E}\left[\phi(s_t)\phi(s_t)^\top\right]=\sum_s d(s)\phi(s)\phi(s)^\top=\Phi^\top D\Phi$$

$$\mathbb{E}\left[\phi(s_t)\phi(s_{t+1})^\top\right]=\sum_{s,s'} d(s)P_{s,s'}\phi(s)\phi(s')^\top=\Phi^\top DP\Phi$$

Thus, we may write

$\mathbb{E}\left[\delta_t\phi(s_t)\,|\,w_t\right]=Aw_t+b$ where $A=\Phi^\top D\left(\gamma P-I\right)\Phi$ and $b=\Phi^\top Dr$.

c. We write

$$w_{t+1}=w_t+\alpha_t\left(\mathbb{E}\left[\delta_t\phi(s_t)\,|\,w_t\right]+\delta_t\phi(s_t)-\mathbb{E}\left[\delta_t\phi(s_t)\,|\,w_t\right]\right),$$

Thus

$$h\left(w_t\right)=Aw_t+b$$

And

$$m_t=\delta_t\phi(s_t)-\mathbb{E}\left[\delta_t\phi(s_t)\,|\,w_t\right].$$

The ODE is therefore

$$\dot{w}=Aw+b.$$

d. Let us write $V=\Pi T^\pi V$ explicitly. We first have $T^\pi V=r+\gamma PV$. We now write the equation explicitly as

$$\Phi w=\Phi\left(\Phi^\top D\Phi\right)^{-1}\Phi^\top D\left(r+\gamma P\Phi w\right) \qquad (2)$$

The equilibrium point of the ODE satisfies

$$\Phi^\top D\left(\gamma P - I\right)\Phi w - \Phi^\top Dr = 0 \qquad (3)$$

Multiplying Eq. (3) by $\Phi\left(\Phi^\top D\Phi\right)^{-1}$ gives Eq. (2). Since ΠT^π is a contraction the solution is unique.

e. Recall that $\|Pz\|_d \le \|z\|_d$. We have

$$\begin{aligned} z^\top \gamma DPz &= \gamma z^\top D^{1/2} D^{1/2} Pz \\ &\overset{c.s.}{\le} \gamma \left\|D^{1/2} z\right\| \left\|D^{1/2} Pz\right\| \\ &= \gamma \|z\|_d \|Pz\|_d \\ &\le \gamma \|z\|_d \|z\|_d = \gamma z^\top Dz \end{aligned}$$

Thus

$$\begin{aligned} z^\top D\left(\gamma P - I\right) z &\le \gamma z^\top Dz - z^\top Dz \\ &= \left(\gamma - 1\right) z^\top Dz < 0. \end{aligned} \qquad (4).$$

Finally, assume by negation that $z^\top \Phi^\top D\left(\gamma P - I\right)\Phi z \ge 0$ for some $z \ne 0$. If Φ is full-rank then $\Phi z \ne 0,$ which contradicts (4).

f. Let the step-sizes satisfy G1. It is easy to show that the noise satisfies N1. Also, clearly h is Lipschitz, and its unique equilibrium point is stable. The boundedness of w_t is not trivial, and may be shown using more advanced techniques (beyond the course material). Alternatively, one may use a projection approach to keep the iterates in some bounded set.

Assuming w_t is bounded w.p. 1, Theorem 1 of Lecture 10 guarantees that w_t converges to the unique fixed point of ΠT^π.

g. In the regular TD(0) algorithm, consecutive states are dependent. Actually, the stochastic approximation theorem may be extended to such cases as well, but it needs to be shown that the dependence is short-ranged, i.e. – the Markov chain has a fast mixing property. This is beyond the scope of this course.