MS&E338 Reinforcement Learning Lecture 7 - April 23, 2018

Approximate Value Iteration and Temporal Difference Learning

Lecturer: Ben Van Roy *Scribe: Shi Dong, Fei Xia*

1 Linear Parametrization and Approximate Value Iteration

Recall from last lecture that, when the state or action space is gigantic, efficiently storing the Q-values of all state-action pairs poses a serious problem. Besides, since the Q-values of different state-action pairs are often correlated, we may wish to infer about certain state-action pairs that have not been visited. Under such circumstances, value function approximation offers a handy solution. To fix ideas, we first reiterate several notations and concepts from last lecture that will be used throughout this lecture.

Let $\phi_1, \cdots, \phi_K : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ be K *feature extractors* of state-action pairs. The Q-function can be parametrized as

$$Q_\theta(s,a) = \sum_{k=1}^{K} \theta_k \phi_k(s,a) = (\Phi\theta)(s,a), \tag{1}$$

where $\theta = (\theta_1, \cdots, \theta_K)^\top \in \mathbb{R}^K$ is the vector of parameters, and

$$\Phi = \begin{bmatrix} | & | & \cdots & | \\ \phi_1 & \phi_2 & \cdots & \phi_K \\ | & | & \cdots & | \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\cdot|\mathcal{A}|\times K}.$$

Here we interpret each ϕ_k as an $|\mathcal{S}|\cdot|\mathcal{A}|$-dimensional vector. The *approximate value iteration* update can be written concisely by

$$Q \leftarrow \Pi F Q, \tag{2}$$

where Π is the projection operator with respect to the $\|\cdot\|_{2,\mu}$-norm, defined by

$$\|Q\|_{2,\mu} = \left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mu(s,a) Q^2(s,a)\right)^{\frac{1}{2}}, \quad \Pi Q = \operatorname*{argmin}_{Q_\theta} \|Q - Q_\theta\|_{2,\mu}, \tag{3}$$

and F is the Bellman operator used in the conventional value iteration algorithm.

To make things simple, we first consider the "autonomous" case, where there is only one action associated with each state. Since there is no other actions, we can omit the dependence of Q on a, and write

$$V_\theta(s) = Q_\theta(s,a) \quad \text{and} \quad \|V\|_{2,\mu} = \left(\sum_{s\in\mathcal{S}} \mu(s) V^2(s)\right)^{\frac{1}{2}}.$$

Notice that in such case the MDP degenerates into a Markov reward process, with transition probability matrix P. We make a simplifying assumption that

$$\sum_{s'\in\mathcal{S}} P_{s,s'} = \alpha, \quad \forall s \in \mathcal{S}.$$

The intuition is that, for each state in $\mathcal{S}$, the system terminates with probability $1-\alpha$. It can be shown that this scenario is equivalent to the "discounted" case in which future rewards are discounted by factor α. If we define

$$\overline{P} = \frac{P}{\alpha}, \tag{4}$$

then $\overline{P}$ is a stochastic matrix.

However, despite that F is a weighted-max norm contraction and Π is non-expansive, ΠF may not be a contraction, which could cause the approximate value iteration (2) to diverge, as is illustrated in Figure 5 of Lecture 6. Fortunately, the following proposition, which is proven in the last lecture, confirms the existence of a norm with respect to which ΠF is contractive.

Proposition 1. *Let $\rho \in \mathbb{R}^{|\mathcal{S}|}$ be such that $\rho^\top \overline{P} = \rho^\top$ and $\rho \succeq 0$, then for any $V, V' \in \mathbb{R}^{|\mathcal{S}|}$,*

$$\|\Pi F V - \Pi F V'\|_{2,\rho} \leq \alpha \|V - V'\|_{2,\rho}.$$

2 Error Bound of Approximate Value Iteration

Let $\mathcal{V} = \{V_\theta : \theta \in \mathbb{R}^K\}$ be the space of all parametrized value functions. Let V^* be the unique fixed point of F, i.e. $V^* = FV^*$. Suppose that $V^* \in \mathcal{V}$, then

$$V^* = \Pi V^* = \Pi F V^*,$$

which implies that V^* is also the fixed point of ΠF. From the contraction property of ΠF, we can guarantee that approximate value iteration eventually converges to V^*.

However, if $V^* \notin \mathcal{V}$, approximate value iteration can only converge to a value function "close" to V^*, as is shown in the following proposition.

Proposition 2. *Let $\tilde{V}$ be the unique fixed point of ΠF, i.e. $\tilde{V} = \Pi F \tilde{V}$. Then*

$$\|\tilde{V} - V^*\|_{2,\rho} \leq \frac{1}{\sqrt{1-\alpha^2}} \|\Pi V^* - V^*\|_{2,\rho},$$

where ρ is defined in Proposition 1.

Proof. Consider the inner product $\langle \cdot, \cdot \rangle_\rho$ on $\mathbb{R}^{|\mathcal{S}|}$ defined by

$$\langle V_1, V_2 \rangle_\rho = \sum_{s \in \mathcal{S}} \rho(s) V_1(s) V_2(s).$$

Apparently $\langle \cdot, \cdot \rangle_\rho$ induces the $\| \cdot \|_{2,\rho}$-norm. Since $\mathcal{V}$ is a linear subspace of $\mathbb{R}^{|\mathcal{S}|}$ and Π is the projection operator onto $\mathcal{V}$, we have that $V^* - \Pi V^* \in \mathcal{V}^\perp$, where $\mathcal{V}^\perp$ is the orthogonal complement of $\mathcal{V}$. Also notice that $\tilde{V} - \Pi V^* \in \mathcal{V}$, which gives

$$\left\langle \tilde{V} - \Pi V^*, V^* - \Pi V^* \right\rangle_\rho = 0.$$

Hence there is

$$\begin{aligned}
\|\tilde{V} - V^*\|_{2,\rho}^2 &= \|\tilde{V} - \Pi V^*\|_{2,\rho}^2 + \|\Pi V^* - V^*\|_{2,\rho}^2 \\
&\overset{(a)}{=} \|\Pi F \tilde{V} - \Pi F V^*\|_{2,\rho}^2 + \|\Pi V^* - V^*\|_{2,\rho}^2 \\
&\overset{(b)}{\leq} \alpha^2 \|\tilde{V} - V^*\|_{2,\rho}^2 + \|\Pi V^* - V^*\|_{2,\rho}^2,
\end{aligned}$$

where (a) follows from that $V^* = FV^*$ and (b) results from Proposition 1. Rearranging, we arrive at the desired result. □

3 Temporal-Difference Learning

3.1 Approximating the Bellman Operator

As we show in the previous sections, approximate value iteration converges to the proximity of V^*, with a provable error bound. However, recall that the Bellman operator is defined by

$$(FQ)(s,a) = \overline{R}(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s') \max_{a' \in \mathcal{A}} Q(s',a'), \tag{5}$$

where $\overline{R}(s,a)$ is the expected reward of state-action pair (s,a) and $P_{s,a}(s')$ is the probability of transitioning to s' under (s,a). From (5), we can see that a lookup table of Q values is still essential if we want to compute FQ accurately. Consequently, to implement (2) we need another approximation approach. To simplify the analysis, we still focus on the "autonomous" case, in which only one action is available at each state, and we use $V(s)$ in place of $Q(s,a)$.

First let us recall the Q-learning update in the tabular case, introduced in Lecture 4. Suppose we have the datapoint (s,r,s'), corresponding to current state, instantaneous reward and next state, respectively. The Q-learning update performs

$$V(s) \leftarrow V(s) + \gamma\big(r + V(s') - V(s)\big). \tag{6}$$

Under value function approximation, the update of parameter is given by

$$\begin{aligned} \theta &\leftarrow \theta + \gamma\big(\nabla_\theta V_\theta(s)\big)\big(r + V_\theta(s') - V_\theta(s)\big) \\ &\overset{(c)}{=} \theta + \gamma\phi(s)\big(r + (\Phi\theta)(s') - (\Phi\theta)(s)\big), \end{aligned} \tag{7}$$

where for each $s \in \mathcal{S}$, we define

$$\phi(s) = \begin{bmatrix} \phi_1(s) \\ \vdots \\ \phi_K(s) \end{bmatrix} \in \mathbb{R}^K,$$

and (c) follows from linear parametrization (1). To acquire intuition for the update (7), consider the tabular case in which $V = \theta$ and Φ is the identity matrix. Then (7) is identical to (6).

Let

$$\mathcal{O}\big(\theta; (s,r,s')\big) = \phi(s) \cdot \big(r + (\Phi\theta)(s') - (\Phi\theta)(s)\big).$$

Then update (7) can be written succinctly by

$$\theta \leftarrow \theta + \gamma\mathcal{O}\big(\theta; (s,r,s')\big),$$

which is the standard form of stochastic approximation, introduced in Lecture 4. Suppose that we have infinitely many datapoints (s,r,s'), operator $\mathcal{O}$ satisfies two technical conditions (cf. Proposition 2 of Lecture 4) and the learning rates $\{\gamma_k\}_{k=1}^\infty$ is chosen such that

$$\sum_{k=1}^\infty \gamma_k = \infty, \quad \sum_{k=1}^\infty \gamma_k^2 < \infty,$$

then, with probability 1, the stochastic approximation converges to the solution of following ODE:

$$\dot{\theta} = \mathbb{E}\big[\mathcal{O}\big(\theta; (s,r,s')\big)\big], \tag{8}$$

where the expectation is taken over datapoints (s,r,s'). Notice that

$$\begin{aligned} \mathbb{E}\big[\mathcal{O}\big(\theta; (s,r,s')\big)\big] &\overset{(d)}{=} \mathbb{E}\Big[\mathbb{E}\big[\phi(s)\big(r + (\Phi\theta)(s') - (\Phi\theta)(s)\big)\big|s\big]\Big] \\ &= \mathbb{E}\Big[\phi(s)\Big(\overline{R}(s) + \alpha(\overline{P}\Phi\theta)(s) - (\Phi\theta)(s)\Big)\Big], \end{aligned} \tag{9}$$

where (d) follows from the tower property of conditional expectation, $\overline{R}(s)$ is the expected reward of state s and $\overline{P}$ is defined as in (4). It is worth noting that the right-hand side of (9) only depends on how we sample the datapoints (s, r, s'). As long as we know the distribution of s, we can evaluate the right-hand side of (9) and solve the ODE (8) to obtain the limit of update (7).

3.2 TD(0) Algorithm

A natural choice for the distribution of s is ρ, the invariant distribution under normalized transition matrix $\overline{P}$. This gives rise to the following algorithm, which we call TD(0), where TD stands for *Temporal-Difference*.

Algorithm 1: Episodic TD(0)

Initialize with some θ
for $\ell = 0, 1, 2 \cdots$ **do**
  Observe $s_0 \sim \rho$
  for $t = 0, 1, \cdots \tau - 1$ **do**
    Observe r_{t+1}, s_{t+1}
    $\theta \leftarrow \theta + \gamma_\ell \cdot \phi(s_t)\big(r_{t+1} + V_\theta(s_{t+1}) - V_\theta(s_t)\big)$

Since we start from the invariant distribution for each episode, ODE (8) becomes

$$\begin{aligned} \dot{\theta} &= \sum_{s \in \mathcal{S}} \rho(s)\phi(s)\Big[\overline{R}(s) + \alpha(\overline{P}\Phi\theta)(s) - (\Phi\theta)(s)\Big] \\ &= \Phi^\top D\big(\overline{R} + \alpha\overline{P}\Phi\theta - \Phi\theta\big), \end{aligned} \tag{10}$$

where $\overline{R} \in \mathbb{R}^{|\mathcal{S}|}$ is the vector of expected reward of each state, and

$$D = \operatorname{diag}\big(\rho(1), \cdots, \rho(|\mathcal{S}|)\big) \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$$

is a diagonal matrix. Notice that, with this notation the inner product $\langle \cdot, \cdot \rangle_\rho$ is just

$$\langle V_1, V_2 \rangle_\rho = V_1^\top D V_2, \quad \forall V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}.$$

Let θ^* be such that $\Pi F V_{\theta^*} = V_{\theta^*}$, and let $d = \|\theta^* - \theta\|_2^2$. Then, we have

$$\begin{aligned} \dot{d} &= -2(\theta^* - \theta)^\top \dot{\theta} \\ &= -2(\theta^* - \theta)^\top \Phi^\top D\big(\overline{R} + \alpha\overline{P}\Phi\theta - \Phi\theta\big) \\ &\overset{(e)}{=} -2(\Phi\theta^* - \Phi\theta)^\top D\big(F\Phi\theta - \Phi\theta\big) \\ &= -2\Big[(\Phi\theta^* - \Phi\theta)^\top D(\Phi\theta^* - \Phi\theta) + (\Phi\theta^* - \Phi\theta)^\top D(F\Phi\theta - \Phi\theta^*)\Big] \\ &\overset{(f)}{=} -2\Big[\|\Phi\theta^* - \Phi\theta\|_{2,\rho}^2 + (\Phi\theta^* - \Phi\theta)^\top D(\Pi F\Phi\theta - \Phi\theta^*)\Big] \\ &\overset{(g)}{\leq} -2\Big[\|\Phi\theta^* - \Phi\theta\|_{2,\rho}^2 - \|\Phi\theta^* - \Phi\theta\|_{2,\rho}\|\Pi F\Phi\theta - \Phi\theta^*\|_{2,\rho}\Big] \\ &\overset{(h)}{\leq} -2\Big[\|\Phi\theta^* - \Phi\theta\|_{2,\rho}^2 - \alpha\|\Phi\theta^* - \Phi\theta\|_{2,\rho}^2\Big] \\ &= -2(1-\alpha)\|\Phi\theta^* - \Phi\theta\|_{2,\rho}^2, \end{aligned} \tag{11}$$

where (e) follows from the definition of Bellman operator F; (g) results from Cauchy-Schwarz inequality; (h) comes from the contraction property of ΠF (Proposition 1). To see (f), note that since Π is the projection

onto the linear subspace $\mathcal{V} = \{\Phi\theta : \theta \in \mathbb{R}^K\}$ and $\Phi\theta^* - \Phi\theta \in \mathcal{V}$, we have

$$\langle \Phi\theta^* - \Phi\theta, F\Phi\theta - \Phi\theta \rangle_\rho = \langle \Phi\theta^* - \Phi\theta, \Pi(F\Phi\theta - \Phi\theta) \rangle_\rho = \langle \Phi\theta^* - \Phi\theta, \Pi F\Phi\theta - \Phi\theta \rangle_\rho .$$

From (11), we can see that the distance between θ and θ^* is non-increasing. If we further assume that the columns of Φ are independent (which is a natural assumption since we can always eliminate redundant features), we can guarantee that $\|\theta^* - \theta\|_2$ strictly decreases until $\theta = \theta^*$. Therefore the updates of TD(0) converges to the optimal parameter θ^*.

3.3 TD(λ) Algorithm

In the update of TD(0), at each stage t we only exploit the approximate value function of next state s_{t+1}. In the following we introduce the TD(λ) algorithm, which incorporates the value estimates of the states that have been visited during the episode.

Algorithm 2: Episodic TD(λ)

Initialize with some $V \in \mathbb{R}^{|\mathcal{S}|}$
for $\ell = 0, 1, 2 \cdots$ **do**
  Observe $s_0 \sim \rho$
  Initialize $z_0 = \phi(s_0)$
  for $t = 0, 1, \cdots \tau - 1$ **do**
    Observe r_{t+1}, s_{t+1}
    $\theta \leftarrow \theta + \gamma_\ell \cdot z_t \cdot \big(r_{t+1} + V_\theta(s_{t+1}) - V_\theta(s_t)\big)$
    $z_{t+1} = \lambda z_t + \phi(s_{t+1})$

Apparently when $\lambda = 0$ the algorithm is identical to TD(0). It can be shown that the argument presented in Section 3.2 also applies to TD(λ) (left as homework).