Policy Evaluation

Chen-Yu Wei

Policy Evaluation

Given: a policy π

Evaluate $V^{\pi}(s)$ or $Q^{\pi}(s, a)$

$\pi_b(\cdot | s)$

On-policy policy evaluation: the learner can execute π to evaluate π

Off-policy/offline policy evaluation: the learner can only execute some $\pi_b \neq \pi$, or can only access some existing dataset to evaluate π

behavior policy

(s, a, r, s')

$(s_1, a_1, r_1, s_2, a_2, \ldots)$

Use cases:

- Approximate policy iteration: $\pi^{(k)}(s) = \underset{a}{\operatorname{argmax}}\, Q^{\pi^{(k-1)}}(s, a)$
- Estimate the value of a policy before deploying it in the real world, e.g., COVID-related border measures, economic recovery policies, or policy changes in recommendation systems.

Value Iteration for V^π / Q^π

$\boxed{Q^*(s,a)}$ $\boxed{V^*(s)}$

$\pi^*(s) = \arg\max_a Q^*(s,a)$

Input: π

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

$V^{(k)}(s) \xrightarrow{k \to \infty} V^\pi(s)$

$$\forall s, \qquad V^{(k)}(s) \leftarrow \sum_a \pi(a|s)\left(R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^{(k-1)}(s') \right)$$

Input: π

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

$Q^{(k)} \to Q^\pi$

$$\forall s, a, \qquad Q^{(k)}(s,a) \leftarrow R(s,a) + \gamma \sum_{s',a'} P(s'|s,a)\, \pi(a'|s') Q^{(k-1)}(s',a')$$

On-Policy Policy Evaluation

LSPE and TD

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

$a_i \sim \pi(\cdot|s_i)$, $\mathbb{E}[r_i] = R(s_i,a_i)$ $\quad s_i' \sim p(\cdot|s_i,a_i)$

Collecting samples $\{(s_i, r_i, s_i')\}_{i=1}^{n}$ using π

$$\theta_k \leftarrow \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(V_\theta(s_i) - r_i - \gamma V_{\theta_{k-1}}(s_i')\right)^2$$

Least-Square
Policy Evaluation (LSPE)

$$V_{\theta_k}(s) \approx \boxed{\sum_a \pi(a|s)\left(R(s,a) + \gamma \mathbb{E}_{s'\sim p(\cdot|s,a)}\left[V_{\theta_{k-1}}(s')\right]\right)}$$

For $i = 1,\ 2, \ldots$

Draw $a_i \sim \pi(\cdot\,|s_i)$

Observe reward r_i and next state s_{i+1}

$$\theta_i \leftarrow \theta_{i-1} - \alpha \nabla_\theta \left(V_\theta(s_i) - r_i - \gamma V_{\theta_{i-1}}(s_{i+1})\right)^2$$

Temporal difference learning

TD learning

TD(0) $\quad$ TD(λ)

LSPEQ and TDQ

TD

For $k = 1,\ 2, \dots$

Collecting samples $\{(s_i, a_i, r_i, s_i')\}_{i=1}^{n}$ using π

$$\theta_k \leftarrow \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \left(Q_{\theta}(s_i, a_i) - r_i - \gamma \sum_{a'} \pi(a' | s_i') \, Q_{\theta_{k-1}}(s_i', a') \right)^2$$

For $i = 1,\ 2, \dots$

Draw $a_i \sim \pi(\cdot\,|s_i)$, observe reward r_i and next state s_{i+1}

$$\theta_i \leftarrow \theta_{i-1} - \alpha \nabla_{\theta} \left(Q_{\theta}(s_i, a_i) - r_i - \gamma \sum_{a'} \pi(a' | s_i') \, Q_{\theta_{k-1}}(s_i', a') \right)^2$$

TD with Linear Function Approximation

$A \succcurlyeq 0$: A is psd

$A \succcurlyeq B \iff A - B$ is psd

BC : $R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} \tilde{Q}(s',a') = \phi(s,a)^\top \theta^\star \quad \forall \tilde{Q}$

Let μ be the stationary state distribution under policy π. Furthermore, assume

(1) $V^\pi(s) = \phi(s)^\top \theta^\star$ (realizability assumption)

imply (set $\tilde{Q} = Q^\star$)

(2) $\mathbb{E}_{s\sim\mu}[\phi(s)\phi(s)^\top] \succcurlyeq \rho I$ for some $\rho > 0$ (coverage assumption)

Realizability in $Q^\star$:

$Q^\star(s,a) = \phi(s,a)^\top \theta^\star$

Then the following TD update:

For $i = 1,2, \ldots$

In fact, even if the samples are generated as $a_i \sim \pi(\cdot|s_i)$, $r_i = R(s_i,a_i)$, $s_{i+1} \sim P(\cdot|s_i,a_i)$

Sample $s \sim \mu$, $a \sim \pi(\cdot|s)$, $r \sim R(s,a)$, $s' \sim P(\cdot|s,a)$

$$\theta_i \leftarrow \theta_{i-1} - \alpha_i \left(\phi(s)^\top \theta_{i-1} - r - \gamma \phi(s')^\top \theta_{i-1}\right)\phi(s)$$

converges to $\theta^\star$ with properly chosen α_i.

ϕ

$$\boxed{V^{\pi}(s) = \phi(s)^T\theta^*}$$

condition θ_i

$$\|\theta_{i+1}-\theta^*\|^2 = \|\theta_i - \alpha(\phi(s)^T\theta_i - r - \gamma\phi(s')^T\theta_i)\phi(s) - \theta^*\|^2$$

where $s\sim\mu$, $a\sim\pi(\cdot|s)$, $\mathbb{E}(r)=R(s,a)$, $s'\sim P(\cdot|s,a)$

$$= \|\theta_i-\theta^*\|^2 - 2\alpha(\theta_i-\theta^*)^T\underbrace{(\phi(s)^T\theta_i - r - \gamma\phi(s')^T\theta_i)\phi(s)}_{g} + \alpha^2\|g\|^2$$

$$\mathbb{E}[\|\theta_{i+1}-\theta^*\|^2] = \|\theta_i-\theta^*\|^2 - 2\alpha(\theta_i-\theta^*)^T\Big(\mathbb{E}[\phi(s)^T\theta_i - \cancel{r} - \gamma\phi(s')^T\theta_i] + \mathbb{E}[-\phi(s)^T\theta^* + \cancel{r} + \gamma\phi(s')^T\theta^*]\Big)\phi(s) + \alpha^2\|g\|^2$$

$$\text{blue} = \mathbb{E}[-V^{\pi}(s) + r + \gamma V^{\pi}(s')]$$

$$= \mathbb{E}\Big[-V^{\pi}(s) + \sum_a \pi(a|s)\big(R(s,a) + \gamma\,\mathbb{E}_{s'\sim P(\cdot|s,a)}(V^{\pi}(s'))\big)\Big]$$

$$= 0$$

$$= \|\theta_i-\theta^*\|^2 - 2\alpha(\theta_i-\theta^*)^T\,\mathbb{E}\big[\phi(s)(\phi(s)^T - \gamma\phi(s')^T)\big](\theta_i-\theta^*) + \alpha^2\|g\|^2$$

Comparison

Why does **Linear TDQ** converge (and converges to the correct solution) but **Linear Q-Learning** diverges?

Comparison

Under coverage assumption
(i.e., the data $\{(s_i, a_i, r_i, s_i')\}$ sufficiently cover every state-action pair / feature space)

<table>
<tr><th></th><th>LSVI</th><th>Watkins's Q-Learning</th><th>On-Policy LSPE(Q) / TD(Q)</th></tr>
<tr><td>Tabular</td><td>$Q^{(k)} \to Q^\star$</td><td>$Q^{(k)} \to Q^\star$</td><td rowspan="2">$V^{(k)} \to V^\pi$ / $Q^{(k)} \to Q^\pi$ under realizability</td></tr>
<tr><td>Linear Approx.</td><td>$Q^{(k)} \to Q^\star$ under Bellman completeness</td><td>Diverges even with Bellman completeness</td></tr>
</table>