Policy Evaluation

Chen-Yu Wei

Policy Evaluation

Given: a policy π

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

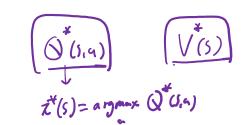**On-policy policy evaluation**: the learner can execute π to evaluate π $\pi_b(\cdot | s)$

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

(s, a, r, s')

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

behavior policy

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

$Q^*(s,a)$ $V^*(s)$

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

Input: π

For $k = 1, 2, \dots$

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

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

Input: π

For $k = 1, 2, \dots$

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

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

On-Policy Policy Evaluation

LSTD and TD

For $k = 1, 2, \ldots$

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

$a_i \sim \pi(\cdot | s_i)$, $\mathbb{E}(r_i) = R(s_i, a_i)$ $s_i' \sim P(\cdot | s_i, a_i)$

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

Least-Square Temporal Difference (LSTD)

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

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)   TD(λ)

LSTDQ and TDQ *TD*

For $k = 1, 2, \ldots$

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

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

For $i = 1, 2, \ldots$

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

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

TD with Linear Function Approximation

$A \succeq 0$: A is psd

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

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

imply $\left(\text{set } \widetilde{Q} = Q^\star \right)$

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

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

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

Then the following TD update:

Realizability in $Q^\star$:

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

For $i = 1, 2, ...$

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

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

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

$$\phi$$

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

$$\|\theta_{i+1} - \theta^*\|^2 = \| \underline{\theta_i} - \alpha \left(\phi(s)^T \theta_i - r - \gamma \phi(s')^T \theta_i \right) \phi(s) \underline{- \theta^*} \|^2 \quad \text{where} \quad \left(s \sim \mu, \; a \sim \pi(\cdot|s), \; \overset{E}{(r)} = R(s,a) \right.$$
$$\left. s' \sim P(\cdot|s,a) \right) \quad \text{condition } \theta_i$$

$$= \|\theta_i - \theta^*\|^2 - 2\alpha \, (\theta_i - \theta^*)^T \left(\phi(s)^T \theta_i - r - \gamma \phi(s')^T \theta_i \right) \phi(s) + \alpha^2 \|g\|^2$$

$$\mathbb{E}\left[\|\theta_{i+1} - \theta^*\|^2 \right] = \|\theta_i - \theta^*\|^2$$

$$- 2\alpha (\theta_i - \theta^*)^T \left(\underbrace{\mathbb{E}\left[\phi(s)^T \theta_i - r - \gamma \phi(s')^T \theta_i \right]}_{g} \phi(s) \right.$$
$$\left. + \mathbb{E}\left[-\phi(s)^T \theta^* + r + \gamma \phi(s')^T \theta^* \right] \right) + \alpha^2 \|g\|^2$$

$$\text{blue} = \mathbb{E}\left[-V^{\pi}(s) + r + \gamma \, \underline{V^{\pi}(s')} \right]$$
$$= \mathbb{E}\left[-V^{\pi}(s) + \sum_a \pi(a|s)\left(R(s,a) + \gamma \underset{s' \sim P(\cdot|s,a)}{\mathbb{E}} (V^{\pi}(s')) \right) \right]$$
$$= 0$$

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

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)

	LSVI	**Watkins's Q-Learning**	**On-Policy LSTD(Q) / TD(Q)**
Tabular	$Q^{(k)} \to Q^\star$	$Q^{(k)} \to Q^\star$	$V^{(k)} \to V^\pi$ / $Q^{(k)} \to Q^\pi$ under realizability
Linear Approx.	$Q^{(k)} \to Q^\star$ under Bellman completeness	Diverges even with Bellman completeness	