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

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

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

$\downarrow$

$\pi^*(s) = \underset{a}{\arg\max}\, Q^*(s,a)$

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

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

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

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

On-Policy Policy Evaluation

LSPE and TD

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

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

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

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

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

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

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

Least-Square
Policy Evaluation (LSPE)

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

Temporal difference learning

TD learning

TD(0) $TD(\lambda)$

LSPEQ and TDQ

TD

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

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

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

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

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

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

TD with Linear Function Approximation

$A \succeq 0$: A is psd

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

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

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 $\tilde{Q} = Q^*$)

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

Realizability in Q^*:

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

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

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

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

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

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

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

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

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

Comparison

Why does **Linear TD and 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>

Monte Carlo Estimation

Start from $s_1 = s^\star$
Execute policy π until the episode ends and obtain trajectory

$$s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_\tau, a_\tau, r_\tau$$

Let $G = \sum_{h=1}^{\tau} \gamma^{h-1} r_h$

G is an unbiased estimator for $V^\pi(s^\star)$

MC estimator: unbiased, higher variance

TD estimator: biased, lower variance

A Family of Estimators

Suppose we have a **value function estimation** $V_\theta(s) \approx V^\pi(s)$
Suppose we also have a **trajectory** $s_1, a_1, r_1, \dots, s_\tau, a_\tau, r_\tau$ generated by π

Then the following are all valid estimators for $V^\pi(s_1)$ besides $V_\theta(s_1)$:

$G_1 = r_1 + \gamma V_\theta(s_2)$
$G_2 = r_1 + \gamma r_2 + \gamma^2 V_\theta(s_3)$
...
$G_\tau = r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots + \gamma^{\tau-1} r_\tau$

Below, we will show

1. A way to combine these estimators
2. A more general policy evaluation method TD(λ) based on these estimators

Striking a Balance Between Bias and Variance

$$
\begin{aligned}
G_\theta(\lambda) &= (1-\lambda)\big(G_1 + \lambda G_2 + \lambda^2 G_3 + \cdots\big) \\
&= (1-\lambda)\big(r_1 + \gamma V_\theta(s_2)\big) + (1-\lambda)\lambda\big(r_1 + \gamma r_2 + \gamma^2 V_\theta(s_3)\big) + (1-\lambda)\lambda^2(\cdots) + \cdots
\end{aligned}
$$

TD(λ)

TD(0): $\theta_{k+1} \leftarrow \theta_k - \alpha \nabla_{\theta} \left(V_{\theta}(s_1) - r_1 - \gamma V_{\theta_k}(s_2) \right)^2$

TD(λ): $\theta_{k+1} \leftarrow \theta_k - \alpha \nabla_{\theta} \left(V_{\theta}(s_1) - G_{\theta_k}(\lambda) \right)^2$

Implementation details:

How to make update before reaching the end of the episode?

(Sutton and Barto Chapter 12)

TD(λ)

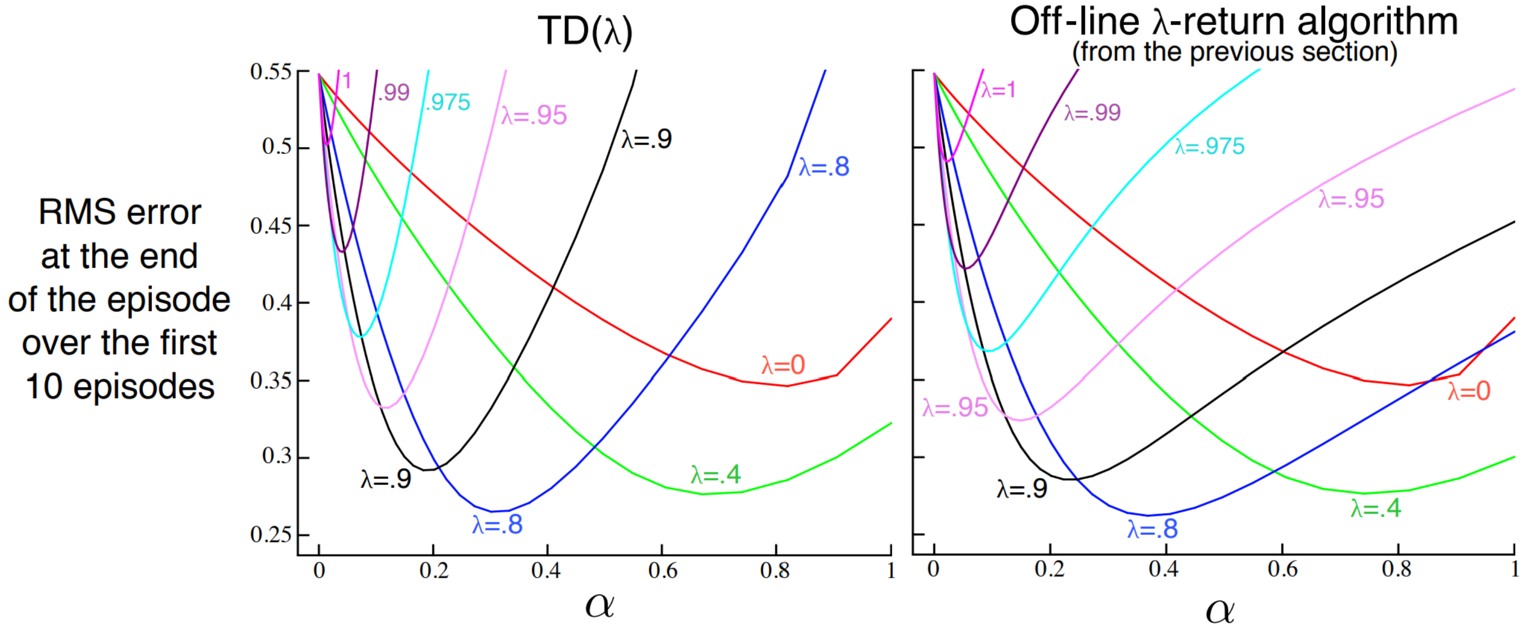
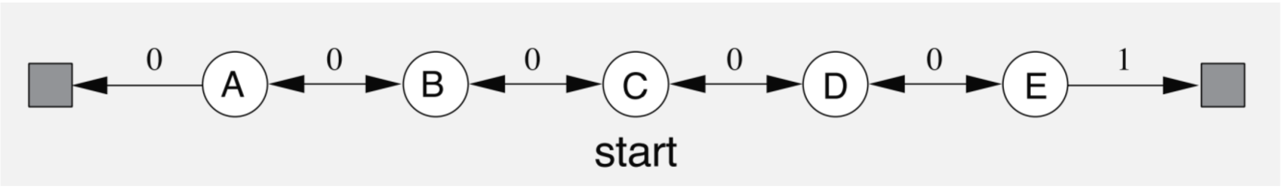


Figure 12.6: 19-state Random walk results (Example 7.1): Performance of TD(λ) alongside that of the off-line λ-return algorithm. The two algorithms performed virtually identically at low (less than optimal) α values, but TD(λ) was worse at high α values.

(Sutton and Barto Chapter 12)



Summary: On-Policy Policy Evaluation

- Double time-scale: **LSPE, LSPEQ**, Single time-scale: **TD, TDQ**
- TD (TD(0)) update:

$$(s, a, r, s') \sim \pi$$

$$\theta_{i+1} \leftarrow \theta_i - \alpha \nabla_\theta \left(V_\theta(s) - r - \gamma V_{\theta_i}(s') \right)^2 \Big|_{\theta=\theta_i}$$

- In the linear case, when realizability and coverage hold, we can show $\theta_i \to \theta^\star$
- Monte Carlo Estimator
- An estimator with parameter λ that balances variance and bias
- TD(λ)

Off-Policy Policy Evaluation

Off-Policy LSPEQ / TDQ

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

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

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

Bellman completeness + coverage will make it work

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

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

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

Like Q-learning, this is not stable

Off-Policy LSPE

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

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

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

Bellman completeness + coverage will make it work

(Sutton and Barto Chapter 11.7 and 11.8 have some better techniques to deal with the V_θ case with less assumptions)