

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')

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

↑
behavior policy

Use cases:

- Approximate policy iteration: $\pi^{(k)}(s) = \operatorname{argmax}_a Q^{\pi^{(k-1)}}(s, a)$
- 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

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

\downarrow
 $\pi^*(s) = \underset{a}{\operatorname{argmax}} Q^*(s,a)$

Input: π

For $k = 1, 2, \dots$

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

$V^{(k)}(s) \xrightarrow{k \rightarrow \infty} V^*(s)$

Input: π

For $k = 1, 2, \dots$

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

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

On-Policy Policy Evaluation

LSPE and TD

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

For $k = 1, 2, \dots$

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

Least-Square
Policy Evaluation (LSPE)

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

For $i = 1, 2, \dots$

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

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

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

$\theta = \theta_{i-1}$ | \therefore note: $V_{\theta}(s) = \phi(s)^T \theta \Rightarrow \nabla_{\theta} V_{\theta}(s) = \phi(s)$

$$2 \left(\underline{V_{\theta}(s_i)} - r_i - \gamma \underline{V_{\theta_{i-1}}(s_{i+1})} \right) \underline{\nabla_{\theta} V_{\theta}(s_i)} \Big|_{\theta = \theta_{i-1}}$$

Temporal difference learning

TD learning

TD(0) $\tau(0, \lambda)$

LSPEQ and TDQ τ

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

For $k = 1, 2, \dots$

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

For $i = 1, 2, \dots$

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

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

TD with Linear Function Approximation

$$A \succeq 0 : A \text{ is psd}$$

$$A \succeq B \Leftrightarrow A - B \text{ is psd}$$

$$\text{BC: } R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} \tilde{Q}(s',a') = \phi(s,a)^T \theta^* \quad \forall \tilde{Q}$$

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

(1) $V^\pi(s) = \phi(s)^T \theta^*$ (realizability assumption)

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

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

Then the following TD update:

Realizability in Q^* :

For $i = 1, 2, \dots$

$$Q^*_{(s,a)} = \phi(s,a)^T \theta^*$$

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

$$\theta_i \leftarrow \theta_{i-1} - \alpha_i (\phi(s)^T \theta_{i-1} - r - \gamma \phi(s')^T \theta_{i-1}) \phi(s)$$

converges to θ^* with properly chosen α_i .

✱

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

$$\|\theta_i - \alpha g - \theta^*\|^2 = \|\theta_i - \theta^*\|^2 - 2\alpha (\theta_i - \theta^*)^T g + \alpha^2 \|g\|^2$$

condition θ_i

$$\|\theta_{i+1} - \theta^*\|^2 = \|\theta_i - \alpha (\phi(s)^T \theta_i - r - \gamma \phi(s')^T \theta_i) \phi(s) - \theta^*\|^2 \text{ where } \begin{cases} s \sim \mu, a \sim \pi(\cdot|s), \mathbb{E}(r) = R(s,a) \\ s' \sim p(\cdot|s,a) \end{cases}$$

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

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

$$\begin{aligned} \text{blue} &= \mathbb{E}[-V^\pi(s) + r + \gamma V^\pi(s')] \\ &= \mathbb{E}[-V^\pi(s) + \sum_a \pi(a|s) (R(s,a) + \gamma \mathbb{E}_{s' \sim p(\cdot|s,a)} (V^\pi(s')))] \\ &= 0 \end{aligned}$$

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

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

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

$$\begin{aligned} & \mathbb{E} \mathbb{E} [\underbrace{(V^\pi(s) - r - \gamma V^\pi(s'))}_{\substack{a \sim \pi(\cdot|s) \\ s' \sim p(\cdot|s,a)}} \phi(s)] \\ & \mathbb{E} \mathbb{E} [V^\pi(s')] \end{aligned}$$

Comparison

$$s \sim \mu, a \sim \pi(\cdot|s) \quad \underline{s' \sim p(\cdot|s,a)}$$

$$\mathbb{E}(a) \mathbb{E}(a) \leq \mathbb{E}(a^2)$$

On-Policy

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

Off-Policy

$$\begin{aligned} & \mathbb{E} \left[\phi(s) \phi(s)^T - \gamma \phi(s) \phi(s')^T \right] \\ &= \underbrace{\mathbb{E}_{s \sim \mu} [\phi(s) \phi(s)^T]} - \gamma \underbrace{\mathbb{E}_{s \sim \mu} [\phi(s)]}_{\mathbb{E}_{s \sim \mu} [\phi(s)]}^T \underbrace{\mathbb{E}_{s \sim \mu} [\phi(s)^T]}_{\mathbb{E}_{s \sim \mu} [\phi(s)^T]} \geq (1-\gamma) \underbrace{\mathbb{E}_{s \sim \mu} [\phi(s) \phi(s)^T]}_{\geq (1-\gamma) \rho I} \end{aligned}$$

$$\begin{aligned} \mathbb{E} \left(\|\theta_{i+1} - \theta^*\|^2 \right) &\leq \|\theta_i - \theta^*\|^2 - 2\alpha (\theta_i - \theta^*)^T \underbrace{\mathbb{E} \left[\phi(s) (\phi(s) - \gamma \phi(s'))^T \right]}_{\geq cI} (\theta_i - \theta^*) + \alpha^2 \|\gamma\|^2 \\ &\leq \|\theta_i - \theta^*\|^2 - 2\alpha c \|\theta_i - \theta^*\|^2 + \alpha^2 \|\gamma\|^2 \\ &\leq (1-2\alpha c) \|\theta_i - \theta^*\|^2 + \alpha^2 \|\gamma\|^2 \end{aligned}$$

(s,a,r,s') from arbitrary policy

Q-learning

$$Q^k(s,a) \leftarrow (1-\alpha) Q^{k-1}(s,a) + \alpha \left(r + \gamma \max_{a'} Q^{k-1}(s',a') \right)$$

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

$$Q^k(s,a) \leftarrow (1-\alpha) Q^{k-1}(s,a) + \alpha \left(r + \gamma \sum_{a'} \pi(a'|s) Q^{k-1}(s',a') \right)$$

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 LSPE(Q) / TD(Q)
Tabular	$Q^{(k)} \rightarrow Q^*$	$Q^{(k)} \rightarrow Q^*$	$V^{(k)} \rightarrow V^\pi / Q^{(k)} \rightarrow Q^\pi$ under realizability
Linear Approx.	$Q^{(k)} \rightarrow Q^*$ under Bellman completeness	Diverges even with Bellman completeness	

Monte Carlo Estimation

Start from $s_1 = s^*$

Execute policy π until the episode ends and obtain trajectory

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

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

$\pi(a)$

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

MC estimator: unbiased, higher variance

TD estimator: biased, lower variance

a sample of arm a 's reward

$\hat{R}(a)$

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 $V_\theta(s_1)$:

$$G_1 = r_1 + \gamma V_\theta(s_2)$$

$$G_2 = r_1 + \gamma r_2 + \gamma^2 V_\theta(s_3)$$

...

$$G_\tau = r_1 + \gamma r_2 + \gamma^2 r_3 + \dots + \gamma^{\tau-1} r_\tau + \square$$

$G_{\tau+1}$

$G_{\tau+2}$

↑ higher bias

↓ higher variance

$$\frac{w_1 G_1 + w_2 G_2 + \dots + w_z G_z}{\sim}$$

$$w_1 + w_2 + \dots + w_z = 1$$

Below, we will show

1. A way to combine these estimators

2. A more general policy evaluation method $TD(\lambda)$ based on these estimators

$TD(0) = TD \text{ learning}$

$TD(1) = MC \text{ estimation}$

$$\lambda \in (0, 1]$$

Striking a Balance Between Bias and Variance

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

TD(λ)

(s_1, a_1, r_1) $s_2, a_2, r_2, \dots, s_L, a_L, r_L$

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

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

$s_1, a_1, r_1, s_2, a_2, r_2, \dots, s_L, a_L, r_L$

Implementation details:

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

([Sutton and Barto](#) Chapter 12)

TD(λ)

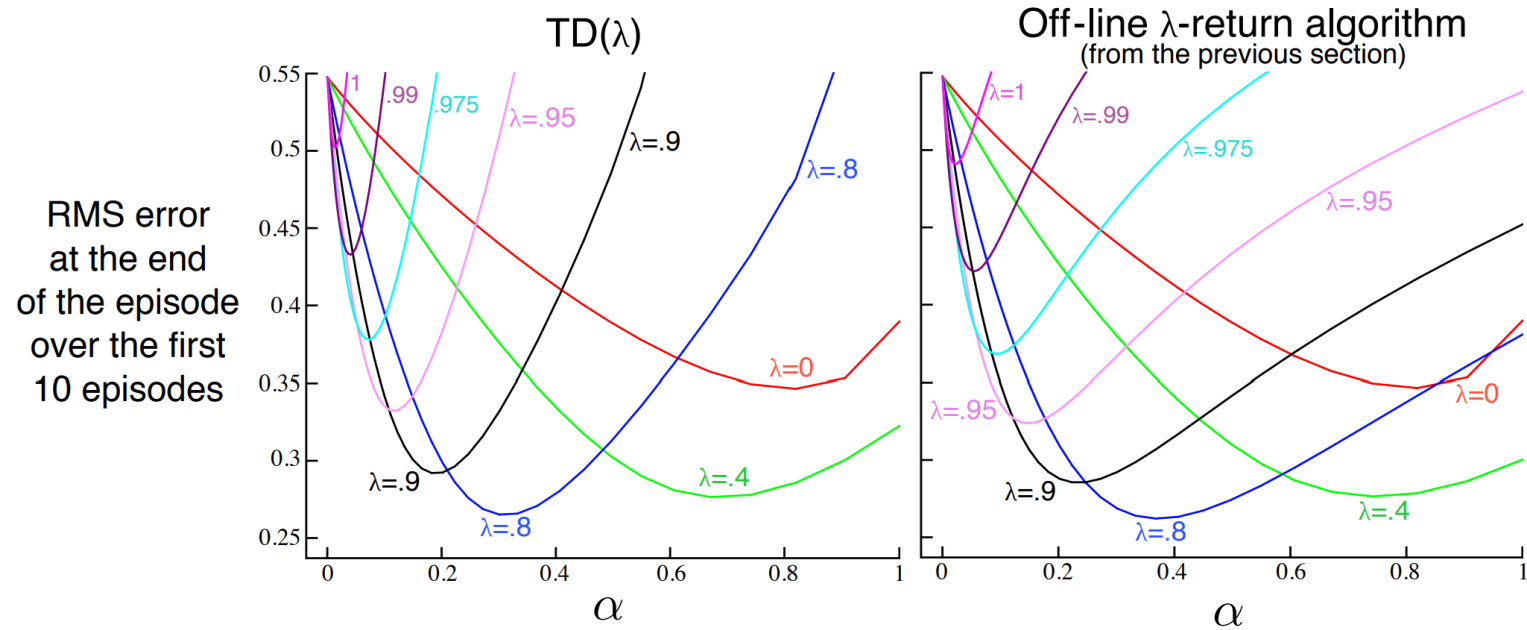
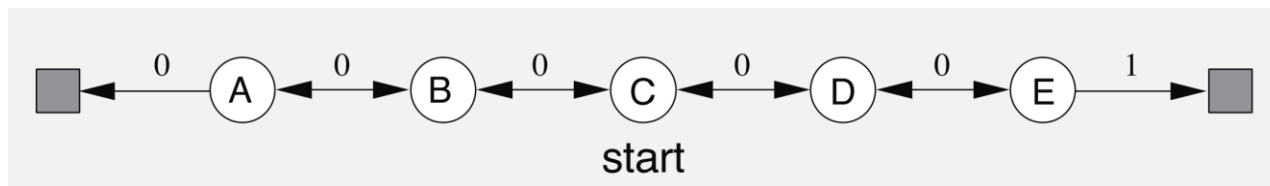


Figure 12.6: 19-state Random walk results (Example 7.1): Performance of TD(λ) alongside that of the off-line λ -return algorithm. The two algorithms performed virtually identically at low (less than optimal) α values, but TD(λ) was worse at high α values.

([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 $\theta_i \rightarrow \theta^*$
- 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 π_b

For $k = 1, 2, \dots$

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

Bellman completeness + coverage will make it work

For $i = 1, 2, \dots$

Draw $a_i \sim \pi_b(\cdot | s_i)$, 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 π_b

For $k = 1, 2, \dots$

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

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)