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}

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

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

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

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

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

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

Least-Square
Policy Evaluation (LSPE)

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

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

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

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

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

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

Temporal difference learning

TD learning

TD(0) $TD(\lambda)$

linear: $V_{\theta}(s) = \phi(s)^T \theta \Rightarrow \nabla_{\theta} V_{\theta}(s) = \phi(s)$

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

$A \succcurlyeq 0$: A is psd

$A \succcurlyeq B \iff A - B$ 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)

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

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

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

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

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

ϕ

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

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

condition θ_i

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

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

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

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

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

$$= 0$$

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

$$\mathbb{E}_{a \sim \pi(\cdot|s)}\ \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]$$

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

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

Comparison

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

$$\boxed{\mathbb{E}(a)\mathbb{E}(a) \le \mathbb{E}(a^2)}$$

On-Policy

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

Off-Policy

$$\mathbb{E}\left[\phi(s)\phi(s)^T - \gamma\phi(s)\phi(s')^T\right]$$

$$= \mathbb{E}_{s\sim\mu}\left[\phi(s)\phi(s)^T\right] - \gamma\,\mathbb{E}_{s\sim\mu}\left[\phi(s)\right]\mathbb{E}_{s\sim\mu}\left(\phi(s)\right)^T \gtrsim (1-\gamma)\,\mathbb{E}_{s\sim\mu}\left[\phi(s)\phi(s)^T\right]$$

$$\ge (1-\gamma)\rho I$$

$$\mathbb{E}\left[\|\theta_{i+1}-\theta^*\|^2\right] \le \|\theta_i-\theta^*\|^2 - 2\alpha(\theta_i-\theta^*)^T\underbrace{\mathbb{E}\left[\phi(s)\left(\phi(s)-\gamma\phi(s')\right)^T\right]}_{\succeq cI}(\theta_i-\theta^*) + \alpha^2\|g\|^2$$

$$\le \|\theta_i-\theta^*\|^2 - 2\alpha c\|\theta_i-\theta^*\|^2 + \alpha^2\|g\|^2$$

$$\le (1-2\alpha c)\|\theta_i-\theta^*\|^2 + \alpha^2\|g\|^2$$

(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') ~~Sadle~~, $a\sim\pi(\cdot|s)$

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

Comparison

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

<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

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

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

$\pi(a)$

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

MC estimator: unbiased, higher variance

TD estimator: biased, lower variance

a sample of arm a's reward

$\boxed{\hat{R}(a)}$

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 $V_\theta(s_1)$:

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

higher bias

higher variance

$w_1 G_1 + w_2 G_2 + \dots + w_\tau G_\tau$

$w_1 + w_2 + \dots w_\tau = 1$

$G_{\tau+1}$ $+ \boxed{0}$

$G_{\tau+2}$

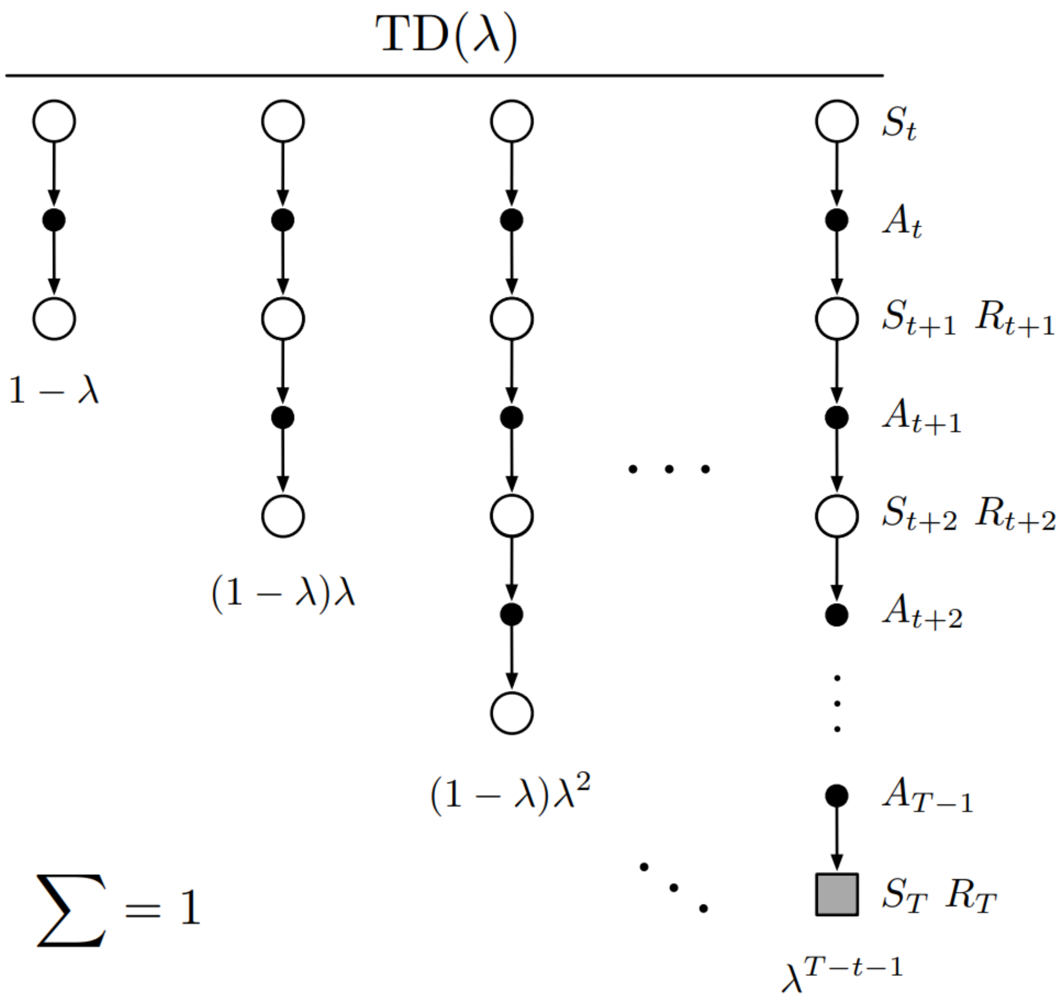
Below, we will show

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

TD(0) = TD learning

TD(1) = MC estimation

$\lambda \in [0,1]$


TD(λ)
S_t
A_t
S_{t+1} R_{t+1}
A_{t+1}
S_{t+2} R_{t+2}
A_{t+2}
A_{T-1}
S_T R_T
$1-\lambda$
$(1-\lambda)\lambda$
$(1-\lambda)\lambda^2$
λ^{T-t-1}
$\sum = 1$

Striking a Balance Between Bias and Variance

$(1-\lambda) + (1-\lambda)\lambda + (1-\lambda)\lambda^2 + \cdots$

$$G_\theta(\lambda) = (1-\lambda)\big(G_1 + \lambda G_2 + \lambda^2 G_3 + \cdots\big)$$

$$= (1-\lambda)\big(r_1 + \gamma V_\theta(s_2)\big) + (1-\lambda)\lambda\big(r_1 + \gamma r_2 + \gamma^2 V_\theta(s_3)\big) + (1-\lambda)\lambda^2(\cdots) + \cdots$$

TD(λ)

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

$G_{\theta_k}(0)$

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

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

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

Implementation details:

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

(Sutton and Barto Chapter 12)

TD(λ)

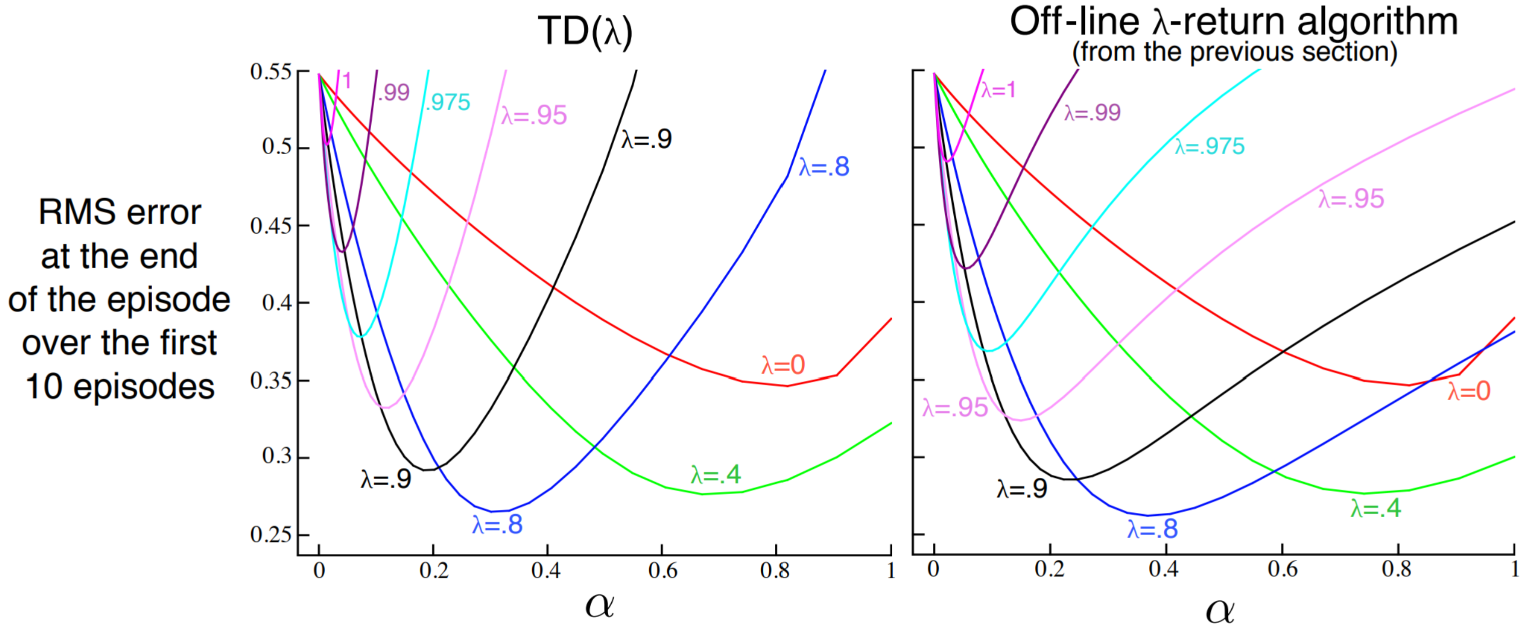
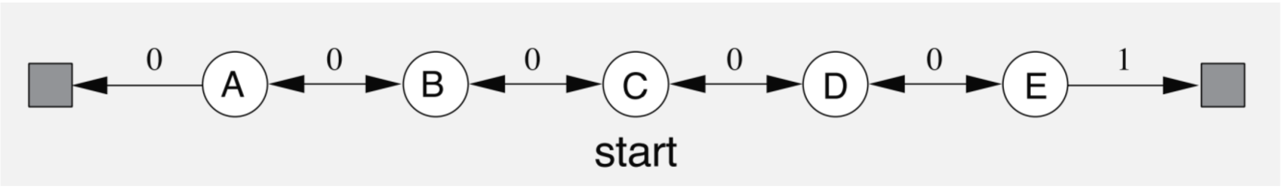


Figure 12.6: 19-state Random walk results (Example 7.1): Performance of TD(λ) alongside that of the off-line λ-return algorithm. The two algorithms performed virtually identically at low (less than optimal) α values, but TD(λ) was worse at high α values.

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

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

Bellman completeness + coverage will make it work

(Sutton and Barto Chapter 11.7 and 11.8 have more techniques to deal with the V_θ case)