Retrace(λ)

Temporal Credit Assignment in Off-Policy Reinforcement Learning

Matteo Papini

28th November 2017

Before we start...

Syllabus

Prediction and Control

- **Prediction**: measure the performance of a given policy π
- **Control**: find the optimal policy π^*

Prediction and Control

- **Prediction**: measure the performance of a given policy π
- **Control**: find the optimal policy π^*

Value Function

Bellman expectation equation:

$$Q^\pi(s,a) = R(s,a) + \gamma \underset{\substack{s' \sim \mathcal{P} \\ a' \sim \pi}}{E} \left[Q^\pi(s',a') \right]$$

Bellman expectation **operator**:

$$\mathcal{T}^\pi Q = R(s,a) + \gamma \underset{\substack{s' \sim \mathcal{P} \\ a' \sim \pi}}{E} \left[Q(s',a') \right]$$

- Q^π is the unique fixed point of $\mathcal{T}^\pi$
- Contraction property: $\|\mathcal{T}^\pi Q - Q^\pi\|_\infty \leq \gamma \|Q - Q^\pi\|_\infty$

Value Function

Bellman expectation equation:

$$Q^\pi(s,a) = R(s,a) + \gamma \underset{\substack{s' \sim \mathcal{P} \\ a' \sim \pi}}{E} \left[Q^\pi(s',a') \right]$$

Bellman expectation **operator**:

$$\mathcal{T}^\pi Q = R(s,a) + \gamma \underset{\substack{s' \sim \mathcal{P} \\ a' \sim \pi}}{E} \left[Q(s',a') \right]$$

- Q^π is the unique fixed point of $\mathcal{T}^\pi$
- Contraction property: $\|\mathcal{T}^\pi Q - Q^\pi\|_\infty \leq \gamma \|Q - Q^\pi\|_\infty$

Bellman expectation equation:

$$Q^\pi(s,a) = R(s,a) + \gamma \underset{\substack{s' \sim \mathcal{P} \\ a' \sim \pi}}{E} \left[Q^\pi(s',a') \right]$$

Bellman expectation **operator**:

$$\mathcal{T}^\pi Q = R(s,a) + \gamma \underset{\substack{s' \sim \mathcal{P} \\ a' \sim \pi}}{E} \left[Q(s',a') \right]$$

- Q^π is the unique fixed point of $\mathcal{T}^\pi$
- Contraction property: $\|\mathcal{T}^\pi Q - Q^\pi\|_\infty \leq \gamma \|Q - Q^\pi\|_\infty$

Value Function

Bellman expectation equation:

$$Q^{\pi}(s,a) = R(s,a) + \gamma \underset{\substack{s' \sim \mathcal{P} \\ a' \sim \pi}}{E} \left[Q^{\pi}(s',a') \right]$$

Bellman expectation **operator**:

$$\mathcal{T}^{\pi} Q = R(s,a) + \gamma \underset{\substack{s' \sim \mathcal{P} \\ a' \sim \pi}}{E} \left[Q(s',a') \right]$$

- Q^{π} is the unique fixed point of $\mathcal{T}^{\pi}$
- Contraction property: $\|\mathcal{T}^{\pi} Q - Q^{\pi}\|_{\infty} \leq \gamma \|Q - Q^{\pi}\|_{\infty}$

Optimal Value Function

Bellman optimality equation:

$$Q^*(s,a) = R(s,a) + \gamma \underset{s' \sim \mathcal{P}}{E} \left[\max_{a' \in \mathcal{A}} Q^*(s',a') \right]$$

Bellman optimality **operator**:

$$\mathcal{T}^* Q = R(s,a) + \gamma \underset{s' \sim \mathcal{P}}{E} \left[\max_{a' \in \mathcal{A}} Q^*(s',a') \right]$$

- Q^* is the unique fixed point of $\mathcal{T}^*$
- Contraction property: $\|\mathcal{T}^* Q - Q^*\|_\infty \leq \gamma \|Q - Q^*\|_\infty$

- Optimal **deterministic** policy:

$$\pi^*(s) = \pi_{\text{greedy}(Q^*)}(s) \doteq \arg\max_{a' \in \mathcal{A}} Q^*(s, a')$$

- Maximization as a special case of expectation:

$$\max_{a' \in \mathcal{A}} Q(s, a') = \underset{a' \sim}{E} \left[Q(s, a') \right]$$

- Optimal **deterministic** policy:

$$\pi^*(s) = \pi_{\mathsf{greedy}(Q^*)}(s) \doteq \arg\max_{a' \in \mathcal{A}} Q^*(s, a')$$

- Maximization as a special case of expectation:

$$\max_{a' \in \mathcal{A}} Q^*(s, a') = \underset{a' \sim \pi^*}{E} \left[Q^*(s, a') \right]$$

- Optimal **deterministic** policy:

$$\pi^*(s) = \pi_{\text{greedy}(Q^*)}(s) \doteq \arg\max_{a' \in \mathcal{A}} Q^*(s, a')$$

- Maximization as a special case of expectation:

$$\max_{a' \in \mathcal{A}} Q(s, a') = \underset{a' \sim \pi_{\text{greedy}(Q)}}{E} \left[Q(s, a') \right]$$

Prediction and Control

- **Prediction**: given policy π, compute Q^π
- **Control**: find π^* such that $Q^{\pi^*} = Q^*$

Temporal Difference Learning

- Idea: repeatedly apply $\mathcal{T}$ to Q
- Approximation: update Q towards a target
- Prediction: fixed π, keep updating Q
- Control: alternate value updates and policy updates

$$\pi_{k+1} = \pi_{\text{greedy}(Q_k)}$$

Temporal Difference Learning

- Idea: repeatedly apply $\mathcal{T}$ to Q
- Approximation: update Q towards a target

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[G_t - Q(s_t, a_t) \right]$$

- Prediction: fixed π, keep updating Q
- Control: alternate value updates and policy updates

$$\pi_{k+1} = \pi_{\text{greedy}(Q_k)}$$

Temporal Difference Learning

- Idea: repeatedly apply $\mathcal{T}$ to Q
- Approximation: update Q towards a target

$$\Delta Q(s_t, a_t) = \alpha \left[G_t - Q(s_t, a_t) \right]$$

- Prediction: fixed π, keep updating Q
- Control: alternate value updates and policy updates

$$\pi_{k+1} = \pi_{\text{greedy}(Q_k)}$$

Temporal Difference Learning

- Idea: repeatedly apply $\mathcal{T}$ to Q
- Approximation: update Q towards a target

$$\Delta Q(s_t, a_t) = \alpha \left[G_t - Q(s_t, a_t) \right]$$

- Prediction: fixed π, keep updating Q
- Control: alternate value updates and policy updates

$$\pi_{k+1} = \pi_{\text{greedy}(Q_k)}$$

- Idea: repeatedly apply $\mathcal{T}$ to Q
- Approximation: update Q towards a target

$$\Delta Q(s_t, a_t) = \alpha\left[G_t - Q(s_t, a_t)\right]$$

- Prediction: fixed π, keep updating Q
- Control: alternate value updates and policy updates

$$\pi_{k+1} = \pi_{\mathsf{greedy}(Q_k)}$$

- Idea: repeatedly apply $\mathcal{T}$ to Q
- Approximation: update Q towards a target

$$\Delta Q(s_t, a_t) = \alpha \left[G_t - Q(s_t, a_t) \right]$$

- Prediction: fixed π, keep updating Q
- Control: alternate value updates and policy updates

$$\pi_{k+1} = (1 - \epsilon)\pi_{\text{greedy}(Q_k)} + \epsilon \pi_{\text{any}}$$

Increasingly greedy policy: $\epsilon \to 0$ as $t \to \infty$

Expected Sarsa

- **Forward view**: look one step forward to compute the target

$$\Delta Q(s_t, a_t) = \alpha \left[r_{t+1} + \gamma \underset{a \sim \pi}{E} [Q(s_{t+1}, a)] - Q(s_t, a_t) \right]$$

- **Backward view**: wait one step to update $Q(s_t, a_t)$

$$\Delta Q(s_{t-1}, a_{t-1}) = \alpha \left[r_t + \gamma \underset{a \sim \pi}{E} [Q(s_t, a)] - Q(s_{t-1}, a_{t-1}) \right]$$

Expected Sarsa

- **Forward view**: look one step forward to compute the target

$$\Delta Q(s_t, a_t) = \alpha \left[r_{t+1} + \gamma \underset{a \sim \pi}{E} \left[Q(s_{t+1}, a) \right] - Q(s_t, a_t) \right]$$

- **Backward view**: wait one step to update $Q(s_t, a_t)$

$$\Delta Q(s_{t-1}, a_{t-1}) = \alpha \left[r_t + \gamma \underset{a \sim \pi}{E} \left[Q(s_t, a) \right] - Q(s_{t-1}, a_{t-1}) \right]$$

Why ”expected” sarsa?

- Short answer: useful later
- Longer answer: Van Seijen et al. [2]

Temporal Credit Assignment Problem

Look far (n steps) in the future:

$$G_t^{(n)} \doteq r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n \underset{a\sim\pi}{E}\left[Q(s_{t+n}, a)\right]$$

$$= \sum_{k=1}^{n} \gamma^{k-1} r_{t+k} + \gamma^n \underset{a\sim\pi}{E}\left[Q(s_{t+n}, a)\right]$$

- $G_t^{(1)}$ is a TD target: high bias, low variance
- $G_t^{(T-t)}$ is a Monte Carlo target: no bias, high variance

Look far (n steps) in the future:

$$
\begin{aligned}
G_t^{(n)} &\doteq r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n \mathop{E}_{a\sim\pi}\left[Q(s_{t+n}, a)\right] \\
&= r_{t+1} + \gamma \mathop{E}_{a\sim\pi}\left[Q(s_{t+n}, a)\right]
\end{aligned}
$$

- $G_t^{(1)}$ is a TD target: high bias, low variance
- $G_t^{(T-t)}$ is a Monte Carlo target: no bias, high variance

Look far (n steps) in the future:

$$
\begin{aligned}
G_t^{(n)} &\doteq r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n \underset{a\sim\pi}{E}\left[Q(s_{t+n}, a)\right] \\
&= \sum_{k=1}^{T-t} \gamma^{k-1} r_{t+k} + \gamma^n \underset{a\sim\pi}{E}\left[Q(s_{t+n}, a)\right]
\end{aligned}
$$

- $G_t^{(1)}$ is a TD target: high bias, low variance
- $G_t^{(T-t)}$ is a Monte Carlo target: no bias, high variance

n-step Expected Sarsa

Look far (n steps) in the future:

$$G_t^{(n)} \doteq r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n \mathop{E}_{a\sim\pi} [Q(s_{t+n}, a)]$$
$$=$$

- $G_t^{(1)}$ is a TD target: high bias, low variance
- $G_t^{(T-t)}$ is a Monte Carlo target: no bias, high variance

How can we use all future returns while keeping the variance low?

Forward View

Average **all** n-step targets:

$$G_t^{\lambda} \doteq (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}G_t^{(n)}$$

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Forward View

Average **all** n-step targets:

$$G_t^{\lambda} = (1-\lambda) \sum_{n=1}^{T-t-1} \lambda^{n-1} G_t^{(n)} + \lambda^{T-t-1} G_t^{(T-t)}$$

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Forward View

Average **all** n-step targets:

$$G_t^{\lambda} = (1-0)0^0 G_t^{(1)} + (1-0) \sum_{n=2}^{T-t-1} 0^{n-1} G_t^{(n)} + 0^{T-t-1} G_t^{(T-t)}$$

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Forward View

Average **all** n-step targets:

$$G_t^\lambda = (1-1)\sum_{n=1}^{T-t-1} 1^{n-1} G_t^{(n)} + 1^{T-t-1} G_t^{(T-t)}$$

- $\lambda = 0$ gives $G_t^{(1)}$, the TD target
- $\lambda = 1$ gives $G_t^{(T-t)}$, the Monte Carlo target

Backward View

Video here

Replacing Traces

Update **all states** at each step t:

$$e(s,a) \leftarrow (1 - \mathbb{1}\{s = s_t, a = a_t\})\gamma\lambda e(s,a) + \mathbb{1}\{s = s_t, a = a_t\}$$

$$\Delta Q(s,a) = \alpha e(s,a)(G_t^{(1)} - Q(s,a))$$

Replacing Traces

Update **all states** at each step t:

$$e \leftarrow (1 - \mathbb{1}_t)\gamma\lambda e(s, a) + \mathbb{1}_t$$
$$\Delta Q = \alpha e \delta_t$$

Back to the forward view

$$\Delta Q(s_t, a_t) =$$

$$\alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} (\gamma\lambda)^{k-j} \mathbb{1}\left\{ s_j, a_j = s_t, a_t \right\} \right]$$

Off-Policy Learning

Two policies:

- **Target** policy π: the one that is evaluated/improved
- **Behavioral** policy μ: the one that is used to interact with the environment

Potentially, both can change!

Advantages:

- Separate exploration from evaluation
- Reuse past experience

Importance Sampling

Correct the update with likelihood ratios

$$\Delta Q(s_t, a_t) = \alpha \prod_{k \geq t} \frac{\pi(a_k \mid s_k)}{\mu(a_k \mid s_k)} \left[G_t^{(n)} - Q(s_t, a_t) \right]$$

Importance Sampling

Correct the update with likelihood ratios

$$\Delta Q(s_t, a_t) = \alpha \prod_{k \geq t} \frac{\pi(a_k \mid s_k)}{\mu(a_k \mid s_k)} \left[G_t^{(n)} - Q(s_t, a_t) \right]$$

Issue: high variance!

Approximate $\mathcal{T}^*$ directly

$$\Delta Q(s_t, a_t) = \alpha \left[r_{t+1} + \gamma \max_{a \in \mathcal{A}} Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$

Q-learning

Approximate $\mathcal{T}^*$ directly

$$\Delta Q(s_t, a_t) = \alpha \left[r_{t+1} + \gamma \max_{a \in \mathcal{A}} Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$

Implicit target policy: $\pi_{\text{greedy}(Q)}$

Approximate $\mathcal{T}^*$ directly

$$\Delta Q(s_t, a_t) = \alpha \left[G_t^* - Q(s_t, a_t) \right]$$

Implicit target policy: $\pi_{\text{greedy}(Q)}$

Maximization as a special case of expectation
$\implies G_t^* = G_t^{(1)}$ when π is a greedy policy
$\implies$ Q-learning is a special case of Expected Sarsa!

Credit Assignment

Watkins' Q(λ)

Cut the eligibility trace each time μ performs a non-greedy action

For all s, a

$$e \leftarrow \gamma\lambda e + \mathbb{1}_t$$

$$\Delta Q(s,a) = \alpha e(s,a)\left[G_t^* - Q(s,a)\right]$$

If $a_t \neq \arg\max_{a' \in \mathcal{A}} Q(s_t, a')$

$$e(s,a) \leftarrow 0$$

Credit is assigned **up to the last greedy action**

Issues:

- Traces are cut too often
- Convergence was an open problem since 1989!

General Off-policy Value Update

Back to the forward view

$$\Delta Q(s_t, a_t) =$$
$$\alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} (\gamma\lambda)^{k-j} \mathbb{1}\left\{s_j, a_j = s_t, a_t\right\} \right]$$

Back to the forward view

$$\Delta Q(s_t, a_t) =$$

$$\alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \lambda^{k-j} \mathbb{1}\left\{ s_j, a_j = s_t, a_t \right\} \right]$$

General Off-policy Value Update

Back to the forward view

$$\Delta Q(s_t, a_t) =$$

$$\alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} c_i \right) \mathbb{1}\left\{ s_j, a_j = s_t, a_t \right\} \right]$$

General Off-policy Value Update

Back to the forward view

$$\Delta Q(s_t, a_t) =$$
$$\alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \mathbf{c_i} \right) \mathbb{1}_{jt} \right]$$

In this formulation we call the $\mathbf{c_i}$ "traces".

General Off-policy Prediction

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \mathbf{c_i} \right) \mathbb{1}_{jt} \right]$$

- $c_i = \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$
- $c_i = \lambda$
- $c_i = \lambda \pi(a_i \mid s_i)$

General Off-policy Prediction

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \mathbf{c_i} \right) \mathbb{1}_{jt} \right]$$

- $c_i = \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)} \implies$ Importance Sampling!
- $c_i = \lambda$
- $c_i = \lambda \pi(a_i \mid s_i)$

General Off-policy Prediction

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \mathbf{c_i} \right) \mathbb{1}_{jt} \right]$$

- $c_i = \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)} \implies$ Importance Sampling!
- $c_i = \lambda \implies Q^{\pi}(\lambda)$
- $c_i = \lambda \pi(a_i \mid s_i)$

General Off-policy Prediction

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \mathbf{c_i} \right) \mathbb{1}_{jt} \right]$$

- $c_i = \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)} \implies$ Importance Sampling!
- $c_i = \lambda \implies Q^\pi(\lambda)$
- $c_i = \lambda\pi(a_i \mid s_i) \implies TB(\lambda)$

$Q^\pi(\lambda)$

- Traces: $c_i = \lambda$
- Idea:
- Low variance
- Issue: not safe
  Convergence only if $\pi \simeq \mu$:

$$\|\pi - \mu\|_1 \leq \frac{1-\gamma}{\lambda\gamma}$$

Tree-Backup $TB(\lambda)$

- Traces: $c_i = \lambda\pi(a_i \mid s_i)$
- Idea:
- Convergence for any π, μ
- Issue: not efficient
  Traces are cut unnecessarily when almost on-policy

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \mathbf{c_i} \right) \mathbb{1}_{jt} \right]$$

Algorithm	Trace c_i	Issue
IS	$\frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)}$	High variance
$Q^\pi(\lambda)$	λ	Not safe off-policy
$TB(\lambda)$	$\lambda\pi(a_i \mid s_i)$	Not efficient on-policy

We want an algorithm that has **low variance**, is **safe** and **efficient**

Theorem (Off-policy prediction)

For any π and μ, assuming the state space is finite and all states are visited infinitely often:

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \quad \textit{then} \quad Q \to Q^{\pi}$$

Safety

Theorem (Off-policy prediction)

For any π and μ, assuming the state space is finite and all states are visited infinitely often:

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \quad \textit{then} \quad Q \to Q^\pi$$

Proof sketch: Define the off-policy operator $\mathcal{R}$:

$$\mathcal{R}Q(s,a) \doteq Q(s,a) + \mathop{E}_{a_t \sim \mu}\left[\sum_{t \geq 0} \gamma^t \left(\prod_{i=1}^{t} c_i\right)\left(G_t^{(1)} - Q(s_t, a_t)\right)\right]$$

Show that Q^π is the unique fixed point of $\mathcal{R}$

Show that $\|\mathcal{R}Q - Q^\pi\|_\infty \leq \gamma \|Q - Q^\pi\|_\infty$

Theorem (Off-policy prediction)

For any π and μ, assuming the state space is finite and all states are visited infinitely often:

$$\textit{If} \quad 0 \le c_i \le \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \quad \textit{then} \quad Q \to Q^\pi$$

Proof sketch: Define the off-policy operator $\mathcal{R}$:

$$\mathcal{R}Q(s,a) \doteq Q(s,a) + \mathop{E}_{a_t \sim \mu}\left[\sum_{t \ge 0} \gamma^t \left(\prod_{i=1}^{t} c_i\right)\left(G_t^{(1)} - Q(s_t, a_t)\right)\right]$$

Show that Q^π is the unique fixed point of $\mathcal{R}$

Show that $\|\mathcal{R}Q - Q^\pi\|_\infty \le \gamma \|Q - Q^\pi\|_\infty$

Theorem (Off-policy prediction)

For any π and μ, assuming the state space is finite and all states are visited infinitely often:

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \quad \textit{then} \quad Q \to Q^\pi$$

Proof sketch: Define the off-policy operator $\mathcal{R}$:

$$\mathcal{R}Q(s,a) \doteq Q(s,a) + \mathop{E}_{a_t \sim \mu}\left[\sum_{t \geq 0} \gamma^t \left(\prod_{i=1}^{t} c_i\right)\left(G_t^{(1)} - Q(s_t, a_t)\right)\right]$$

Show that Q^π is the unique fixed point of $\mathcal{R}$

Show that $\|\mathcal{R}Q - Q^\pi\|_\infty \leq \gamma \|Q - Q^\pi\|_\infty$

Theorem (Off-policy prediction)

For any π and μ, assuming the state space is finite and all states are visited infinitely often:

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \quad \textit{then} \quad Q \to Q^\pi$$

Proof sketch: Define the off-policy operator $\mathcal{R}$:

$$\mathcal{R}Q(s,a) \doteq Q(s,a) + \underset{a_t \sim \mu}{E}\left[\sum_{t \geq 0} \gamma^t \left(\prod_{i=1}^{t} c_i\right)\left(G_t^{(1)} - Q(s_t, a_t)\right)\right]$$

Show that Q^π is the unique fixed point of $\mathcal{R}$

Show that $\|\mathcal{R}Q - Q^\pi\|_\infty \leq \gamma \|Q - Q^\pi\|_\infty$

Theorem (Off-policy prediction)

For any π and μ, assuming the state space is finite and all states are visited infinitely often:

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \quad \textit{then} \quad Q \to Q^\pi$$

- **Safety**: ensured
- **Variance**: maximal when $c_i = \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$
- **Efficiency**: minimal when $c_i = 0$

Retrace(λ) for prediction

$$a_t \sim \mu(\cdot \mid s_t)$$

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[\left(G_k^{(1)} - Q(s_k, a_k) \right) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \mathbf{c_i} \right) \mathbb{1}_{jt} \right]$$

$$c_i = \lambda \min \left\{ 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right\}$$

Retrace(λ) for prediction

$$a_t \sim \mu(\cdot \mid s_t)$$

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} c_i \right) \mathbb{1}_{jt} \right]$$

$$c_i = \lambda \min \left\{ 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right\}$$

- **Safe** since $0 \leq c_i \leq \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$
- **Low variance** since $c_i \leq 1$
- **Efficient on-policy** since $c_i \to 1$ as $\mu \to \pi$

Retrace(λ) for prediction

$$a_t \sim \mu(\cdot \mid s_t)$$

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} c_i \right) \mathbb{1}_{jt} \right]$$

$$c_i = \lambda \min \left\{ 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right\}$$

- **Safe** since $0 \leq c_i \leq \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$
- **Low variance** since $c_i \leq 1$
- **Efficient on-policy** since $c_i \to 1$ as $\mu \to \pi$

Retrace(λ) for prediction

$$a_t \sim \mu(\cdot \mid s_t)$$

$$\Delta Q(s_t, a_t) = \alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} c_i \right) \mathbb{1}_{jt} \right]$$

$$c_i = \lambda \min \left\{ 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right\}$$

- **Safe** since $0 \leq c_i \leq \frac{\pi(a_i|s_i)}{\mu(a_i|s_i)}$
- **Low variance** since $c_i \leq 1$
- **Efficient on-policy** since $c_i \to 1$ as $\mu \to \pi$

Retrace(λ) for control

Theorem (Off-policy control)

For any μ_k, if π_k is ***increasingly greedy*** *w.r.t to Q_k:*

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi_k(a_i \mid s_i)}{\mu_k(a_i \mid s_i)} \quad \textit{then} \quad Q_k \to Q^*$$

Theorem (Off-policy control)

For any μ_k, if π_k is ***increasingly greedy*** *w.r.t to Q_k:*

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi_k(a_i \mid s_i)}{\mu_k(a_i \mid s_i)} \quad \textit{then} \quad Q_k \to Q^*$$

Proof sketch: Show that

$$\|\mathcal{R}Q_k - Q^*\|_\infty \leq \gamma \|Q_k - Q^*\|_\infty + \epsilon_k \|Q_k\|_\infty$$

Then $Q_k \to Q^*$ as $\epsilon_k \to 0$

Theorem (Off-policy control)

*For any μ_k, if π_k is **increasingly greedy** w.r.t to Q_k:*

$$\text{If} \quad 0 \leq c_i \leq \frac{\pi_k(a_i \mid s_i)}{\mu_k(a_i \mid s_i)} \quad \text{then} \quad Q_k \to Q^*$$

Proof sketch: Show that

$$\|\mathcal{R} Q_k - Q^*\|_\infty \leq \gamma \, \|Q_k - Q^*\|_\infty + \epsilon_k \, \|Q_k\|_\infty$$

Then $Q_k \to Q^*$ as $\epsilon_k \to 0$

Retrace(λ) for control

Theorem (Off-policy control)

*For any μ_k, if π_k is **increasingly greedy** w.r.t to Q_k:*

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi_k(a_i \mid s_i)}{\mu_k(a_i \mid s_i)} \quad \textit{then} \quad Q_k \to Q^*$$

Proof sketch: Show that

$$\|\mathcal{R}Q_k - Q^*\|_\infty \leq \gamma \|Q_k - Q^*\|_\infty + \epsilon_k \|Q_k\|_\infty$$

Then $Q_k \to Q^*$ as $\epsilon_k \to 0$

Retrace(λ) for control

Theorem (Off-policy control)

*For any μ_k, if π_k is **increasingly greedy** w.r.t to Q_k:*

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi_k(a_i \mid s_i)}{\mu_k(a_i \mid s_i)} \quad \textit{then} \quad Q_k \to Q^*$$

Proof sketch: Show that

$$\|\mathcal{R} Q_k - Q^*\|_\infty \leq \gamma \, \| Q_k - Q^* \|_\infty + \epsilon_k \, \| Q_k \|_\infty$$

Then $Q_k \to Q^*$ as $\epsilon_k \to 0$

Retrace(λ) for control

Theorem (Off-policy control)

For any μ_k, if π_k is ***increasingly greedy*** *w.r.t to Q_k:*

$$\textit{If} \quad 0 \leq c_i \leq \frac{\pi_k(a_i \mid s_i)}{\mu_k(a_i \mid s_i)} \quad \textit{then} \quad Q_k \to Q^*$$

Remarks

- No GLIE assumption on μ

Solving open problems as a corollary

$$\alpha_k \sum_{k \geq t} \left[\left(G_k^{(1)} - Q(s_k, a_k) \right) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \lambda \min \left\{ 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right\} \right) \mathbb{1}_{jt} \right]$$

$$\alpha_k \sum_{k \geq t} \left[\left(G_k^{(1)} - Q(s_k, a_k) \right) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \lambda \min \left\{ 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right\} \right) \mathbb{1}_{jt} \right]$$

- when π is a greedy policy, $G_t^{(1)}$ is just G_t^*
- when π is a greedy policy, $c_i = \mathbb{1}\left\{ \mu_k(s_i) = \pi_{\text{greedy}}(s_i) \right\}$

Solving open problems as a corollary

$$\alpha_k \sum_{k \geq t} \left[(G_k^{(1)} - Q(s_k, a_k)) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \lambda \min \left\{ 1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)} \right\} \right) \mathbb{1}_{jt} \right]$$

- when π is a greedy policy, $G_t^{(1)}$ is just G_t^*
- when π is a greedy policy, $c_i = \mathbb{1}\left\{\mu_k(s_i) = \pi_{\text{greedy}}(s_i)\right\}$

$$\alpha_k \sum_{k \geq t} \left[\left(G_k^{(1)} - Q(s_k, a_k)\right) \sum_{j=t}^{k} \gamma^{k-j} \left(\prod_{i=j+1}^{k} \lambda \min\left\{1, \frac{\pi(a_i \mid s_i)}{\mu(a_i \mid s_i)}\right\} \right) \mathbb{1}_{jt} \right]$$

- when π is a greedy policy, $G_t^{(1)}$ is just G_t^*
- when π is a greedy policy, $c_i = \mathbb{1}\left\{\mu_k(s_i) = \pi_{\text{greedy}}(s_i)\right\}$

Watkins' $Q(\lambda)$ is a special case of Retrace(λ)

Convergence of Watkins' $Q(\lambda)$ proved after 27 years!

The Tortoise and the Hare

Retrace(λ) is more general than Watkins' $Q(\lambda)$
For instance, π_k and μ_k can both be increasingly greedy policies, with π_k converging faster than μ_k

Harm Seijen and Rich Sutton.
True online td (lambda).
In *International Conference on Machine Learning*, pages 692–700, 2014.

Harm Van Seijen, Hado Van Hasselt, Shimon Whiteson, and Marco Wiering.
A theoretical and empirical analysis of expected sarsa.
In *Adaptive Dynamic Programming and Reinforcement Learning, 2009. ADPRL'09. IEEE Symposium on*, pages 177–184. IEEE, 2009.

Accumulating Traces

Update **all states** at each step t:

$$e(s,a) \leftarrow \gamma\lambda e(s,a) + \mathbb{1}\{s = s_t, a = a_t\}$$
$$\Delta Q(s,a) = \alpha e(s,a)(G_t^{(1)} - Q(s,a))$$

Accumulating Traces

Update **all states** at each step t:

$$e \leftarrow \gamma\lambda e + \mathbb{1}_t$$
$$\Delta Q = \alpha e \delta_t$$

Dutch Traces

Update at step t:

$$e \leftarrow (1 - \alpha \mathbb{1}_t)\gamma\lambda e + \mathbb{1}_t$$
$$\Delta Q = \alpha e \delta_t$$

Seijen and Sutton, 2014 [1]

Title B1

- Lorem ipsum dolor sit amet, consectetur adipiscing elit.
- Nulla id ex ornare, gravida nisi in, ornare risus.
    1. Aenean eu posuere purus.
    2. Etiam maximus convallis libero, ac venenatis nunc sagittis nec.
- Suspendisse orci ex, pharetra vitae aliquam ac, rutrum in dui.

Theorem (Th. Name)

This is a theorem

- *Property 1;*
- *Property 2.*

Proof.

$$a + b = c \tag{1}$$

$$a = c - b \tag{2}$$

$$answer = 42 \tag{3}$$

□

Proof.

Another proof style. □

Theorem (Th. Name)

This is a theorem

- *Property 1;*
- *Property 2.*

Proof.

$$a + b = c \tag{1}$$

$$a = c - b \tag{2}$$

$$answer = 42 \tag{3}$$

□

Proof.

Another proof style. □

Title B2

Theorem (Th. Name)

This is a theorem

- *Property 1;*
- *Property 2.*

Proof.

$$a + b = c \tag{1}$$

$$a = c - b \tag{2}$$

$$answer = 42 \tag{3}$$

□

Proof.

Another proof style. □

Title C1

First column.

Second column.

Third column.

Appears with third column

Title C1

First column.

Appears with third column

Second column.

Third column.

Title C1

First column.

Appears with third column

Second column.

Third column.

Title C2

Image:

Title C3

1. lorem
2. Ipsus
    1. sub1
    2. sub3
        1. sub4
        2. sub5

Video