Lecture 5. Model free Control

Prediction : $V_\pi(s)$

Control : improve policy $V_{\pi'}(s) \geq V_\pi(s)$ $\pi' \geq \pi$

5.1 Model free

① boat, wave, wind, complicate

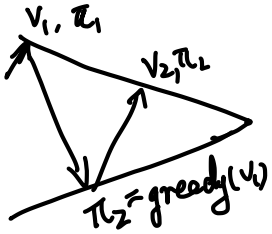
② go : know every detail, space too large

5.2 on-policy :. learn on job
learn about π from samples from π

off-policy: lean policy $\pi \neq$ sample policy μ

5.3 use V to improve

V compute accurately



policy iteration ① MC. to evaluate V

② TD evaluate V

MDP

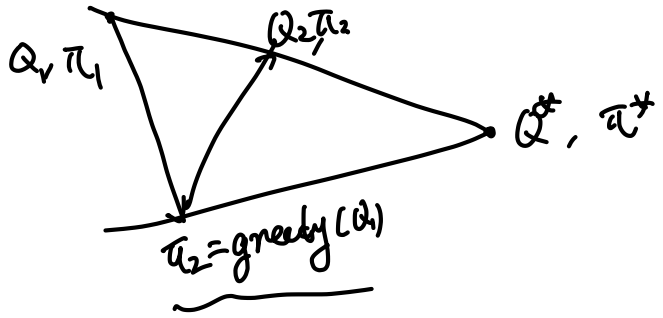
$$\pi'(s) = \arg\max_a R_s^a + \gamma P_{ss'}^a V(s')$$

Model free

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

state action value function

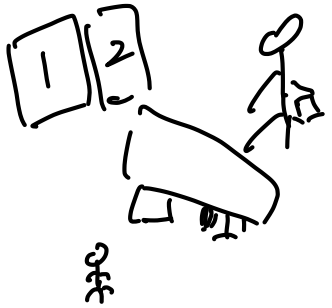
5.4 Policy iteration with action-value function



① MC
② TD

① 终点
② 奖励

example



Greedy:

left Reward 0

$V(\text{left}) = 0$

right R: +1

$V(\text{right}) = 1$

right R: +3

$V(\text{right}) = \frac{1+3}{2} = 2$

⋮

right

巨大问题

5.5 ε greedy exploration

$$\pi(a|s) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{m} & \text{if } a^* = \arg\max_a Q(s,a) \\ \frac{\varepsilon}{m} & \text{others} \end{cases}$$

1-ε greedy

ε uniform

Theorem ε-greedy policy improvement

ε-greedy π, π', ε-greedy w.r.t q_π

$\Rightarrow \quad V_{\pi'}(s) \geq V_\pi(s)$

Pf: $V_{\pi'}(s) \geq q_\pi(s, \pi'(s)) = \sum_a \pi'(a|s)\, q_\pi(s,a)$

Lect 3 DP

$$\begin{aligned} v_\pi(s) \leq q_\pi(s,\pi'(s)) &= \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s\right] \\ &\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi'(S_{t+1})) \mid S_t = s\right] \\ &\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 q_\pi(S_{t+2}, \pi'(S_{t+2})) \mid S_t = s\right] \\ &\leq \mathbb{E}_{\pi'}\left[R_{t+1} + \gamma R_{t+2} + \ldots \mid S_t = s\right] = v_{\pi'}(s) \end{aligned}$$

$V_\pi(s) \leq V_{\pi'}(s)$

$V_{\pi'}(s) \geq q_\pi(s, \pi'(s)) = \sum_a \pi'(a|s)\, q_\pi(s,a)$

$= \frac{\varepsilon}{m} \sum_a q_\pi(s,a) + (1-\varepsilon) \max_a q(s,a)$

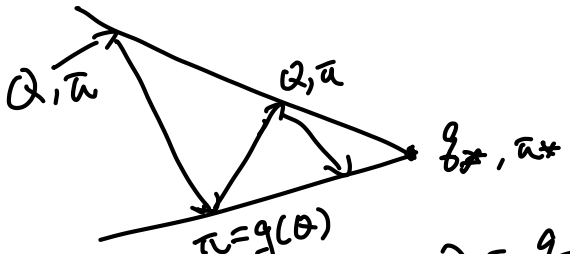
$\sum_a \pi(a|s) q_\pi(s,a) \qquad \geq \frac{\varepsilon}{m} \sum_a q_\pi(s,a) + (1-\varepsilon) \sum_a \frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon} q_\pi(s,a)$

$\sum_a \frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon} \qquad = \sum_a \pi(a|s)\, q_\pi(s,a) = V_\pi(s)$

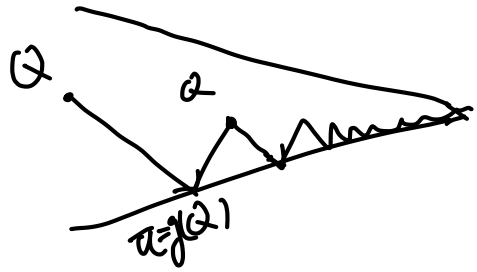
$= \frac{1-\varepsilon}{1-\varepsilon}$

$= 1$

5.6. MC. policy iteration

① 

$Q = q_\pi$

② 

$Q \approx q_\pi$

ε-greedy

Definition (GLIE)

5.6.1 Greedy in the limit with infinite exploration

① $N_k(s,a) = +\infty$ k: episode

$k \to \infty$

② $\lim_{k\to\infty} \pi_k(a|s) = 1(a = \arg\max_{a'} Q_k(s,a'))$

5.6.2 GLIE MC control

① sample kth episode using π: $\{S_1, A_1, R_2, S_2, A_2, R_3, \ldots\}$

② S_t, A_t.

$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$

$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)}(G_t - Q(S_t, A_t))$

③ $\boxed{\varepsilon \leftarrow \frac{1}{k}}$

$\pi \leftarrow \varepsilon$-greedy$(Q)$

Theorem: GLIE MC control converges to the optimal action-value function

$$Q(s,a) \rightarrow q_*(s,a)$$

5.7. MC vs TD control

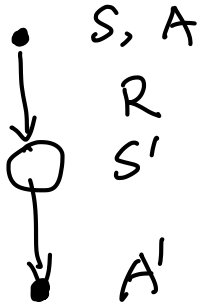
TD: lower variance

: online

: incomplete sequences

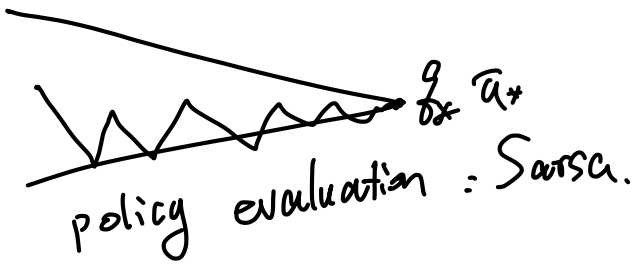
use TD to replace MC

5.7.1



SARSA

$$Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma Q(S',A') - Q(S,A))$$



policy evaluation : Sarsa.

Sarsa Algorithm for On-Policy Control

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Choose A from S using policy derived from Q (e.g., ε-greedy)
  Repeat (for each step of episode):
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma Q(S',A') - Q(S,A)\big]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

Theorem convergence of Sarsa

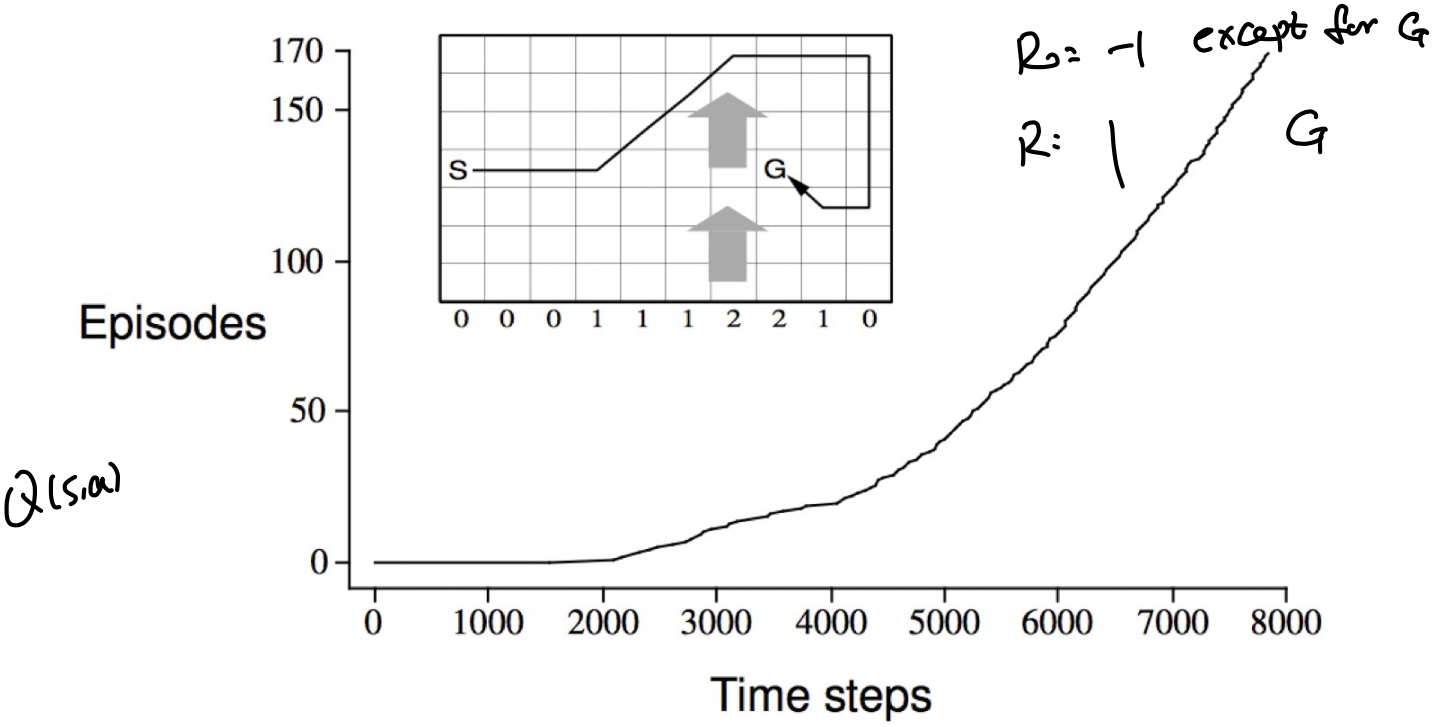
Condition:

① GLIE

② $\sum_{t=1}^{\infty} \alpha_t = +\infty$

$\sum_{t=1}^{\infty} \alpha_t^2 < +\infty$

Sarsa on the Windy Gridworld



$R_t = -1$ except for G

R: 1 G

Q(S,a)

5.8. n-step Sarsa.

$n=1$ (Sarsa) $\quad q_t^{(1)} = R_{t+1} + \gamma Q(S_{t+1})$

$n=2$ $\quad q_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q(S_{t+2})$

$\vdots$

$n=\infty$ MC $\quad q_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-1} R_T$

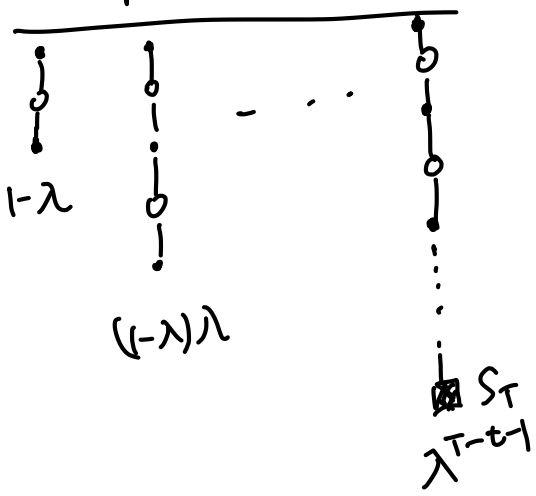
Define n-step Q-return

$$q_t^{(n)} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n Q(S_{t+n})$$

n-step Sarsa

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(q_t^{(n)} - Q(S_t, A_t) \right)$$

5.9. Forward view Sarsa(λ)



$$1 = \sum_{n=1}^{\infty} (1-\lambda)\lambda^{n-1} \qquad \frac{1-\lambda^{\infty}}{1-\lambda} \quad 1-\lambda$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(q_t^{\lambda} - Q(S_t, A_t) \right)$$

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

5.10 Backward View Sarsa (λ)

Sarsa(λ) Algorithm

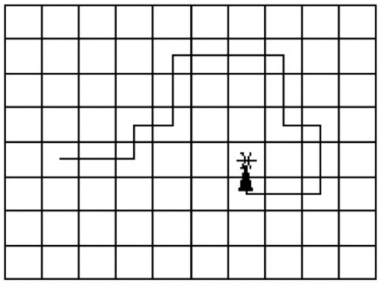
eligibility trace

Initialize $Q(s,a)$ arbitrarily, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
Repeat (for each episode):
  $E(s,a) = 0$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$
  Initialize S, A
  Repeat (for each step of episode):
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $\delta \leftarrow R + \gamma Q(S', A') - Q(S, A)$
    $E(S, A) \leftarrow E(S, A) + 1$
    For all $s \in \mathcal{S}, a \in \mathcal{A}(s)$:
      $Q(s,a) \leftarrow Q(s,a) + \alpha \delta E(s,a)$
      $E(s,a) \leftarrow \gamma \lambda E(s,a)$
    $S \leftarrow S'$; $A \leftarrow A'$
  until S is terminal

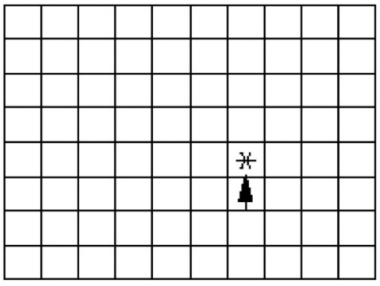
Sarsa(λ) Gridworld Example

R=0

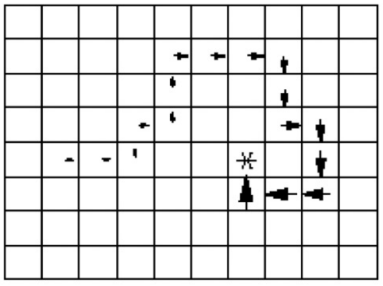
Path taken



Action values increased by one-step Sarsa



Action values increased by Sarsa(λ) with λ=0.9



5.11 off-policy learning

target policy $\pi(a|s) \rightarrow V_\pi(s)$ or $q_\pi(s,a)$

$\{S_1, A_1, R_2, S_2, A_2, \cdots S_T\} \sim \mu$

① learn from observing other agents

② Re-use experience

③ learn about optimal policy while following exploratory policy

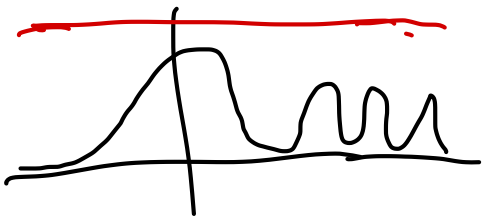
④ Train multiple policies

5.12. Importance sampling

$$E_{X\sim p}[f(X)] = \sum p(X) f(X)$$

$$= \sum f(X) \frac{p(X)}{q(X)} q(X)$$

$$= E_{X\sim q(X)}\left[f(X)\frac{p(X)}{q(X)}\right]$$



5.13. Imp. sampling for off-policy

$G_t \quad \mu(A_t|S_t) \quad \mu(A_{t+1}|S_{t+1}) \cdots$

$$G_t^{\pi/\mu} = \frac{\pi(A_t|S_t)}{\mu(A_t|S_t)} \frac{\pi(A_{t+1}|S_{t+1})}{\mu(A_{t+1}|S_{t+1})} \cdots \frac{\pi(A_T|S_T)}{\mu(A_T|S_T)} G_t$$

MC $\quad V(S_t) \leftarrow V(S_t) + \alpha\left(G_t^{\pi/\mu} - V(S_t)\right)$

TD $V(S_t) \leftarrow V(S_t) + \alpha\left(\frac{\pi(A_t|S_t)}{\mu(A_t|S_t)}(R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right)$

S.14 $\pi(S_{t+1}) = \arg\max_{a'} Q(S_{t+1}, a')$

$R_{t+1} + \gamma Q(S_{t+1}, A')$

$= R_{t+1} + \gamma Q(S_{t+1}, \arg\max_{a'} Q(S_{t+1}, a'))$

$= R_{t+1} + \gamma \max_{a'} \gamma Q(S_{t+1}, a')$

Q-learning Control : (no policy explicitly)

$(Q(S,A) \leftarrow Q(S,A) + \alpha(R + \gamma \max_{a'} Q(S', a') - Q(S,A))$

Q-Learning Algorithm for Off-Policy Control

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Repeat (for each step of episode):
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S,A) \leftarrow Q(S,A) + \alpha\left[R + \gamma \max_a Q(S',a) - Q(S,A)\right]$
    $S \leftarrow S'$;
  until S is terminal

Relationship Between DP and TD (2)

Full Backup (DP)	*Sample Backup (TD)*
Iterative Policy Evaluation	TD Learning
$V(s) \leftarrow \mathbb{E}\left[R + \gamma V(S') \mid s\right]$	$V(S) \overset{\alpha}{\leftarrow} R + \gamma V(S')$
Q-Policy Iteration	Sarsa
$Q(s,a) \leftarrow \mathbb{E}\left[R + \gamma Q(S',A') \mid s,a\right]$	$Q(S,A) \overset{\alpha}{\leftarrow} R + \gamma Q(S',A')$
Q-Value Iteration	Q-Learning
$Q(s,a) \leftarrow \mathbb{E}\left[R + \gamma \max_{a' \in \mathcal{A}} Q(S',a') \mid s,a\right]$	$Q(S,A) \overset{\alpha}{\leftarrow} R + \gamma \max_{a' \in \mathcal{A}} Q(S',a')$

where $x \overset{\alpha}{\leftarrow} y \equiv x \leftarrow x + \alpha(y - x)$