Generalized Policy Iteration (Page1-12)
N-Step TD method (Page 13)
Off-Policy MC method (Page 14-16)

Prof. Subrahmanya Swamy

Challenges of Policy Iteration in Model-Free Context

Policy Iteration

$\pi_0 \rightarrow V_{\pi_0} \rightarrow \pi_1 \rightarrow V_{\pi_1} \rightarrow \pi_2 \rightarrow V_{\pi_2} \rightarrow \ldots$

~~Policy Evaluation~~ MC / TD

Policy Improvement

$V_{\pi_1} \geq V_{\pi_0}$

Repeat till
$\pi_{k+1} = \pi_k$
$\Downarrow$
$\pi_k = \pi_*$

- ***Policy Evaluation:*** $V_{k+1}(s) = R_s^\pi + \sum_{s'} P_{ss'}^\pi V_k(s')$
- *Policy Improvement:* $\pi_{i+1}(s) := argmax_a \ R_s^a + \sum_{s'} P_{ss'}^a V_{\pi_i}(s')$

1. Policy Evaluation requires model dynamics

Solution:

- Don't use Iterative ~~Policy~~ Policy Evaluation to estimate V_π
- Instead use MC/TD methods to estimate V_π

Challenges of Policy Iteration in Model-Free Context

Policy Iteration

$$\pi_0 \xrightarrow{\text{PE}} V_{\pi_0} \xrightarrow{\text{PI}} \pi_1 \xrightarrow{\text{PE}} V_{\pi_1} \xrightarrow{\text{PI}} \pi_2 \xrightarrow{\text{PE}} V_{\pi_2} \rightarrow \dots$$

$Q_{\pi_0}(s,a)$

$Q_\pi(s,a)$

- *Policy Evaluation:* $V_{k+1}(s) = R_s^\pi + \sum_{s'} P_{ss'}^\pi V_k(s')$
- ***Policy Improvement:*** $\pi_{i+1}(s) := argmax_a R_s^a + \sum_{s'} P_{ss'}^a V_{\pi_i}(s')$

2. Policy Improvement requires model dynamics

$\pi_{i+1}(s) = \underset{a}{\operatorname{argmax}}\, Q_{\pi_i}(s,a)$

Solution:

$$\textbf{PI} \quad \pi_{i+1}(s) := argmax_a \left(R_s^a + \sum_{s'} P_{ss'}^a V_{\pi_i}(s') \right)$$

$$= argmax_a\, Q_{\pi_i}(s,a)$$

- If Q_{π_i} is known, model dynamics not required for PI
- Hence, estimate Q_π instead of V_π in the PE step

How to estimate $Q_\pi(s, a)$?

π

V_π MC / TD

MC method to estimate V_π

- $V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$ $\rightarrow$ $Q_\pi(s, a) = E_\pi[G_t \mid S_t = s, A_t = a]$

- Generate multiple episodes starting from s
  - Episode 1: $S_0 = s,\ A_0 \sim \pi,\ R_1,\ S_1,\ A_1 \sim \pi,\ R_2,\ S_2,\ \ldots,\ S_T$ $\quad G^{(1)}$
  - Episode 2: $S_0 = s,\ A_0 \sim \pi,\ R_1,\ S_1,\ A_1 \sim \pi,\ R_2,\ S_2,\ \ldots,\ S_T$ $\quad G^{(2)}$
  - …
  - …

  $\vdots$

  $G^{(N)}$

- Compute sample returns of each episode from state s
  - $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$

- $V_\pi(s) \approx$ sample avg of the returns

MC method to estimate Q_π

TD

Q_π ?

- $Q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$

(s,a)

- Generate multiple episodes starting from (s,a)
  - Episode 1: $S_0 = s,\ A_0 = a,\ R_1,\ S_1,\ A_1 \sim \pi,\ R_2,\ S_2,\ \ldots,\ S_T$ → $G^{(1)}$
  - Episode 2: $S_0 = s,\ A_0 = a,\ R_1,\ S_1,\ A_1 \sim \pi,\ R_2,\ S_2,\ \ldots,\ S_T$ → $G^{(2)}$
  - …
  - …

$\vdots$

$G^{(N)}$

- Compute sample returns of those episodes starting from (s,a)
  - $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$

- $Q_\pi(s,a) \approx$ sample avg of the returns

TD Method for Q_π

$E[R_{t+1} | S_t = s, A_t = a]$

$$Q_\pi(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a V_\pi(s')$$

(s', a')

$V_\pi(s') = \left(\sum_{a'} Q_\pi(s', a') \pi(a'|s')\right)$

$Q_\pi(s,a) = R_s^a + \gamma \sum_{s'} P_{ss'}^a \left(\sum_{a'} Q_\pi(s', a') \pi(a'|s')\right)$

$= R_s^a + \gamma \sum_{s'} \sum_{a'} P_{ss'}^a Q_\pi(s', a') \pi(a'|s')$

$= E[R_{t+1} | S_t = s, A_t = a]$

$+ \gamma E_{s', a' \sim P_{ss'}^a, \pi(a'|s')}[Q_\pi(s', a')]$

$Q_\pi(s,a) = E[$

TD Method for Q_π (SARSA)

$V_\pi(s) = E_\pi[G_t \mid S_t = s]$

$$
\begin{aligned}
Q_\pi(s,a) &= \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] + \gamma \mathbb{E}[V_\pi(S_{t+1})] \\
&= \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] + \gamma \mathbb{E}[\mathbb{E}[Q_\pi(S_{t+1}, A_{t+1})]] \\
&\approx R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1})
\end{aligned}
$$

$Q_\pi(S_t, A_t)$

$V_{new}(s) = V_{old}(s) + \alpha[G_t - V_{old}(s)]$

$$Q_{new}(S_t, A_t) = Q_{old}(S_t, A_t) + \alpha\left(R_{t+1} + \gamma Q_{old}(S_{t+1}, A_{t+1}) - Q_{old}(S_t, A_t)\right)$$

SARSA

$S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}$

Q_π

MC | TD

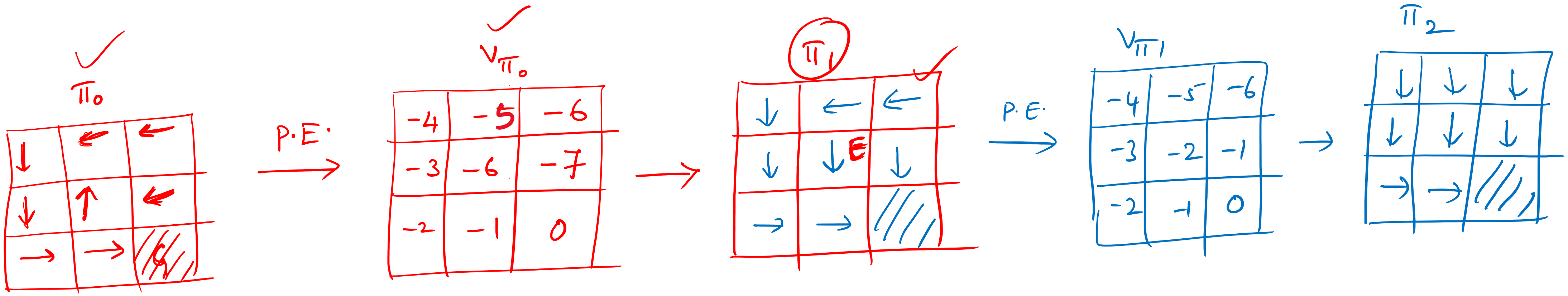
GPI

$$\pi_0 \xrightarrow[\text{P.E.}]{\text{MC/TD}} Q_{\pi_0} \xrightarrow{\text{P.I.}} \pi_1 \longrightarrow \cdots\cdots$$

$$\pi_1(s) = \underset{a}{\operatorname{argmax}}\ Q_{\pi_0}(s, a)$$

Q_π Estimation: Challenges

- **Observation:** Only Deterministic policies are encountered in Policy Iteration



$$\pi_1(s) = \arg\max_{a} \left\{ R_s^a + \gamma \sum_{s'} P^a_{ss'} V_{\pi_0}(s') \right\}$$

$$\pi_1(E) = \arg\max \begin{Bmatrix} L: & -1 + V_{\pi_0}(F) \\ R: & -1 + V_{\pi_0}(D) \\ U: & -1 + V_{\pi_0}(B) \\ D: & -1 + V_{\pi_0}(H) \end{Bmatrix} = \arg\max_a \begin{Bmatrix} L: & -1-3 \\ R: & -1-7 \\ U: & -1-5 \\ D: & -1-1 \end{Bmatrix}$$

Issue with Deterministic Policy

ϵ-greedy policies   "ϵ"

$\rightarrow \pi_1(s) = \underset{a}{\text{argmax}}\ Q_{\pi_0}(s,a)$ "$1-\epsilon$"

- Consider 3 states A, B, C
- 4 actions in each state: Left, Right, Up, Down
- Consider a deterministic policy

$$\pi(A) = Right$$
$$\pi(B) = Right$$
$$\pi(C) = Left$$

Q_π $V_\pi(A)$ $V_\pi(B)$ $V_\pi(C)$

3×4

$Q_\pi(s,a)$

$Q_\pi(A, left)$	$Q(B, L)$	$Q(C, L)$
$Q_\pi(A, R)$	$Q(B, R)$	$Q(C, R)$
$Q_\pi(A, UP)$	$Q(B, UP)$	$Q(C, U)$
$Q_\pi(A, Down)$	$Q(B, D)$	$Q(C, D)$

- Sample Episode

Q_π $\pi(B)$ $\pi(A)$ $\pi(C)$

$$A, Left, -1, B, Right, -1, A, Right, -1, C, Left, -1, \ldots$$

"S_0" "A_0"

(s, a)

$Q_\pi(B, Right)$

Generalized Policy Iteration (GPI)

Policy Iteration

$$\pi_0 \rightarrow \cancel{V_{\pi_0}} \rightarrow \pi_1 \rightarrow V_{\pi_1} \rightarrow \pi_2 \rightarrow V_{\pi_2} \rightarrow \ldots$$

~~**Policy Evaluation**~~

Policy Improvement

$$\boxed{\pi_1(s) = \arg\max_a Q_{\pi_0}(s,a)}$$

GPI

$$\pi_0 \rightarrow Q_{\pi_0} \rightarrow \pi_1 \rightarrow Q_{\pi_1} \rightarrow \pi_2 \rightarrow Q_{\pi_2} \rightarrow \ldots \quad \pi^*$$

Approx. Policy Evaluation

MC / TD

ϵ −greedy Policy Improvement

$$a^* = \arg\max_a Q_{\pi_0}(s,a)$$

$$\pi_1(a^*|s) = 1 - \varepsilon + \frac{\varepsilon}{N}$$

$$\pi_1(a|s) = \frac{\varepsilon}{N}$$

N-Step TD Method

Q_π

$\pi_0 \rightarrow Q_{\pi_0}$

MC/TD

$\rightarrow G_t = \underbrace{R_{t+1}}_{\text{Immediate reward}} + \gamma \underbrace{G_{t+1}}_{\text{remaining return}}$

$Q_\pi(s,a) \approx R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1})$

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$\boxed{G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 G_{t+2}}$

2-step TD $\quad Q_\pi(s,a) = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q_\pi(S_{t+2}, A_{t+2})$

Off–Policy MC

$$\left.\begin{array}{l} MC \\ TD \\ n\text{-step } TD \end{array}\right\} \text{on-policy}$$

V_π

π → Target policy

"b" → behaviour policy

Q_π

π

$P(\cdot x)$

$x_1, x_2, \dots x_N$

$b(x)$

$x_1, \dots x_N$

"b"

M.C

$$\mathop{E}_{x\sim p}[f(x)] \approx \frac{f(x_1) + f(x_2) + \cdots f(x_N)}{N}$$

Importance Sampling

$X \sim p \qquad f(x) = x^2$

$$\mathop{E}_{p}[f(X)] = \sum_x P(X=x)f(x)$$

Off–Policy MC

$$E_p[f(x)] = \sum_x f(x)\, P(X=x)$$

$$= \sum_x f(x)\, b(X=x) \times \frac{P(X=x)}{b(X=x)}$$

$$= \underbrace{\sum_x \left(f(x) \cdot \frac{P(X=x)}{b(X=x)} \right)}_{g(x)} b(X=x)$$

MC $V_\pi \mid Q_\pi$

$b(\cdot)$

$x_1, \ldots x_N$

$$E_p[f(x)] = \sum_x g(x)\, b(X=x) = E_b[g(x)] \approx \frac{g(x_1) + g(x_2) + \cdots g(x_N)}{N}$$

E_p

$$= \frac{1}{N}\left(f(x_1)\frac{P(x_1)}{b(x_1)} + f(x_2)\frac{P(x_2)}{b(x_2)} + \cdots \right)$$

Importance sampling Tech.

Off–Policy MC

MC for V_π.

$V_\pi(s)$

$=$

Ep1 $\quad S_0 = s, \quad A_0 \sim \pi, \quad R_1, \quad S_1, \quad A_1 \sim \pi, \quad \ldots \quad G^{(1)}$

Ep2 $\quad G^{(2)}$

$$V_\pi(s) \approx \frac{G^{(1)} + G^{(2)} + \cdots G^{(N)}}{N}$$

behaviour b.

Ep1 $\quad S_0 = s, \quad A_0 \sim b, \quad R_1, \quad S_1, \quad A_1 \sim b, \quad \ldots \quad G^{(1)}$

Ep2

$\vdots$

$G^{(N)}$

$$V_\pi(s) \approx \frac{1}{N}\left(\left(G^{(1)} \cdot \frac{P_\pi(Ep1)}{P_b(Ep1)} \right) + \left(G^{(2)} \frac{P_\pi(Ep2)}{P_b(Ep1)} \right) + \cdots \right)$$

$$\frac{P_\pi(Ep1)}{P_b(Ep1)} = \frac{\pi(A_0|S_0)\ \pi(A_1|S_1)\ \pi(A_2|S_2) \cdots}{b(A_0|S_0)\ b(A_1|S_1)\ b(A_2|S_2) \cdots}$$