Generalized Policy Iteration (Page1-10)
N-Step TD method (Page 11)
Off-Policy MC method (Page 12-14)

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

Policy Improvement

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:* $\pi_{i+1}(s) \coloneqq argmax_a \; R_s^a + \sum_{s'} P_{ss'}^a V_{\pi_i}(s')$

1. Policy Evaluation requires model dynamics

Solution:

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

Challenges of Policy Iteration in Model-Free Context

Policy Iteration

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

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

2. Policy Improvement requires model dynamics

Solution:

PI $$\pi_{i+1}(s) \coloneqq argmax_a\, R_s^a + \sum_{s'} P_{ss'}^a V_{\pi_i}(s') = argmax_a\, Q_{\pi_i}(s,a)$$

- 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 $Q_\pi(s, a)$?

MC method to estimate V_π

- $V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$

- Generate multiple episodes starting from s
  - Episode 1: $S_0 = s,\ A_0 \sim \pi,\ R_1,\ S_1,\ A_1 \sim \pi,\ R_2,\ S_2,\ \ldots,\ S_T$
  - Episode 2: $S_0 = s,\ A_0 \sim \pi,\ R_1,\ S_1,\ A_1 \sim \pi,\ R_2,\ S_2,\ \ldots,\ S_T$
  - …
  - …

- 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

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

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

- 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 (SARSA)

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

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

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

SARSA

Q_π Estimation: Challenges

- Observation: Only Deterministic policies are encountered in Policy Iteration

π_0

↓	←	←
↓	↑	←
→	→	G

P.E. →

V_{π_0}

-4	-5	-6
-3	-6	-7
-2	-1	0

→

π_1

↓	←	←
↓	↓	↓
→	→	

P.E. →

V_{π_1}

-4	-5	-6
-3	-2	-1
-2	-1	0

→

π_2

↓	↓	↓
↓	↓	↓
→	→	

$$\pi_1(s) = \text{argmax} \left\{ R_s^a + \gamma \sum_{s'} P_{ss'}^a V_{\pi_0}(s') \right\}$$

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

Issue with Deterministic Policy

- Consider 3 states A, B, C

- 4 actions in each state: Left, Right, Up, Down

- Consider a deterministic policy

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

- Sample Episode

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

Generalized Policy Iteration (GPI)

Policy Iteration

$$\pi_0 \xrightarrow{\text{Policy Evaluation}} V_{\pi_0} \xrightarrow{\text{Policy Improvement}} \pi_1 \rightarrow V_{\pi_1} \rightarrow \pi_2 \rightarrow V_{\pi_2} \rightarrow \ldots$$

GPI

$$\pi_0 \xrightarrow{\text{Approx. Policy Evaluation}} Q_{\pi_0} \xrightarrow{\epsilon-\text{greedy Policy Improvement}} \pi_1 \rightarrow Q_{\pi_1} \rightarrow \pi_2 \rightarrow Q_{\pi_2} \rightarrow \ldots$$

N-Step TD Method

$\pi_0 \xrightarrow[MC/TD]{} Q_{\pi_0}$

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

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

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

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

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

MC
TD
n-step TD $\Big\}$ on-policy

$Q_\pi \to$ Target policy

V_π

"b" $\to$ behaviour policy

Q_π

$\xrightarrow[\pi]{}$

$P(\cdot x)$

$x_1, x_2, \dots x_N$

$b(x)$

$x_1, \dots x_N$

M.C

"b"

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

Importance Sampling

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

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

Importance Sampling

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

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

MC $V_\pi | Q_\pi$

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

$b(\cdot)$

$x_1, \ldots x_N$

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

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

Importance sampling Tech.

Off–Policy MC

MC for V_π.

$V_\pi(s)$
$=$

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

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

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

behaviour b.

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

Ep2

$\vdots$

$G^{(N)}$

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

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