

Generalized Policy Iteration (Page 1-10)

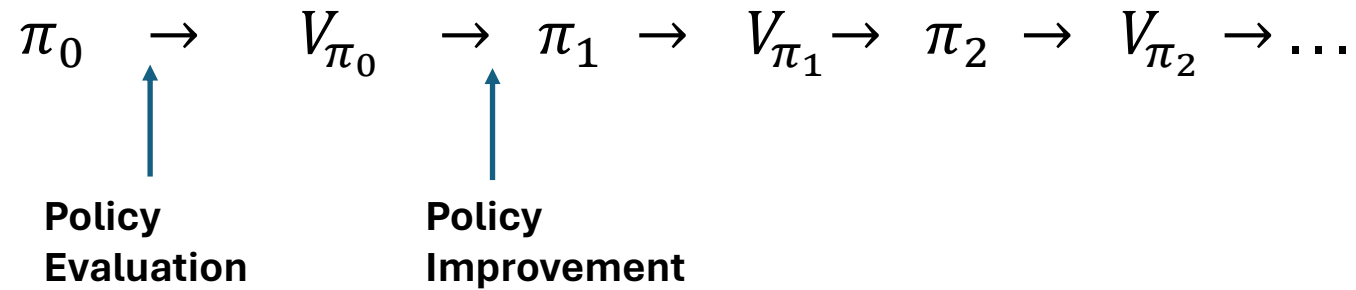
N-Step TD method (Page 11)

Off-Policy MC method (Page 12-14)

Prof. Subrahmanya Swamy

Challenges of Policy Iteration in Model-Free Context

Policy Iteration



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:** $\pi_{i+1}(s) := \operatorname{argmax}_a R_s^a + \sum_{s'} P_{ss'}^a V_{\pi_i}(s')$

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

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

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

2. Policy Improvement requires model dynamics

Solution:

$$\begin{aligned} \text{PI} \quad \pi_{i+1}(s) &:= \operatorname{argmax}_a R_s^a + \sum_{s'} P_{ss'}^a V_{\pi_i}(s') \\ &= \operatorname{argmax}_a Q_{\pi_i}(s, a) \end{aligned}$$

- 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: $S_0 = s, A_0 \sim \pi, R_1, S_1, A_1 \sim \pi, R_2, S_2, \dots, S_T$
 - Episode 2: $S_0 = s, A_0 \sim \pi, R_1, S_1, A_1 \sim \pi, R_2, S_2, \dots, S_T$
 - ...
 - ...
- Compute sample returns of each episode from state s
 - $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$
- $V_\pi(s) \approx$ 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: $S_0 = s, A_0 = a, R_1, S_1, A_1 \sim \pi, R_2, S_2, \dots, S_T$
 - Episode 2: $S_0 = s, A_0 = a, R_1, S_1, A_1 \sim \pi, R_2, S_2, \dots, S_T$
 - ...
 - ...
- Compute sample returns of those episodes starting from (s, a)
 - $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$
- $Q_\pi(s, a) \approx$ sample avg of the returns

TD Method for Q_π (SARSA)

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

$$\begin{aligned} Q_\pi(s, a) &= \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] + \gamma \mathbb{E}[V_\pi(S_{t+1})] \\ &= \mathbb{E}[\textcolor{blue}{R}_{t+1} \mid S_t = s, A_t = a] + \gamma \mathbb{E}[\mathbb{E}[\textcolor{blue}{Q}_\pi(\textcolor{blue}{S}_{t+1}, \textcolor{blue}{A}_{t+1})]] \\ &\approx \textcolor{blue}{R}_{t+1} + \gamma \textcolor{blue}{Q}_\pi(\textcolor{blue}{S}_{t+1}, \textcolor{blue}{A}_{t+1}) \end{aligned}$$

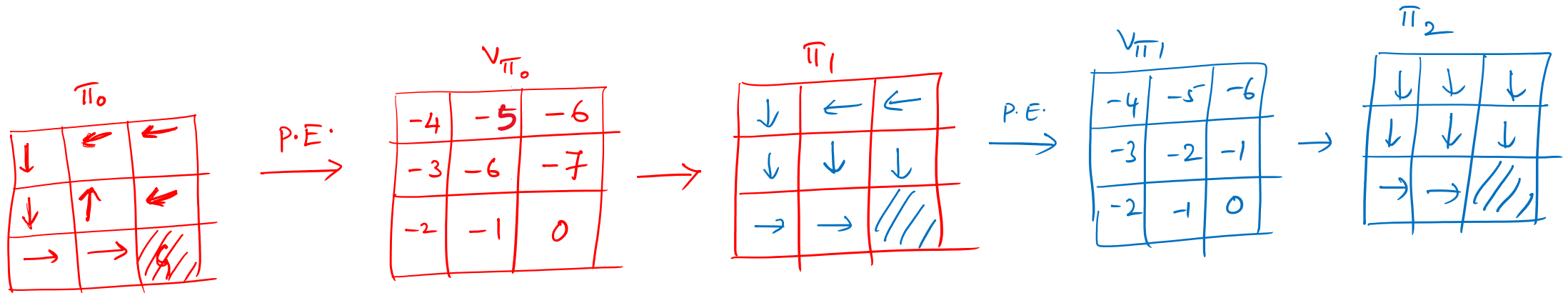
$$\textcolor{red}{Q}_{new}(S_t, A_t) = \textcolor{red}{Q}_{old}(S_t, A_t) + \alpha (\textcolor{blue}{R}_{t+1} + \gamma \textcolor{blue}{Q}_{old}(\textcolor{blue}{S}_{t+1}, \textcolor{blue}{A}_{t+1}) - \textcolor{red}{Q}_{old}(S_t, A_t))$$

SARSA



Q_π Estimation: Challenges

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



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

$$\pi_1(E) = \operatorname{argmax} \left\{ \begin{array}{l} L: -1 + V_{\pi_0}(F) \\ R: -1 + V_{\pi_0}(D) \\ U: -1 + V_{\pi_0}(B) \\ D: -1 + V_{\pi_0}(H) \end{array} \right\} = \operatorname{argmax}_a \left\{ \begin{array}{l} L: -1 - 3 \\ R: -1 - 7 \\ U: -1 - 5 \\ D: -1 - 1 \end{array} \right\}$$

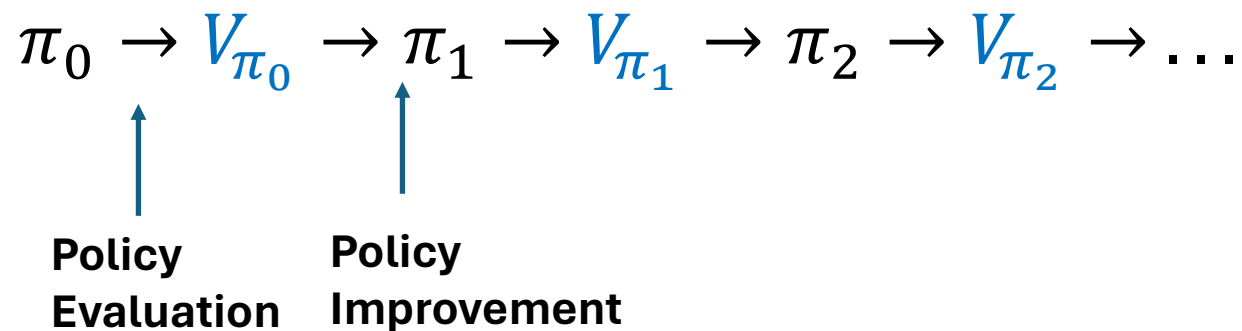
Issue with Deterministic Policy

- Consider 3 states A, B, C
- 4 actions in each state: Left, Right, Up, Down
- Consider a deterministic policy
$$\pi(A) = \textit{Right}$$
$$\pi(B) = \textit{Right}$$
$$\pi(C) = \textit{Left}$$
- Sample Episode

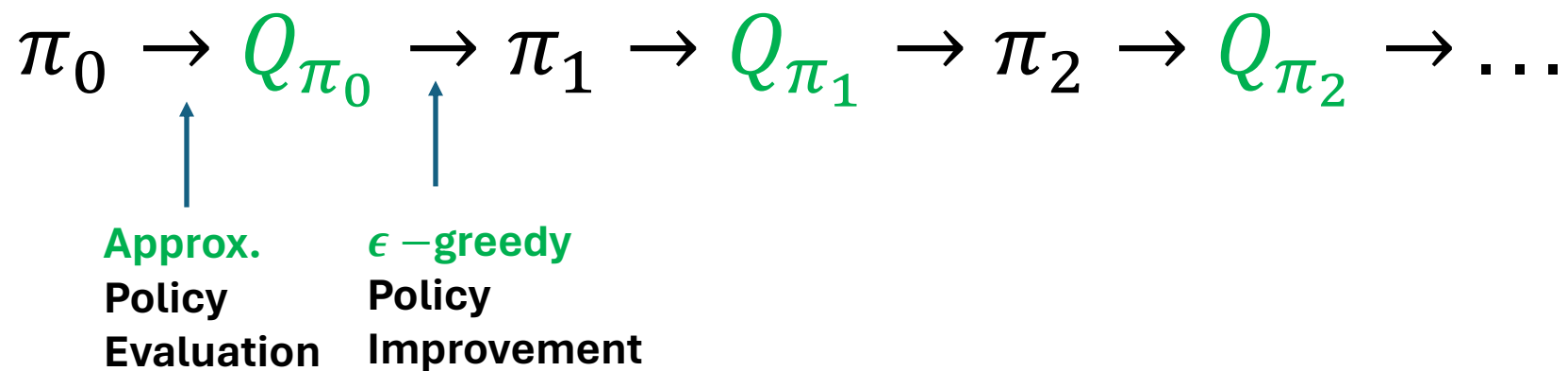
$A, \textit{Left}, -1, B, \textit{Right}, -1, A, \textit{Right}, -1, C, \textit{Left}, -1, \dots$

Generalized Policy Iteration (GPI)

**Policy
Iteration**



GPI



N-Step TD Method

$$\pi_0 \xrightarrow{\text{MC/TD}} Q_{\pi_0}$$

$$\rightarrow G_t = R_{t+1} + \gamma G_{t+1}$$

\downarrow Immediate reward \downarrow remaining return

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

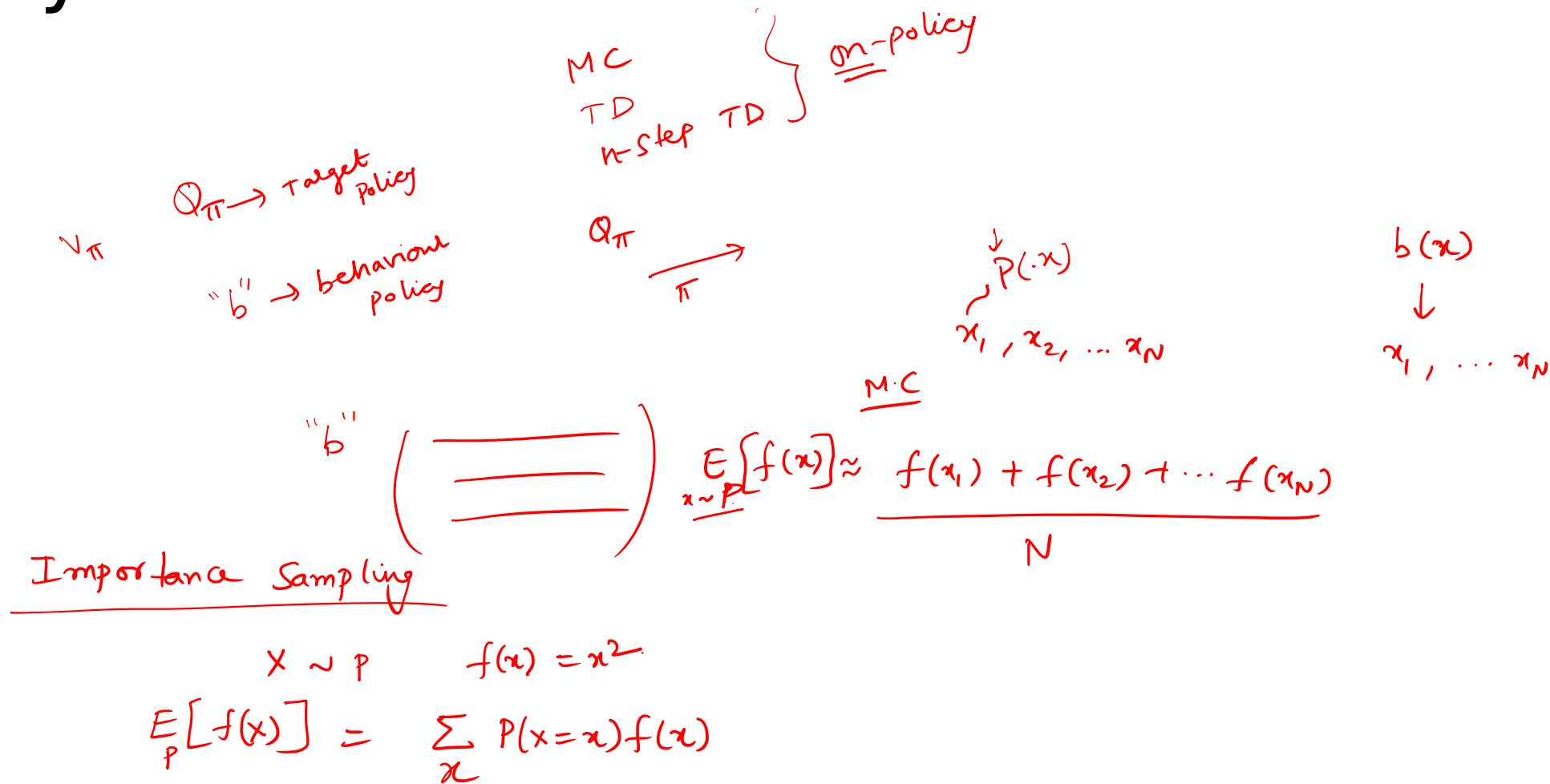
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots$$

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 \underbrace{G_{t+2}}_{\uparrow}$

2-step
TD

$$Q_{\pi}(s, a) = R_{t+1} + \gamma R_{t+2} + \gamma^2 Q_{\pi}(s_{t+2}, A_{t+2})$$

Off-Policy MC



Importance Sampling

$$\begin{aligned}
 E_P[f(x)] &= \sum_x f(x) P(x=x) \\
 &= \sum_x f(x) b(x=x) \times \frac{P(x=x)}{b(x=x)} \\
 &= \sum_x \left(f(x) \cdot \frac{P(x=x)}{b(x=x)} \right) b(x=x)
 \end{aligned}$$

MC $\sqrt{\pi} / Q\pi$

$$E_P[f(x)] = \sum_x g(x) b(x=x) = E_b[g(x)] \approx \frac{g(x_1) + g(x_2) + \dots + g(x_N)}{N}$$

\downarrow \downarrow \downarrow
 $b(\cdot)$
 x_1, \dots, x_N

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

Importance Sampling
Tech.

$V_\pi(s)$

$\Sigma p1 \quad S_0 = s, \underline{\underline{A_0 \sim \pi}}, R_1, S_1, A_1 \sim \pi, \dots \dots \dots G^{(1)}$

$\Sigma p2 \quad \dots \dots \dots G^{(2)}$

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

$\Sigma p_1 \quad S_0 = s, A_0 \sim \underline{\underline{b}}, R_1, S_1, A_1 \sim b, \dots G^{(1)}$

$$V_{\pi}(S) \approx \frac{1}{N} \left(\left(G^{(1)} \cdot \frac{P_{\pi}(\epsilon p 1)}{P_b(\epsilon p 1)} \right) + \left(G^{(2)} \cdot \frac{P_{\pi}(\epsilon p 2)}{P_b(\epsilon p 1)} \right) + \dots \right)$$

$$\frac{P(\Sigma P_i)}{P_\Sigma(\Sigma P_i)} = \frac{\pi(A_0|S_0) \pi(A_1|S_1) \pi(A_2|S_2) \dots}{b(A_0|S_0) b(A_1|S_1) b(A_2|S_2) \dots}$$