Lecture 3

Review

- $p(s', r \mid s, a)$
  $= P(S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a)$
- $MDP = \langle S, A, R, T, \gamma \rangle$

- Trajectory

$$= \{S_0, A_0, R_1, S_1, A_1, \ldots\}$$

- Return

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

- Value

$$V^{\pi}(s) = E_{\pi,T}[G_t \mid S_t = s]$$

- Q-Value

$$Q^{\pi}(s,a) = E[G_t \mid S_t = s, A_t = a$$

- Optimal-Value

$$V^*(s) = \max_{\pi} V^{\pi}(s)$$

- $Q^*(s,a) = \max_\pi Q^\pi(s,a)$
- $V^*(s) = \max_a Q^*(s,a)$
- $Q^*(s,a) = \mathbb{E}\left[R_{t+1} + \gamma V^*(s_{t+1}) \mid S_t = s, A_t = a \right]$
- Bellman Eq

$$V^\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) \cdot \left[r + \gamma V^\pi(s') \right]$$

- Bellman Optimality Eq

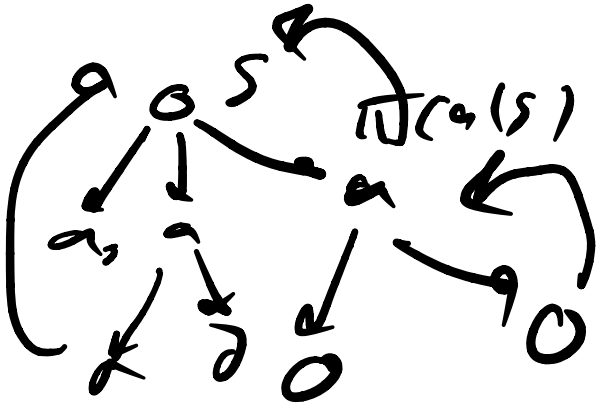
$$V^*(s) = \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma V^*(s')\right]$$

$$Q^*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} Q^*(s',a')\right]$$

- Dynamic Programming

→ Policy Evaluation

$$V_{k+1}(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \gamma V_k(s')\right]$$



→ Policy improvement

$$Q^{\pi}(s, \pi'(s)) \geq V^{\pi}(s)$$

$$\pi' = \arg\max_{a} Q^{\pi}(s, a)$$

→

Policy Iteration

1.) Initialise $V(s)$ and $\pi(s)$

2.) Policy Evaluation $\Rightarrow$ $V^{\pi}(s)$

3) Policy Improvement

$$\pi \longrightarrow \pi'$$

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

4) Repeat

• Value Iteration

1) Initial $V(s)$

2) { one-step policy evaluati
   { one-step policy improvem

$$V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)\left[r + \gamma V(s')\right]$$

3) Repeat

Monte Carlo Method

- Previously, we assume knowledge over MDP

  i.e. $p(s', r \mid s, a)$

- Now, let's say we don't know $p(s', r \mid s, a)$

We want to find

$V^{\pi}(s)$, $V^{*}(s)$, $Q^{\pi}(s, a)$, $Q^{*}(s, a)$, π^{*}

- We could estimate

$$P(s', r \mid s, a)$$

from data

This is called

"model-based" method

- We could learn

$V^{\pi}(s)$, $\pi(s)$, etc

directly from data

"Model-free" method

$$V^{\pi}(s) = \mathbb{E}[G_t \mid S_t = s]$$

$$V^{\pi}(s) = \mathbb{E}[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots]$$

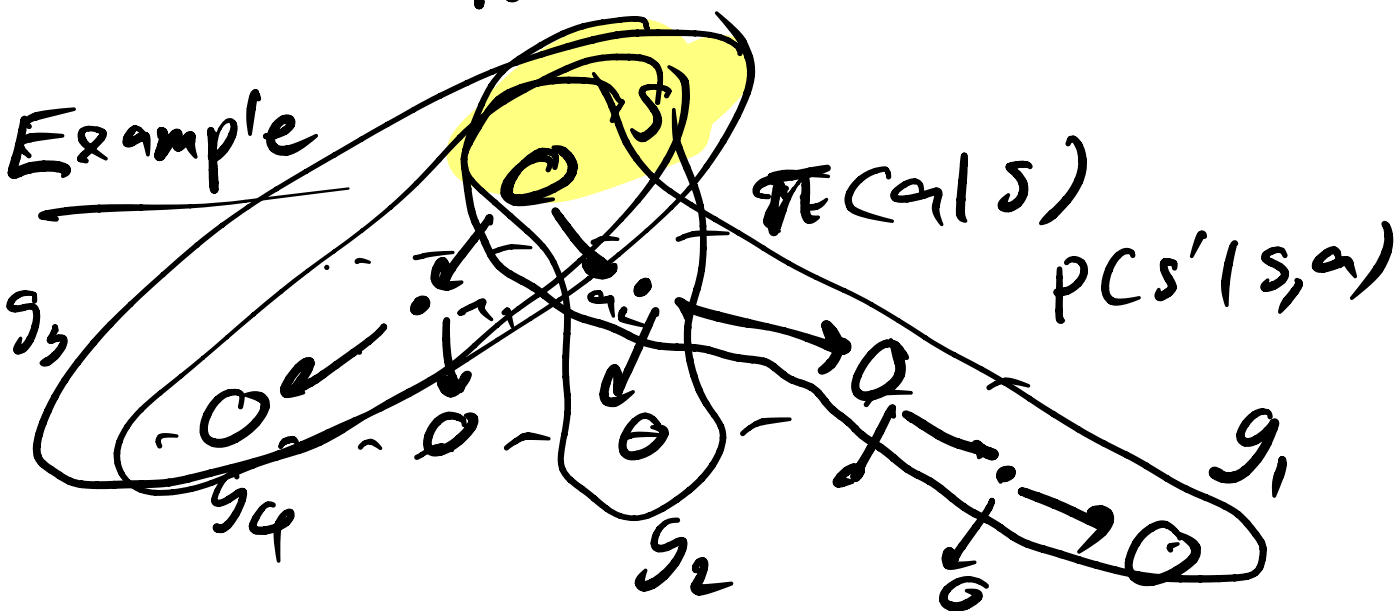
$$\mathbb{E}[G] = \sum_{g_t} p(g_t|s)\, g_t$$

- $$\mathbb{E}[G] \approx \frac{1}{N} \sum_{i}^{N} g_i$$

a Monte-Carlo Estimation

- $$V^{\pi}(s) \approx \frac{1}{N} \sum_{i}^{N} \left(r_t^i + \gamma r_{t+1}^i + \gamma^2 r_{t+2}^i + \dots \right)$$

Example



$$V^{\pi}(s) \approx \frac{1}{\varphi}(g_1 + g_2 + g_3 + g_{\varphi})$$

First-Visit MC estimation V^{π}

0.) input π

1.) Initialise

$V(s)$, $G_s = []$

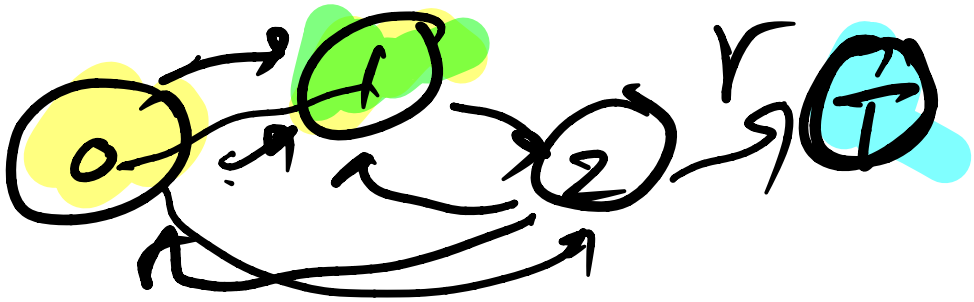
2.) Generate a trajectory with π

3.) For each s appear

- $g(s) = \sum_i \gamma^i r$
- Append $g(s) \rightarrow G_s$

4.) Compute $V(s) \leftarrow \text{Average}[G_s]$

Example



$$\tau = \{ s_0 = 0, a_0 = 1, r_1 = 1$$

$$s_1 = 1, a_1 = 0, r_2 = 0$$

$$s_2 = 0, a_2 = 1, r_3 = 1$$

$$s_3 = 2, a_3 = 1, r_4 = 1 \}$$

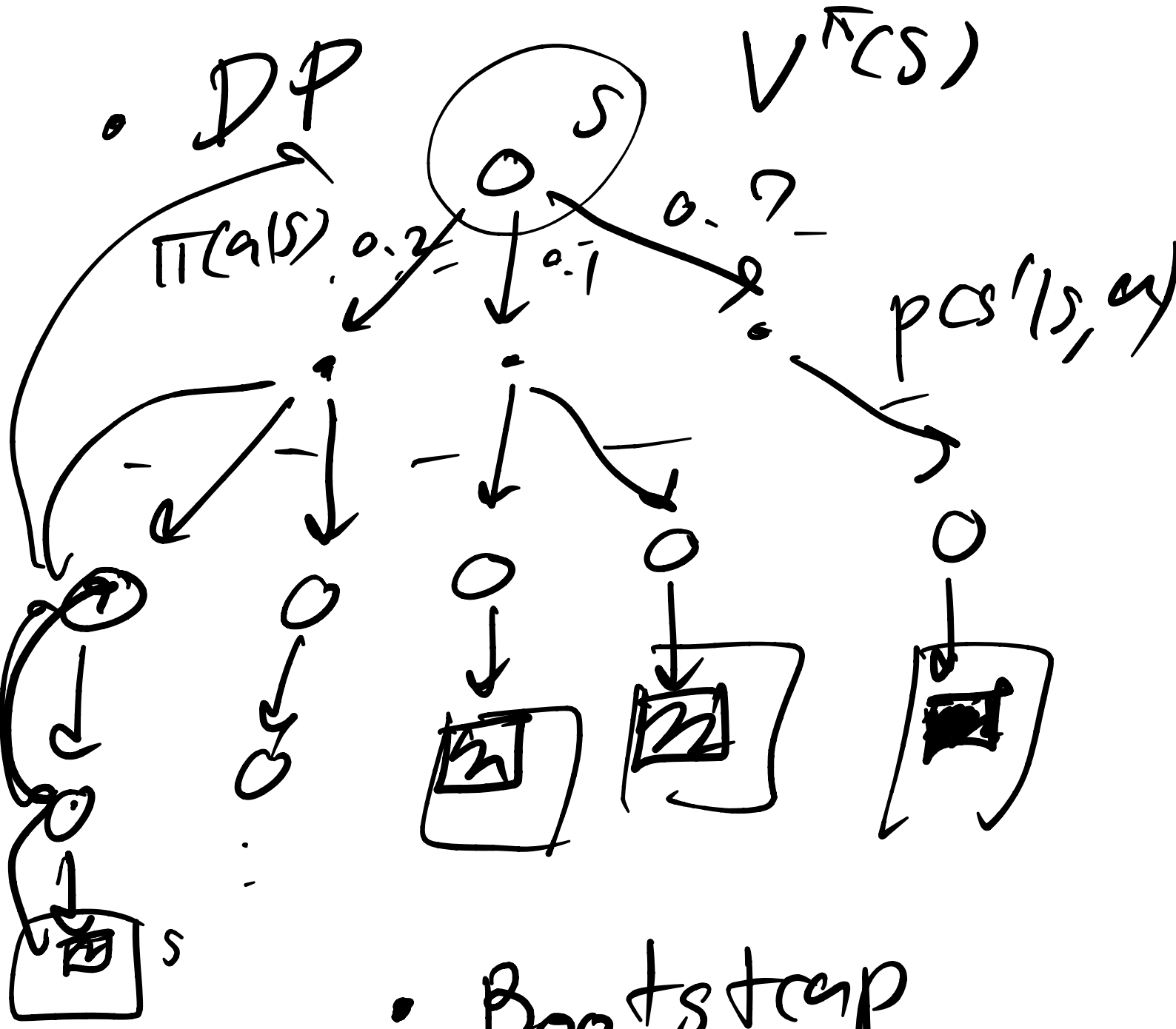
$s_4 =$ Terminate

$$g(0) = r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4$$

$$= 1 + \gamma(0) + \gamma^2(1) + \gamma^3(1)$$

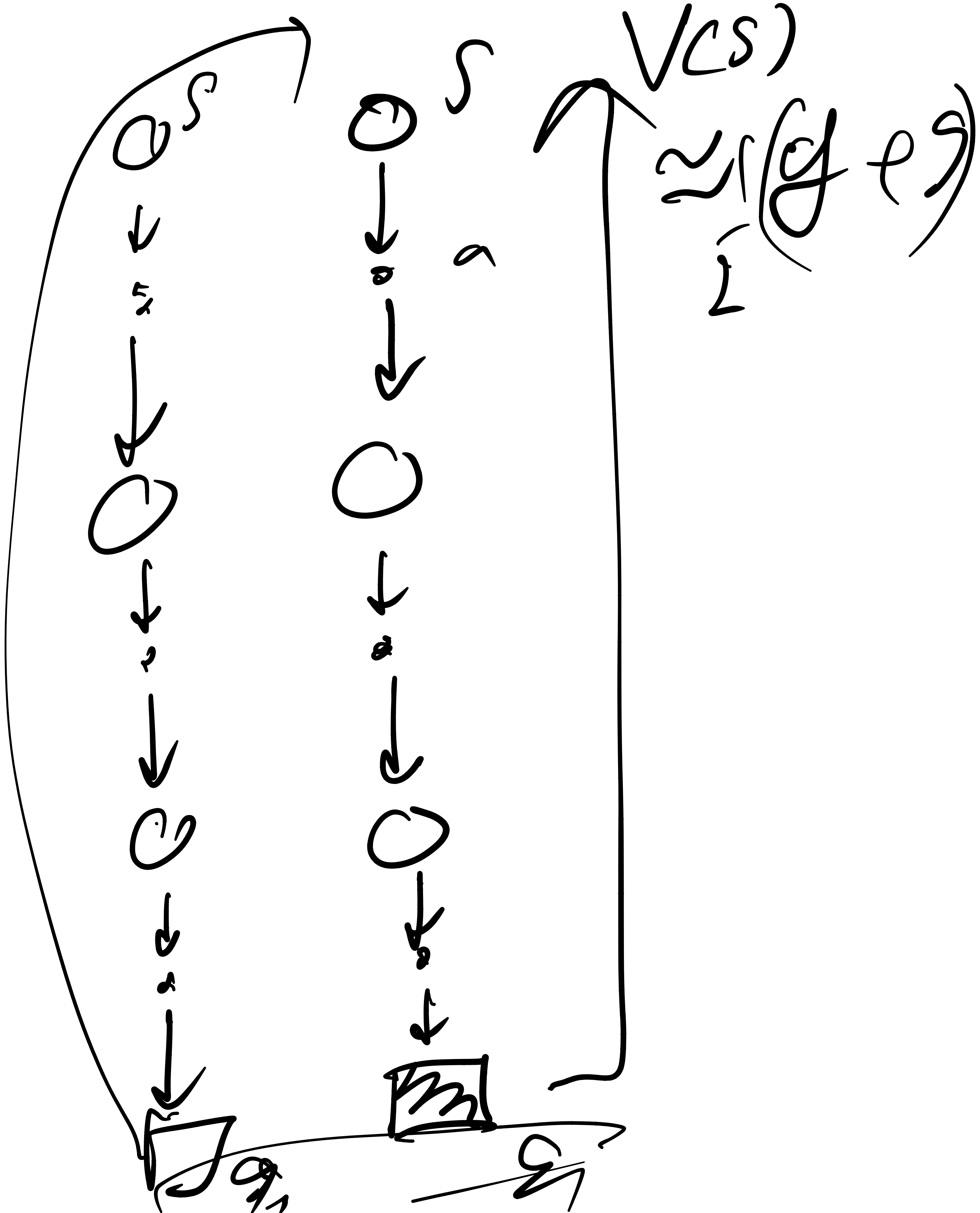
$$G_1 = [g(0)]$$

Backup diagram

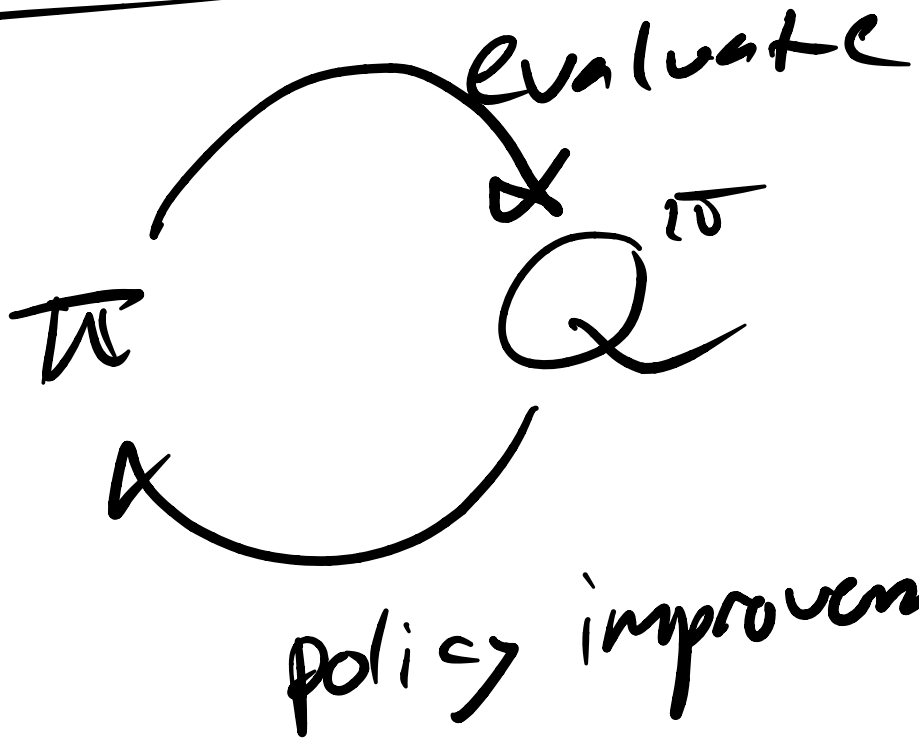


- Bootstrap
- Complete knowledge

• MC backup



MC - Control



1.) Initialize

$Q(s,a)$, $\pi(s)$

$G_{(s,a)} = []$

2-) Generate trajectory with π

- For each s, a
  - $g(s,a) \leftarrow \sum_i \gamma^i r$
  - Append $g(s,a) \rightarrow G_{s,a}$
  - Compute $Q(s,a) = Avg[G_{s,a}]$

- For each s in the traj
  $\pi(s) \leftarrow \underset{a}{\text{argmax}}\, Q(s,a)$

Exploration in MC

- We want to estimate

  $V^{\pi}(s)$ or $Q^{\pi}(s,a)$

  $\forall s$ or $\forall (s,a)$

- For a particular π, (s,a) may never be visited.

- One way to cheat

  "Exploration Start assumption"

- We could consider using $\pi(a|s)$

→ e.g. ε-greedy policy

$$a^* = \arg\max_a Q(s,a)$$