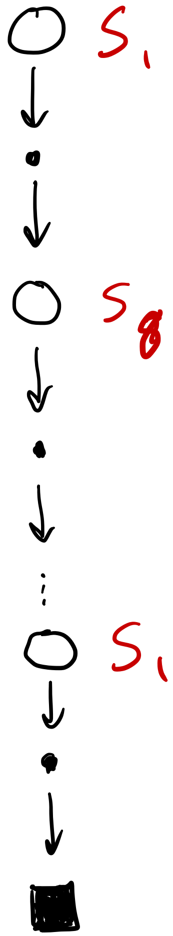
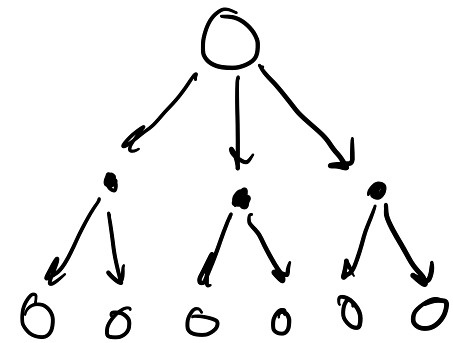




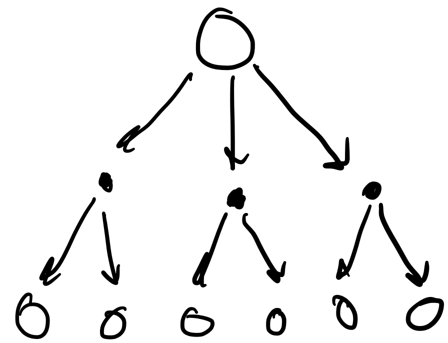
Bellman eqn. for V_π,
Dynamic programming

Monte Carlo
est. of V_π

Sample of return:

$$G_{s_t} = \sum_{i=t}^{T-1} r_i \implies V(s) = E\left[G_{s_t}\right]$$

Monte Carlo est. of V_π



Bellman eqn. for V_π, Dynamic programming

MC:

- only sampled transitions
- All the way to end of episode
- no bootstrapping

DP:

- All possible transitions
- only one step
- bootstrapping

Sample of return:

$$G_{s_t} = \sum_{i=t}^{T-1} r_i \quad \Rightarrow \quad V(s) = E\left[G_{s_t}\right]$$

First-visit MC prediction, for estimating $V \approx v_\pi$

Input: a policy π to be evaluated

Initialize:

$\quad V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$

$\quad Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):

$\quad$ Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

$\quad G \leftarrow 0$

$\quad$ Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$\quad\quad G \leftarrow \gamma G + R_{t+1}$

$\quad\quad$ Unless S_t appears in $S_0, S_1, \ldots, S_{t-1}$:

$\quad\quad\quad$ Append G to $Returns(S_t)$

$\quad\quad\quad V(S_t) \leftarrow \text{average}(Returns(S_t))$

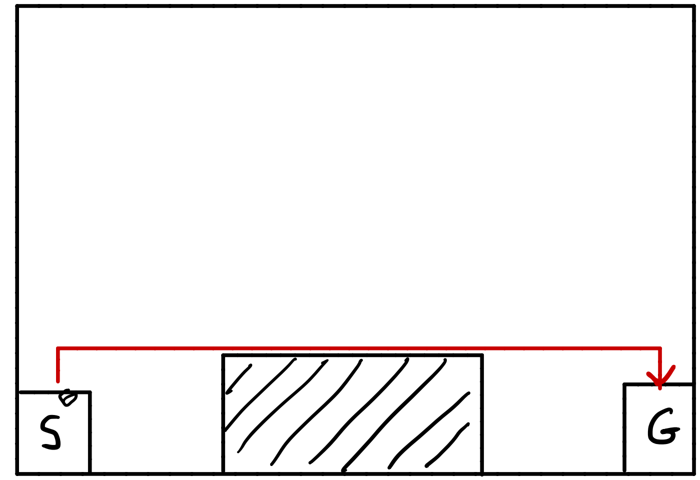
$G_i^{s,\pi}$: i^{th} return starting from state s, collected from policy π

On policy prediction :

$$V_\pi(s) = \frac{1}{N}\sum_{i=1}^{N} G_i^{s,\pi}$$

$$Q_\pi(s,a) = \frac{1}{N}\sum_{i=1}^{N} G_i^{s,a,\pi}$$

$G_i^{s,\pi}$: i^{th} return starting from state s, collected from policy π

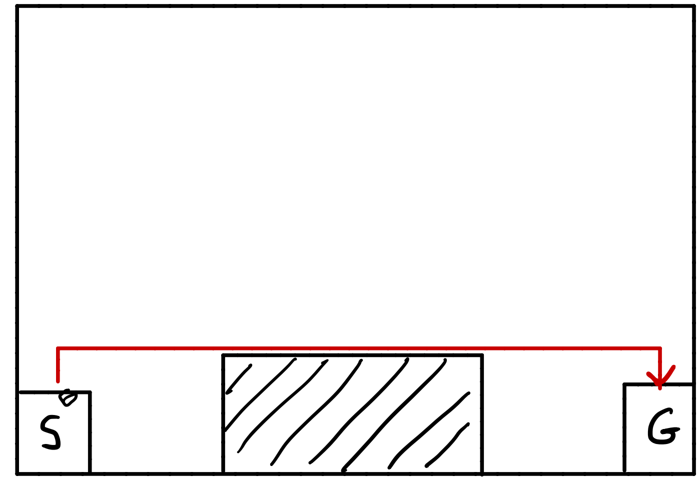


On policy :
prediction

$$V_\pi(s) = \frac{1}{N} \sum_{i=1}^{N} G_i^{s,\pi}$$

$$Q_\pi(s,a) = \frac{1}{N} \sum_{i=1}^{N} G_i^{s,a,\pi}$$

$G_i^{s,\pi}$: i^{th} return starting from state s, collected from policy π



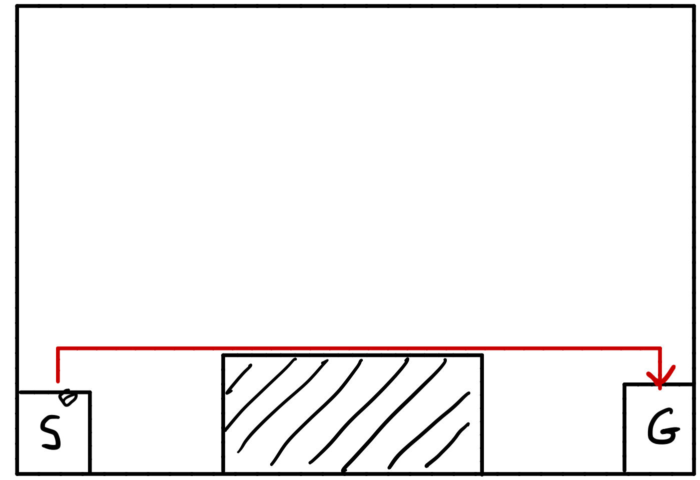
On policy : $V_\pi(s) = \frac{1}{N} \sum_{i=1}^{N} G_i^{s,\pi}$
prediction

$Q_\pi(s,a) = \frac{1}{N} \sum_{i=1}^{N} G_i^{s,a,\pi}$

Off policy : $V_{\pi'}(s) = \frac{1}{N} \sum_{i=1}^{N} G_i^{s,\pi} \cdot \rho_i$ where $\rho_i = \prod_{k=t_i}^{T_i - 1} \frac{\pi'(A_k | S_k)}{\pi(A_k | S_k)}$
prediction

$Q_{\pi'}(s,a) = \frac{1}{N} \sum_{i=1}^{N} G_i^{s,a,\pi} \cdot \rho_i$

$G_i^{s,\pi}$: i^{th} return starting from state s, collected from policy π



On policy prediction: $V_\pi(s) = \frac{1}{N}\sum_{i=1}^{N} G_i^{s,\pi}$

$Q_\pi(s,a) = \frac{1}{N}\sum_{i=1}^{N} G_i^{s,a,\pi}$

Off policy prediction: $V_{\pi'}(s) = \frac{1}{N}\sum_{i=1}^{N} G_i^{s,\pi} \cdot \rho_i$ where $\rho_i = \prod_{k=t_i}^{T_i - 1} \frac{\pi'(A_k|S_k)}{\pi(A_k|S_k)}$

$Q_{\pi'}(s,a) = \frac{1}{N}\sum_{i=1}^{N} G_i^{s,a,\pi} \cdot \rho_i$

Consider: On-policy control vs. off-policy control w/ ϵ-greedy exploration

What are $\underbrace{V_\star \text{ and } \pi_\star}_{\text{off-policy}}$ vs. $\underbrace{\tilde{V}_\star \text{ and } \tilde{\pi}_\star}_{\text{on-policy}}$?

Safe off policy evaluation:

Return probabilistic lower bound V_π^{lb} such that:

$V_\pi > V_\pi^{lb}$ with prob. $1-\delta$ Given: π, δ, data from π_b

without ever running policy π!

Safe off policy evaluation:

Return probabilistic lower bound V_π^{lb} such that:

$V_\pi > V_\pi^{lb}$ with prob. $1-\delta$ Given: π, δ, data from π_b

without ever running policy π!

Confidence bounds: Chernoff - Hoeffding inequality

with probability at least $1-\delta$:

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\log(1/\delta)}{2n}} \qquad \text{for } 0 \leq X_i \leq b$$

Safe off policy evaluation:

Return probabilistic lower bound V_π^{lb} such that:

$V_\pi > V_\pi^{lb}$ with prob. $1-\delta$ Given: $\pi, \delta,$ data from π_b

without ever running policy π!

Confidence bounds: Chernoff-Hoeffding inequality

with probability at least $1-\delta$:

$$\mu \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\log(1/\delta)}{2n}} \quad \text{for } 0 \leq X_i \leq b$$

$$\downarrow$$

$$V_\pi \geq \frac{1}{n}\sum_{i=1}^{n} G_i^{\pi_b} \cdot e_i^{\pi,\pi_b} - G_{max}\sqrt{\frac{\log(1/\delta)}{2n}} \quad \text{for } 0 \leq G_i \leq G_{max}$$

Given returns $G_1 \cdots G_n$ from policy π_b AND $\rho_i = \prod_{k=t_i}^{T_i - 1} \frac{\pi(A_k|S_k)}{\pi_b(A_k|S_k)}$ THEN: $V_\pi(s) =$

Given returns $G_1 \cdots G_n$ from policy π_b AND $\rho_i = \prod_{k=t_i}^{T_i - 1} \frac{\pi(A_k|S_k)}{\pi_b(A_k|S_k)}$ THEN: $V_\pi(s) =$

OIS :

$$\frac{1}{n} \sum_{i=1}^{n} G_i \rho_i$$

Given returns $G_1 \cdots G_n$ from policy π_b AND $\rho_i = \prod_{k=t_i}^{T_i - 1} \frac{\pi(A_k|S_k)}{\pi_b(A_k|S_k)}$ THEN: $V_\pi(s) =$

OIS:

$$\frac{1}{n} \sum_{i=1}^{n} G_i \rho_i$$

WIS:

$$\sum_{i=1}^{n} \frac{G_i \rho_i}{\sum_{j=1}^{n} \rho_j}$$

Given returns $G_1 \cdots G_n$ from policy π_b AND $\rho_i = \prod_{k=t_i}^{T_i - 1} \frac{\pi(A_k|S_k)}{\pi_b(A_k|S_k)}$ THEN: $V_\pi(s) =$

OIS:

$$\frac{1}{n} \sum_{i=1}^{n} G_i \rho_i$$

WIS:

$$\sum_{i=1}^{n} \frac{G_i \rho_i}{\sum_{j=1}^{n} \rho_j}$$

PDIS:

$$\frac{1}{n} \sum \tilde{G}_i \text{, where:}$$

$$\tilde{G}_i = \rho_{1:1} R_1 + \gamma \rho_{1:2} R_2 + \dots + \gamma^{n-1} \rho_{1:t} R_t$$

and

$$\rho_{a:b} = \prod_{k=a}^{b} \frac{\pi(A_k|S_k)}{\pi_b(A_k|S_k)}$$