① Monte Carlo

Recap:
- MDP [extensions: multi agent, imperfect information]
- Bellman equations
  - ↳ VI / PI

[Tabular assumption:
- X · S states is FINITE and TABULATED
- ∽ · A actions is FINITE and TABULATED
- X · $\Delta : S \times A \to \mathcal{D}ist(S \times \mathbb{R})$ is KNOWN

[Now! Statistical framework
$s, a \longmapsto$ Sample$(\Delta(s,a)) \longmapsto (s', r)$

$\sigma : S \to A$ (partial function)

Evaluation (σ) task
- ↳ $val_\sigma(s)$ for s all states σ is defined on

$$v_{\sigma}(s_0) = \mathop{\mathbb{E}}_{\sigma, s_0}\left[\underbrace{r_0 + \gamma r_1 + \gamma^2 r_2 + \dots}_{G \text{ total return}} \right]$$

- ↳ $q_\sigma(s_0, a)$ for s all states, $a \in A$ actions

Most natural idea: Monte Carlo

Sample a number N of trajectories from s_0 using σ

and $\widehat{val_\sigma(s_0)} = \frac{1}{N} \sum_{i=1}^{N} G_i$

ith trajectory:

$\pi_i = s_0^i, a_0^i, r_0^i, s_1^i, a_1^i, r_1^i, s_2^i, \dots$

$G_i = r_0^i + \gamma r_1^i + \gamma^2 r_2^i + \dots$

↳ total reward.

truncate to large K

Works for s_0. What about

$val_\sigma(s)$ for all $s \in S$ that σ "knows"

$$val_\sigma(s) = \sum_{s', r} \Delta(s, \sigma(s))(s', r)\,(r + \gamma\, val_\sigma(s'))$$

↳ dynamic programming is not available in our statistical setting!

idea: one sample → many trajectories!

$s_0, a_0, r_0, s_1, a_1, r_1, s_2, \dots$

from s_0

from s_1

from s_2

How to use them? $s \in S$

$\widehat{val_\sigma(s)}$

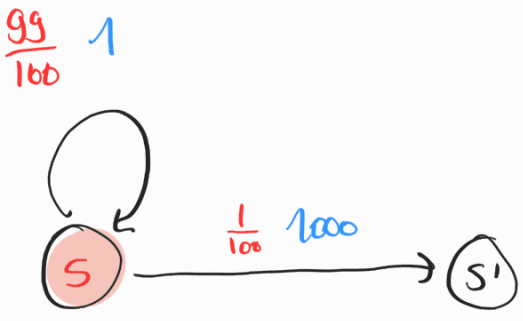
$\widehat{val}_\sigma(s) =$ s_0 ... s ... s ... s ...

(1) average over all trajectories which start in the first occurrence of s

(2) average over all trajectories which start in any occurrence of s

(3) average over all trajectories which start in the last occurrence of s

$\gamma = 1$



probability

Reward

$$\widehat{val}_\sigma(s) = \begin{cases} \text{last visit} & 1000 \quad \times \\ \text{all visits} & 1099 \quad \checkmark \\ \text{first visit} & 1099 \quad \checkmark \end{cases}$$

Samples:

- $s\ s \cdots s\ s'$ (1000)
- $s\ s \cdots s\ s'$ (1000)
- $s \cdots s\ s'$
- ⋮

actual value

$$val_\sigma(s) = \frac{99}{100} \cdot (1 + val_\sigma(s)) + \underbrace{\frac{1}{100} \cdot 1000}_{10}$$

$$\| \quad v$$

$$100\, v = 99\,(1 + v) + 1000$$

$$v = 1099$$

last visit

all trajectories have total reward 1000

first visit

"typical sample":

$S, 1, S, 1, S, 1, \ldots, S, 1000, S'$

$\hookleftarrow \sim 99 \hookrightarrow$

value: $\underbrace{1+1+\ldots+1}_{99} + 1000 \simeq 1099$

all visits

calculations are hard to do by hand, but it does converge to 1099...

CONCLUSION:

BOTH all visits and first visits converge to the right value

last visit does NOT.