Last Time

MDP $(S, A, T, R, \gamma, [P_0])$

Policy: $\pi : S \to A$ $\quad a = \pi(s)$

MC Policy Eval.

MDP: Balance immediate and future rewards

Today

- When is enough MC sims? (SEM)
- Policy Search (Cross Entropy)
- Value Function
- Policy Iter. } Optimal Policies!
- Value Iter. } "Dynamic Programming"

MC Evaluation

$$\hat{u} = \sum_{t=0}^{T-1} \gamma^t r_t$$

$$\bar{u}_m = \frac{1}{m} \sum_{k=1}^{m} \hat{u}_k$$

$$\sum_{t=0}^{\infty} \gamma^t r_t = \sum_{t=0}^{T-1} \gamma^t r_t + \sum_{t=T}^{\infty} \gamma^t r_t$$

$$= \sum_{t=0}^{T-1} \gamma^t r_t + \gamma^T \sum_{t=T}^{\infty} \gamma^{t-T} r_t$$

$$\leq \gamma^T \frac{\bar{r}}{1-\gamma}$$

Central Limit Theorem

as $m \to \infty$ $\quad \bar{u}_m \xrightarrow{d} N\left(\mu, \sigma_{\hat{u}}/\sqrt{m}\right)$

(μ: true mean return; $\sigma_{\hat{u}}$: std. dev of $\hat{u}$)

$$\boxed{SEM(\bar{u}_m) \equiv \frac{std(\hat{u})}{\sqrt{m}}}$$

Policy Search (Ch 10)

① Parameterize policy $\quad \pi(s) = f(s;\theta)$

② Optimize θ w/ favorite $\quad \pi(s) = -k_\theta s$

0th order optimization alg.

fmincon

$|S| < \infty \quad |A| < \infty$

Value Function

$$\pi(s) = \underset{a \in A}{\operatorname{argmax}}\ R(s,a)$$

Not good

"Myopic"

"Greedy"

$$V^\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_t = \pi(s_t)\right]$$

Hard

$$V^\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_t = \pi(s_t)\right]$$

$$= R(s, \pi(s)) + E\left[\sum_{t=1}^{\infty} \gamma^t r_t \,\middle|\, s_1 \sim T(s,a), a_t = \pi(s_t)\right]$$

$$= \text{"} \quad + \gamma E\left[\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s', a_t = \pi(s_t)\right]$$

$$\rightarrow \quad V^\pi(s) = R(s, \pi(s)) + \gamma \underset{s' \sim T(s,a)}{E}\left[V^\pi(s')\right]$$

$$= \text{"} \quad + \gamma \sum_{s' \in S} T(s'|s,a) V^\pi(s')$$

$T^\pi_{i,j} = T(j|i, \pi(i))$

$R^\pi_i = R(i, \pi(i))$

$$\rightarrow \quad V^\pi = R^\pi + \gamma T^\pi V^\pi$$

$$V^\pi - \gamma T^\pi V^\pi = R^\pi$$

$$\boxed{V^\pi = (I - \gamma T^\pi)^{-1} R^\pi}$$

matrix policy evaluation

Policy Iteration

initialize π, π'

while $\pi \neq \pi'$

$\pi = \pi'$

$\rightarrow \quad V^\pi \leftarrow (I - \gamma T^\pi)^{-1} R^\pi$ ← policy eval

$\pi'(s) \leftarrow \underset{a \in A}{\text{argmax}} \left(R(s,a) + \gamma \sum_{s'} T(s'|s,a) V^\pi(s')\right)$ ∀ $s \in S$

↑ policy update

π is optimal!

---

know V^*

(1) $\pi^*(s) = \underset{a \in A}{\text{argmax}} R(s,a) + \gamma \underset{s' \sim T(s,a)}{E}\left[V^*(s')\right]$

if know V^*, we know π^*

Bellman's Equation

(2) $$V^*(s) = \max_{a \in A} \left[R(s,a) + \gamma \, E_{s' \sim T(s,a)} \left[V^*(s') \right] \right]$$

Value Iteration

Why does VI converge?

initialize V, V'

while $V \neq V'$

$V = V'$

$V'(s) \leftarrow \max_a R(s,a) + \gamma \, E_{s' \sim T(s,a)} \left[V(s') \right]$

$\forall s \in S$

when terminates, $V = V^*$.

Fact: There can be many optimal policies, but only one optimal value function!

"Dynamic Programming"

↑ Because it sounds cool

Break down a large opt. into many smaller ones.