$$= \max_{s} \left| \max_{a} \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot V(s')\right) - \max_{a} \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot V'(s')\right) \right|$$

$$= \max_{s} \left| \max_{a} \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot V(s')\right) - \sum_{s'} \text{---------} + \gamma \cdot V'(s') \right|$$

claim $\Rightarrow \left| \max_{x} f(x) - \cancel{g(x)} \max_{x} g(x) \right| \leq \max_{x} \left| f(x) - g(x) \right|$

$$\forall x \quad f(x) - g(x) \leq |f(x) - g(x)|$$

$$f(x) \leq |f(x) - g(x)| + g(x)$$

$$\max_{x} f(x) \leq \max_{x} |f(x) - g(x)| + g(x)$$

$$\leq \max_{x} |f(x) - g(x)| + \max_{x} g(x)$$

$$\therefore \max_{x} f(x) - \max_{x} g(x) \leq \max_{x} |f(x) - g(x)|$$

Lecture - 11

- $T_v(s) = \max_a \sum_{s'} P(s,a,s')\left(R(s,a,s') + \gamma \cdot v(s')\right)$

- Contraction mapping

  $f: X \to X$ is a contraction mapping if there exists a $\lambda \in [0,1]$ such that $\forall x, y$

$$d(f(x), f(y)) \leq \lambda\, d(x,y)$$

- Show

$$\| T_v - T_{v'} \| \leq \gamma \cdot \| v - v' \|$$

$$\| T_v - T_{v'} \| = \max_s | T_v(s) - T_{v'}(s) |$$

$$= \max_s \left| \max_a \sum_{s'} P(s a s')\left(R(s a s') + \gamma \cdot v(s')\right) - \max_a \sum_{s'} P(s,a,s')\left(R(s a s') + \gamma \cdot v'(s')\right) \right|$$

$$\leq \max_s \left| \max_a \left| \sum_{s'} P(s a s')\left(\gamma \cdot v(s') - \gamma \cdot v'(s')\right) \right| \right|$$

$$\leq \gamma \cdot \max_s \max_a \left| \max_{s'} | v(s') - v'(s') | \right|$$

$$\leq \gamma \cdot \max_{s'} \left| v(s') - v'(s') \right|$$

$$\leq \gamma \cdot \| v - v' \|$$

Overview till today —

- MDP — environment
- CEM — first agent
- Value function, Bellman eqⁿ
- Dynamic programming
  - Policy evaluatⁿ
  - Policy improvement
  - — " — iteratⁿ
  - Value iteratⁿ
    - Bellman operator

<u>Monte-Carlo method</u> —

- Named after a Casino : Monte Carlo Casino

Idea — Use randomness to solve problems that are deterministic in principle.

Policy evaluation using Monte-Carlo —

- History $H = (S_0, A_0, R_0, \ldots)$ generated by policy $\underline{\pi}$

Estimate of $v^{\pi}(s)$

$$= \sum_{t=0}^{\infty} \gamma^t R_t$$

First visit MC

- Generate many episodes
- For each state, average discounted returns after it was visited the first time in each episode

Pseudo code —

- Initialize : $\pi \leftarrow$ policy to be evaluated

  $V \leftarrow$ an arbitrary value-function estimate

  $R \leftarrow$ an empty list for each state s

- Repeat forever

- ① Generate an episode using π

② for each state s appearing in the episode

$G_t \leftarrow$ Return following the first occurance of s

if t is the first where $s_t = s$, then

$$G = G_t \triangleq \sum_{k=0}^{\infty} \gamma^k \cdot R_{t+k}$$

Append G to Returns(s)

$V(s) \longleftarrow$ average (Returns(s))