Lecture 16 - March 14, 2023

- Temporal Difference Learning
  - TD(0)
  - SARSA
  - Q-Learning
  - On-Policy Vs. off-policy

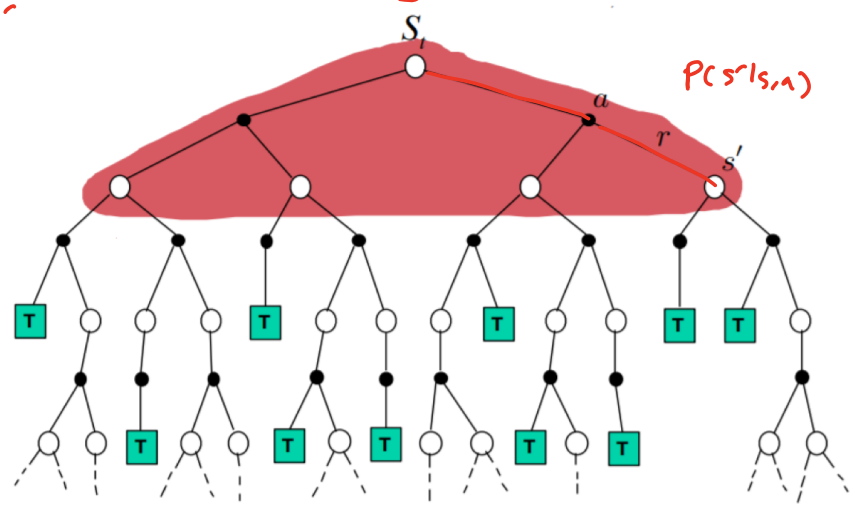
Project 3 is posted → Due April 14

HW3 → Due March 17

TA's office hour: Wendsdays, 2pm-3pm (in-person)
Fridays, 2pm-3pm (virtual)

DP: $V_{k+1}(s) = \max_{a \in A} \sum_{s'} P(s'|s,a)[R + \gamma V_k(s')]$

Model of system

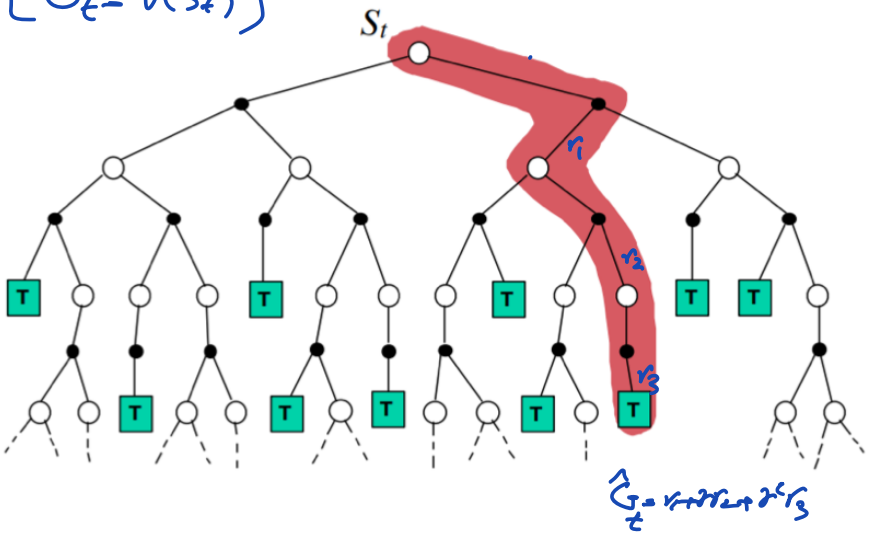


MC: $V(S_t) = V(S_t) + \alpha[\hat{G}_t - V(S_t)]$

Model X

Delay ✓

Interactive X

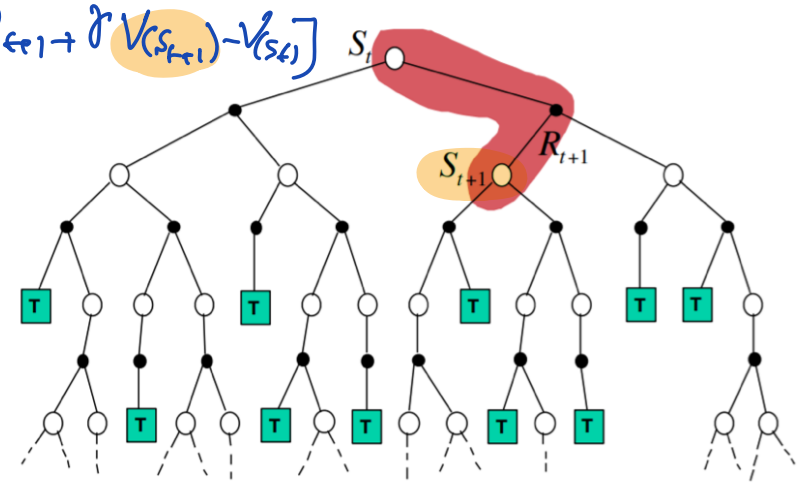


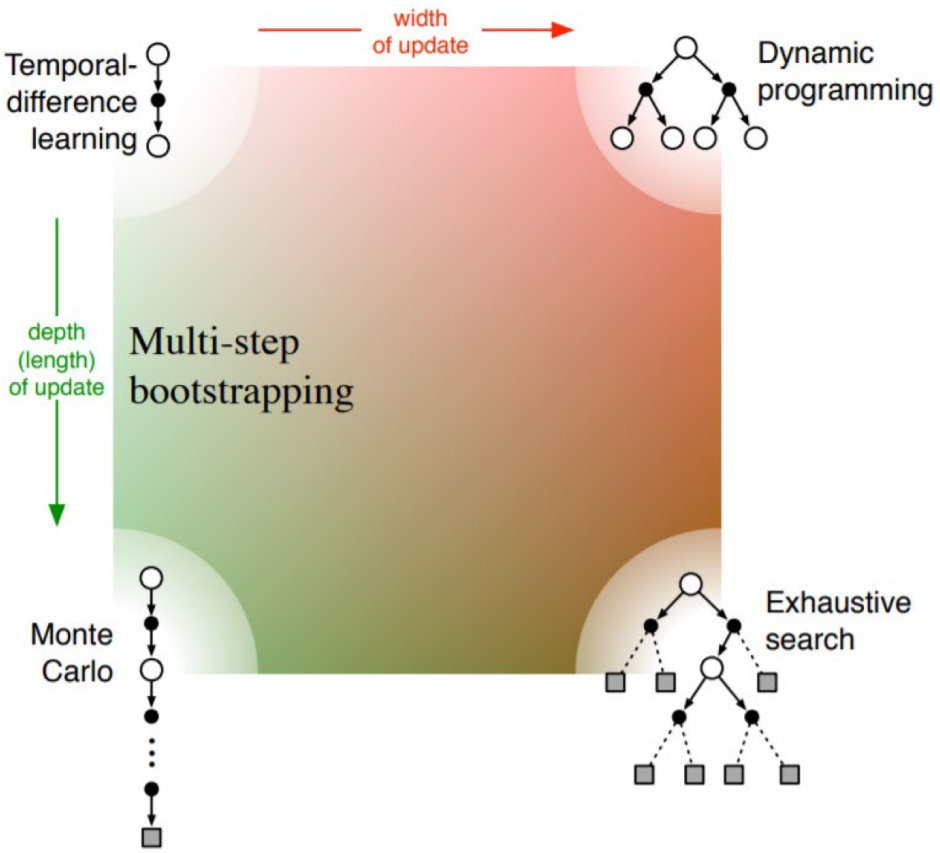
TD: $V(S_t) = V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$

Model X

Delay one step

Interactive ✓



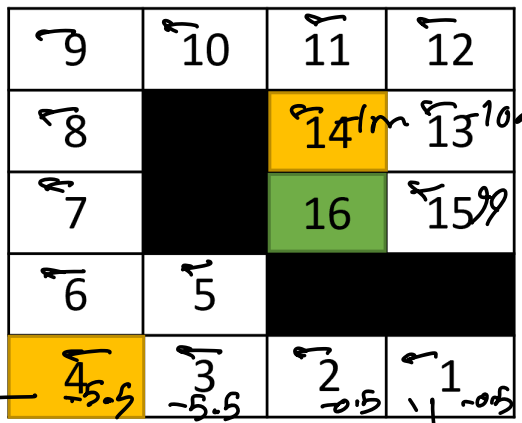

width
of update
Temporal-
difference
learning
Dynamic
programming
depth
(length)
of update
Multi-step
bootstrapping
Monte
Carlo
Exhaustive
search

TD(0)

$\pi \longrightarrow V^{\pi}$

$$V(S_t) = V(S_t) + \alpha \left[R_{t+1}^{\pi(S_t)} + \gamma V(S_{t+1}) - V(S_t) \right]$$

$R_{t+1}^{\pi(S_t)} \rightarrow R(S_t, \pi(S_t), S_{t+1})$



V^{π}

$V(1) = V(1) + \alpha [-1 + \gamma V(2) - V(1)] = -0.5$

$V(4) = V(4) + \alpha [-11 + \gamma V(4) - V(4)] = -9.8$

Episode

1 → 2 → 3 → 4 → 1 → 4 → 4 — — —

Random state

8

1000

$\pi \longrightarrow V^{\pi}$ TD(0)

$$\pi'(s) = \underset{a}{\operatorname{argmax}} \sum_{s'} P(s'|s,a)\left[R + \gamma V_{\pi}(s')\right]$$

Policy Improvement

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

TD (0) $\Rightarrow$ Most application for Policy optimization

$\pi \Rightarrow$ $V_{\pi}(s_t) = V_{\pi}(s_t) + \alpha\left[R_{t+1} + \gamma V_{\pi}(s_{t+1}) - V_{\pi}(s_t)\right]$

How to improve the Policy

Resolving Issues of TD(0)

- Learn Q
-

$$V_\pi(S_t) = V_\pi(S_t) + \alpha\left[R_{t+1} + \gamma V_\pi(S_{t+1}) - V_\pi(S_t)\right] \Rightarrow TD(0)$$

$$Q_\pi(S_t, \pi(S_t)) = Q_\pi(S_t, \pi(S_t)) + \alpha\left[R_{t+1} + \gamma Q_\pi(S_{t+1}, \pi(S_{t+1})) - Q_\pi(S_t, \pi(S_t))\right]$$

stochastic Policy

$$Q_\pi(S_t, \pi(S_t)) = Q_\pi(S_t, \pi(S_t)) + \alpha\left[R_{t+1} + \gamma Q_\pi(S_{t+1}, \pi(S_{t+1})) - Q_\pi(S_t, \pi(S_t))\right]$$

$$\pi_{\varepsilon\text{-greedy}}(S_t) = \begin{cases} \arg\max_a Q(S_t, a) & 1-\varepsilon \\ \text{Random} & \varepsilon \end{cases}$$

π

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

$Q_\pi(S, a)$

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

$Q(S,a)=0$

9	10	11	12
8		14	13
7		16	15
6	5		
4	3	2	1

Wall Bump Goal

0.25 0.25 0.25

$\frac{0.9}{3}$ $\frac{0.9}{3}$ $\frac{0.9}{3}$

argmax Q L D

$\frac{0.1}{4}$ $\frac{0.1}{4}$ $\frac{0.1}{4}$ $\frac{0.1}{4}$

-0.5

0 0 0

argmax $Q(S=1,a)$: U, L, R, D $1-\varepsilon$

Random ε

$\pi(S=1) = 0.25$ U L R D

stochastic Policy

$$Q_\pi(S_t, \pi(S_t)) = Q_\pi(S_t, \pi(S_t)) + \alpha\left[R_{t+1} + \gamma Q_\pi(S_{t+1}, \pi(S_{t+1})) - Q_\pi(S_t, \pi(S_t))\right]$$

$$\pi^{\text{Egreedy}}(S_t) = \begin{cases} \text{argmax}_a Q(S_t, a) & 1-\varepsilon \\ \text{Random} & \varepsilon \end{cases}$$

$$1 \xrightarrow{\pi^{Eg}(1) \sim V} 1 \xrightarrow{\pi^{Eg}(1) \sim L}$$

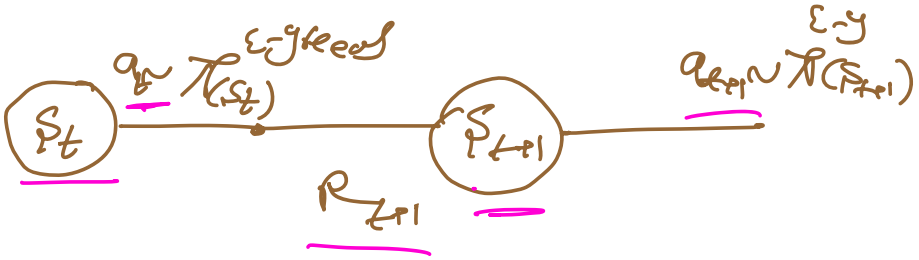
$$Q(1, V) = \underset{0}{Q(1, U)} + \alpha\left[-1 + \gamma \underset{0}{Q(S_{t+1}, L)} - \underset{0}{Q(1, U)}\right]$$

α: 0.5

$= -0.5$

SARSA

Initialize $Q(s,a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Choose A from S using policy derived from Q (e.g., ε-greedy)
  Repeat (for each step of episode):
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S',A') - Q(S,A)]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal



$$Q(S_t, a_t) = Q(S_t, a_t) + \alpha\left[R_{t+1} + \gamma Q(S_{t+1}, a_{t+1}) - Q(S_t, a_t)\right]$$

State – Action – Reward – State – Action (SARSA)

$$\pi^{\varepsilon}(s) = \begin{cases} \underset{a \in A}{\operatorname{argmax}}\, Q(s,a) & 1-\varepsilon \\ \text{Random} & \varepsilon \end{cases}$$

Final SARSA: $\pi^*(s) = \underset{a \in A}{\operatorname{argmax}}\, Q(s,a)$

Example

A	B

Intuitively
$\pi^*(A) = a^2$
$\pi^*(B) = a^1$

$$M(a^1) = \begin{array}{c} \\ A \\ B \end{array}\begin{bmatrix} A & B \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$M(a^2) = \begin{array}{c} \\ A \\ B \end{array}\begin{bmatrix} A & B \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Reward → 5 B
−1 a^2

SARSA

Initial Q

$$\begin{bmatrix} Q(A, a^1) \\ Q(B, a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} Q(A, a^2) \\ Q(B, a^2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

episode 1

Initial state = A (Random)

$$\pi^{\varepsilon\text{-}g}(A) = \begin{cases} \underset{a \in A}{\operatorname{argmax}}\, Q(A, a) & \overset{0.5\ 0.5}{a^1, a^2} \\ \text{Random} & \sim a^2 \end{cases}$$

$A \xrightarrow[R=4]{a^2 \sim \pi^{\varepsilon\text{-}g}(A)} B \xrightarrow{a^2 \sim \pi^{\varepsilon\text{-}g}(B)}$

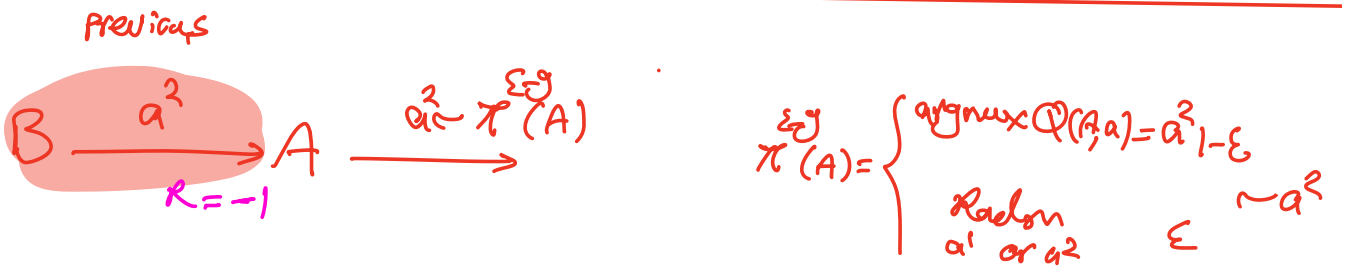
$$\pi^{\varepsilon\text{-}g}(B) = \begin{cases} 0.5 & a^1 \sim a^2 \\ 0.5 & a^2 \end{cases}$$

$$Q(A, a^2) = \underset{0}{Q(A, a^2)} + \underset{0.5}{\alpha}\left[\underset{4}{R} + \underset{0.9}{\gamma}\underset{0}{Q(B, a^2)} - \underset{0}{Q(A, a^2)}\right] = 2$$

$$\begin{bmatrix} Q(A, a^1) \\ Q(B, a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} Q(A, a^2) \\ Q(B, a^2) \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$$

Closest Deterministic Policy to π_{greedy}

$\pi(A) = a^2$ $\quad$ $\pi(B) = a^1$ or a^2



$$Q(B, a^2) = \underbrace{Q(B, a^2)}_{0} + \alpha \left[\underbrace{R}_{-1} + \underbrace{\gamma}_{0.9} \underbrace{Q(A, a^2)}_{2} - \underbrace{Q(B, a^2)}_{0} \right] = 0.4$$

$$\begin{bmatrix} Q(A, a^1) \\ Q(B, a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} Q(A, a^2) \\ Q(B, a^2) \end{bmatrix} = \begin{bmatrix} 2 \\ 0.4 \end{bmatrix}$$

Closest Deterministic Policy to π_{greedy}

$\pi(A) = a^2$ $\quad$ $\pi(B) = a^2$

$A \xrightarrow[R=4]{a^2} B \xrightarrow{a' \sim \pi^{\varepsilon\text{-greedy}}(B)}$ ← $\begin{cases} 0.05 & a^1 \\ 0.95 & a^2 \end{cases} \sim a'$

$$Q(A,a^2) = \underbrace{Q(A,a^2)}_{2} + \alpha\left[\underbrace{R}_{4} + \gamma \underbrace{Q(B,a')}_{0} - \underbrace{Q(A,a^2)}_{2}\right] = 3$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 3 \\ 0.4 \end{bmatrix}$$

Closest Deterministic Policy to ε-greedy

$\pi(A) = a^2$ $\quad$ $\pi(B) = a^2$

$B \xrightarrow[R=5]{a'} B \xrightarrow{a^2 \sim \pi^{\varepsilon\text{-greedy}}(B)}$ $\quad \pi^{\varepsilon\text{-greedy}}(B) = \begin{cases} a^1 & 0.05 \\ a^2 & 0.95 \end{cases} \sim a^2$

$$Q(B,a') = \underbrace{Q(B,a')}_{0} + \underbrace{\alpha}_{0.5}\left[\underbrace{R}_{5} + \underbrace{\gamma}_{0.9} \underbrace{Q(B,a^2)}_{0.4} - \underbrace{Q(B,a')}_{0}\right] = 2.68$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2.68 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 3 \\ 0.4 \end{bmatrix}$$

Closest Deterministic Policy to $\pi^{\epsilon\text{-greedy}}$

$\pi(A) = a^2 \qquad \pi(B) = a^1$

$B \xrightarrow[R=-1]{a^2} A \xrightarrow{a^2 \sim \pi^{\epsilon\text{-g}}(A)} \begin{cases} 0.95 & a^2 \\ 0.05 & a^1 \end{cases}$

$$Q(B,a^2) = \underbrace{Q(B,a^2)}_{0.4} + \underbrace{\alpha}_{0.5}\left[\underbrace{R}_{-1} + \underbrace{\gamma}_{0.9}\underbrace{Q(A,a^2)}_{3} - \underbrace{Q(B,a^2)}_{0.4}\right] = 1.05$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2.68 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 3 \\ 1.05 \end{bmatrix}$$

Closest Deterministic Policy to $\pi^{\epsilon\text{-greedy}}$

$\pi(A) = a^2 \qquad \pi(B) = a^1$

$A \xrightarrow[R=4]{a^2} B \xrightarrow{a' \sim \pi^{\varepsilon\text{-}g}(B)} \begin{cases} 0.95 & a^1 \\ 0.05 & a^2 \end{cases}$

$$Q(A, a^2) = \underbrace{Q(A,a^2)}_{3} + \alpha \Big[\underbrace{R}_{4} + \gamma \underbrace{Q(B,a^1)}_{2.68} - \underbrace{Q(A,a^2)}_{3} \Big]$$

$\alpha = 0.5, \quad \gamma = 0.9$

$$= 4.7$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 2.68 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 4.7 \\ 1.05 \end{bmatrix}$$

Closest Deterministic Policy to $\pi^{\varepsilon\text{-greedy}}$

$\pi(A) = a^2 \qquad \pi(B) = a^1$

No learning

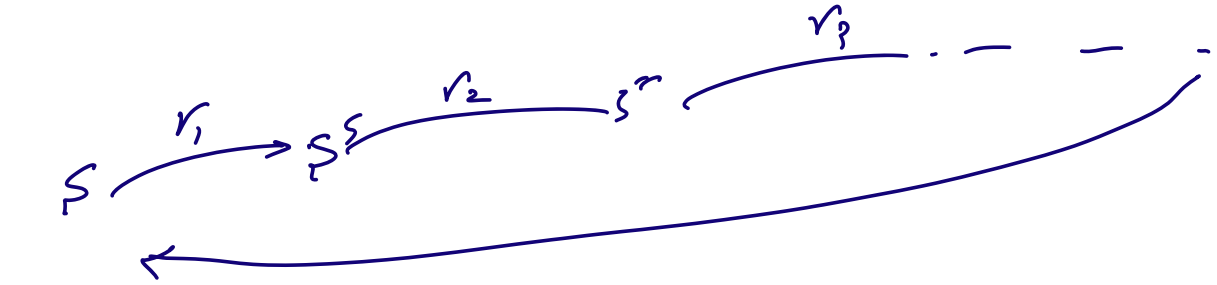
Execution → greedy form of ε-greedy

SARSA

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

$\pi^{SARSA}(A) = a^2$

$\pi^{SARSA}(B) = a^1$



$Q(s, a) = \sum \gamma r_i$