

Why  $\frac{1}{1-\gamma} \approx \text{"horizon"}$ ?

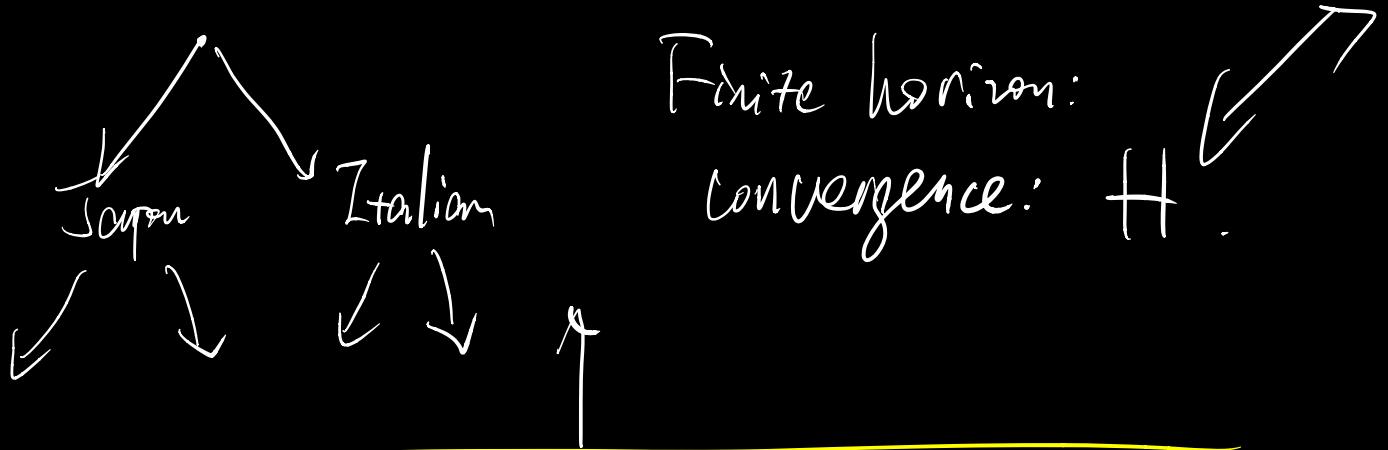
$$\sum_{t=1}^{\infty} \gamma^{t-1} r_t \in [0, \frac{R_{\max}}{1-\gamma}]$$

$$\sum_{t=1}^H r_t \in [0, R_{\max} \cdot H]$$

Convergence speed Value iteration.

$$\|f_k - Q^*\|_\infty \leq \gamma^k \|f_0 - Q^*\|_\infty \quad \text{"effective horizon".}$$

$$\|f_k - Q^*\|_\infty \leq \varepsilon. \Rightarrow k \approx \underbrace{\log \frac{1}{\varepsilon}}_{\text{const.}} = O\left(\frac{1}{1-\gamma}\right)$$



What if no absorbing (terminal) states in discounted MDP? How to do MC policy eval?

Ans: Stop (truncate) traj' at  $H \approx O(\frac{1}{1-\gamma})$   
incur small bias  $\frac{\gamma^H}{1-\gamma} R_{\max}$ .

from each  $(s, a)$ , we have sampled  $\{(r_i, \underline{s}_i^1)\}_{i=1}^n$

$$\Rightarrow \hat{R}(s, a) = \frac{1}{n} \sum_{i=1}^n r_i \quad \text{model estimation}$$

$$\hat{P}(s' | s, a) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(s'_i = s')$$

$$(\mathcal{T}f)(s, a) = \hat{R}(s, a) + \gamma \mathbb{E}_{\substack{s' \sim P(\cdot | s, a)}} \left[ \max_{a'} f(s', a') \right].$$

$$= \frac{1}{n} \sum_{i=1}^n r_i + \gamma \sum_{s' \in S} \hat{P}(s' | s, a) \left( \max_{a'} f(s', a') \right).$$

$$= \frac{1}{n} \sum_{i=1}^n r_i + \gamma \sum_{s' \in S} \left( \frac{1}{n} \sum_{i=1}^n \mathbb{I}(s'_i = s') \right) \max_{a'} f(s', a')$$

$$= \frac{1}{n} \sum_{i=1}^n r_i + \frac{\gamma}{n} \sum_{s' \in S} \mathbb{I}(s'_i = s') \max_{a'} f(s', a')$$

$$= \frac{1}{n} \sum_{i=1}^n r_i + \frac{\gamma}{n} \sum_{i=1}^n \max_{a'} f(s'_i, a') \quad \text{verif.}$$

$$= \frac{1}{n} \sum_{i=1}^n \left( r_i + \gamma \max_{a'} f(s'_i, a') \right). \approx (\mathcal{T}f)(s, a)$$

empirical Bellman update:

$$\mathbb{E} \left[ \underbrace{r_i + \gamma \max_{a'} f(s'_i, a')}_{\substack{r_i \sim R(\cdot | s, a) \\ s'_i \sim P(\cdot | s, a)}} \right] = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[ \max_{a'} f(s', a') \right] = (\mathcal{T}f)(s, a)$$