We should never discount when approximating policies!

γ is ok it there is a start state/distribution

The average-reward setting

- Maximize the reward rate (reward per step):

$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_\pi[R_t] = \sum_s d_\pi(s) \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) r$$

where $d_\pi(s) \doteq \lim_{t\to\infty} \mathbb{P}\{S_t = s\}$

- Learn to approximate $r(\pi)$ and new “differential” values, in which all rewards are compared to the reward rate:

$$\tilde{v}_\pi(s) = \sum_{k=1}^{\infty} \mathbb{E}_\pi[R_{t+k} - r(\pi) \mid S_t = s]$$

$$\tilde{q}_\pi(s,a) = \sum_{k=1}^{\infty} \mathbb{E}_\pi[R_{t+k} - r(\pi) \mid S_t = s, A_t = a]$$

Average-reward Q-learning (R-learning)

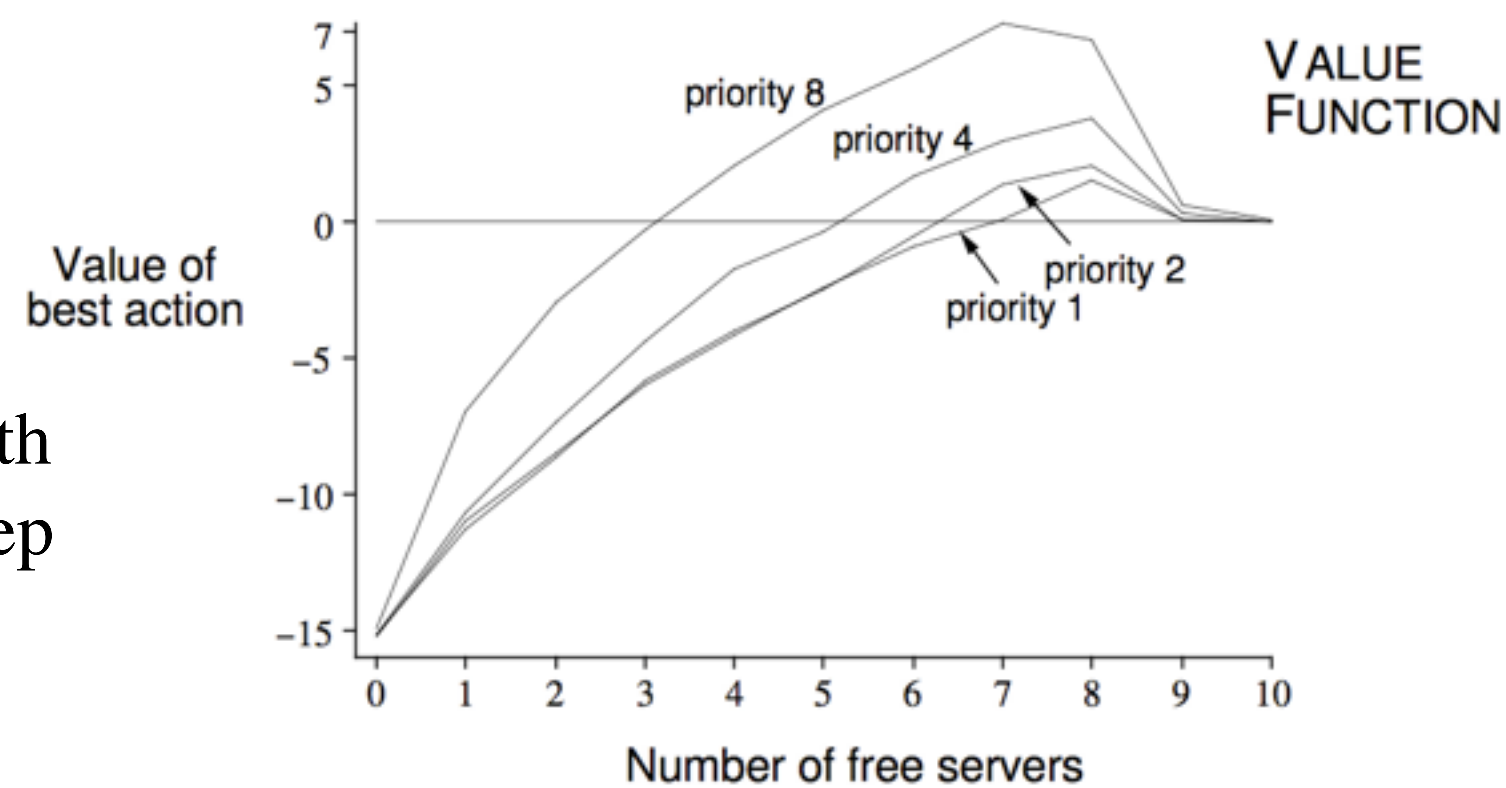
Initialize $\bar{R}$ and $Q(s, a)$, for all s, a, arbitrarily
Repeat forever:
    $S \leftarrow$ current state
    Choose action A in S using behavior policy (e.g., ϵ-greedy)
    Take action A, observe R, S'
    $\delta \leftarrow R - \bar{R} + \max_a Q(S', a) - Q(S, A)$
    $Q(S, A) \leftarrow Q(S, A) + \alpha\delta$
    If $Q(S, A) = \max_a Q(S, a)$, then:
        $\bar{R} \leftarrow \bar{R} + \beta\delta$

Access-Control Queuing Task

- n servers
- Customers have four different priorities, which pay reward of 1, 2, 4, or 8, if served
- At each time step, customer at head of queue is accepted (assigned to a server) or removed from the queue
- Proportion of randomly distributed high priority customers in queue is h
- Busy server becomes free with probability p on each time step
- Statistics of arrivals and departures are unknown

Apply R-learning

n=10, h=.5, p=.06



On-policy average-reward
with traces and linear FA

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \delta_t \mathbf{e}_t$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_{t+1} - \boldsymbol{\theta}_t^\top \boldsymbol{\phi}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta \delta_t$$