MARKOV DECISION PROCESSES

· Planning in uncertain domain.

MDP: It is a discrete time state transition system. It has 4 components

1) S : states.

Ex: For a robot the state could be room, or the (x, y) position.

- play the role of outcomes in the game with environment viewpoint.

2) A : actions

Usually chosen from a finite set.

3) $\mathbb{P}r(s_{t+1} \mid s_t, a_t)$: transition probabilities

. describe the dynamics.

$p(s'|s,a)$: the probability of going to state s' from state s if action a is played.

$$a_1 \quad \begin{array}{c|c} & s \qquad\qquad |S| \\ \hline s' & \begin{pmatrix} & & \\ & 0.3 & \\ & & \end{pmatrix} \\ |S| & \end{array} \longrightarrow pr(s'|s,a_1)$$

$$a_2 \quad \begin{array}{c|c} & s \\ \hline s' & \begin{pmatrix} & & \\ & 0.05 & \\ & & \end{pmatrix} \end{array} \longrightarrow pr(s'|s,a_2)$$

$$\vdots$$

$$a_K \quad \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}$$

- By construction next state only depends upon current state and action!

(4) Reward function on states.

$R : S \to \mathbb{R}$

HISTORY: tuples $\tau_t = (s_0, a_0, s_1, a_1, \ldots)$

- Sometimes we specify that s_0 is sampled from a distribution μ on states.

WRITE : $\boxed{M = (S, A, \mathbb{P}, R, \gamma, \mu).}$

Example from Barto Sutton.

Recycling robot:

- Collects empty cans.
- Decisions to be made based on current level of "charge"

$S = \{\text{high}, \text{low}\}$.

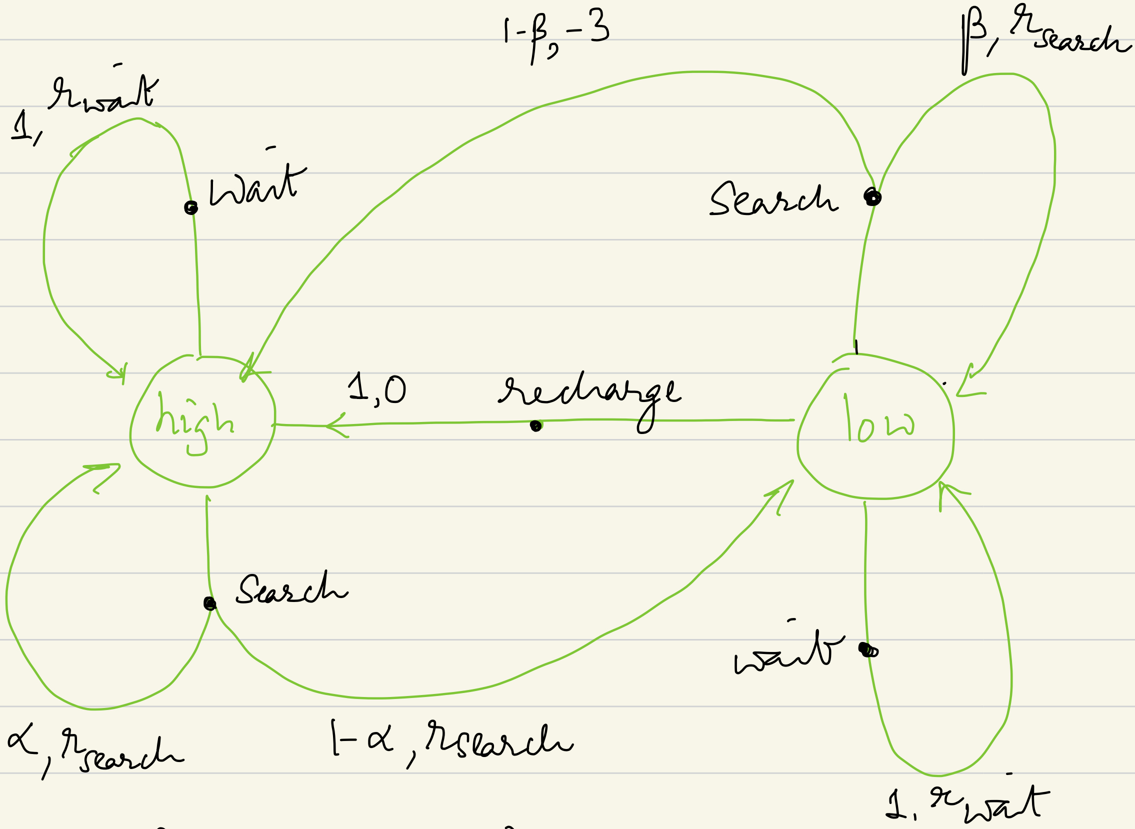
Agent decides
- Search
- Wait
- Recharge

high $\longrightarrow$ {search, wait}

low $\longrightarrow$ {search, wait, recharge}

Rewards: 0 most times

- positive if a can is found
- negative if battery drains.



- $p(\text{low} \mid \text{low}, \text{search}) = \beta$

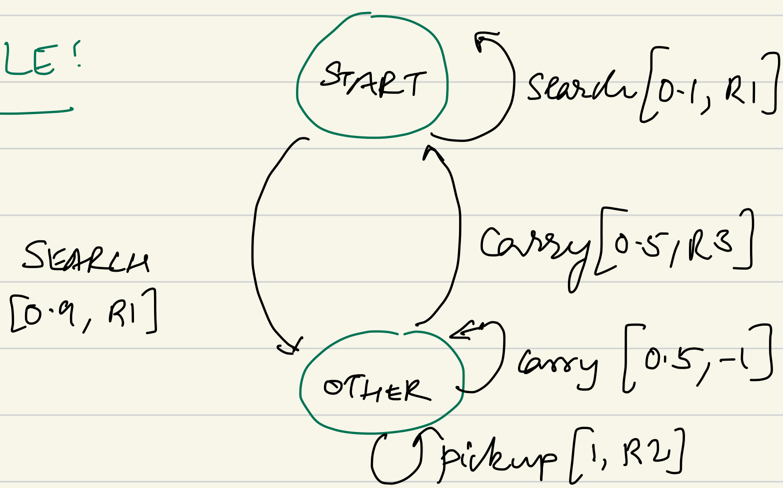
  $r(\text{low}, \text{search}, \text{low}) = r_{search}$

Figure depicts.

- State nodes
- action nodes

GOAL: Compute the value function and Q- value.

EXAMPLE!



Def:

a) Policy: $\pi : \mathcal{H} \to A$ is a map from histories to actions.

b) Deterministic, Stationary policy: is a strategy based on the current state $a_t = \pi(s_t)$

$$\pi : S \to A$$

c) Stochastic policy:

$$\pi : S \to \Delta(A)$$

distribution on actions

We write:

$$a_t \sim \pi(\cdot \mid s_t)$$

the action on the t-th step is drawn from the distribution $\pi(s_t)$

Policy value:

Expected reward when starting at s and following policy π.

- Finite horizon:

$$V_\pi(s) \triangleq \mathbb{E}_{a_t \sim \pi(s_t)} \left[\sum_{t=0}^{T} r(s_t, a_t) \,\middle|\, s_0 = s \right]$$

The expectation taken over the random selection of actions, and random state s_t reach.

- Infinite horizon:

Value function: For a policy π, the value function is the average, discounted sum of future rewards

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

Note: a) Assuming $r(s,a)$ is bounded b/w 0 & 1, each term is at most $1+\gamma+\gamma^2+\cdots$

- bounded by $\frac{1}{1-\gamma}$.

b) So we may normalize it

$$V^{\pi}(s) = (1-\gamma)\,\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, s_0 = s\right]$$

- Note we are working over an infinite horizon.

- for a single time step clearly the action giving max instantaneous reward is best.

- If we fix a horizon k:
  - we need to find a policy that gives max expected returns from time 0 to k

Could give us a policy that depends upon time a non-stationary policy.

- Popular to consider infinite horizon.

- $\boxed{\text{There is an optimal stationary policy}}$

VALUE FUNCTION for Recycling robot:

$$v^*(h) = \max \begin{cases} \alpha\left[r_s + \gamma v^*(h)\right] + (1-\alpha)\left[r_s + \gamma v^*(\ell)\right], \\ \left[r_w + \gamma v^*(h)\right] \end{cases}$$

$$v^*(\ell) = \max \begin{cases} \beta r_s - 3(1-\beta) + \gamma\left[(1-\beta)v^*(h) + \beta v^*(\ell)\right] \\ r_w + \gamma v^*(\ell) \\ \gamma v^*(h) \end{cases}$$

Def: Q-value:

The Q-value function is defined as

$$Q^{\pi}(s,a) = (1-\gamma)\, \mathbb{E}\left[\sum \gamma^{t} r(s_t, a_t) \,\middle|\, \pi,\ s_0 = s, a_0 = a \right]$$

Note that $a_0 = a$ in the above

GOAL: Given $s \in S$, find π maximizing

$$V^{\pi}(s)$$

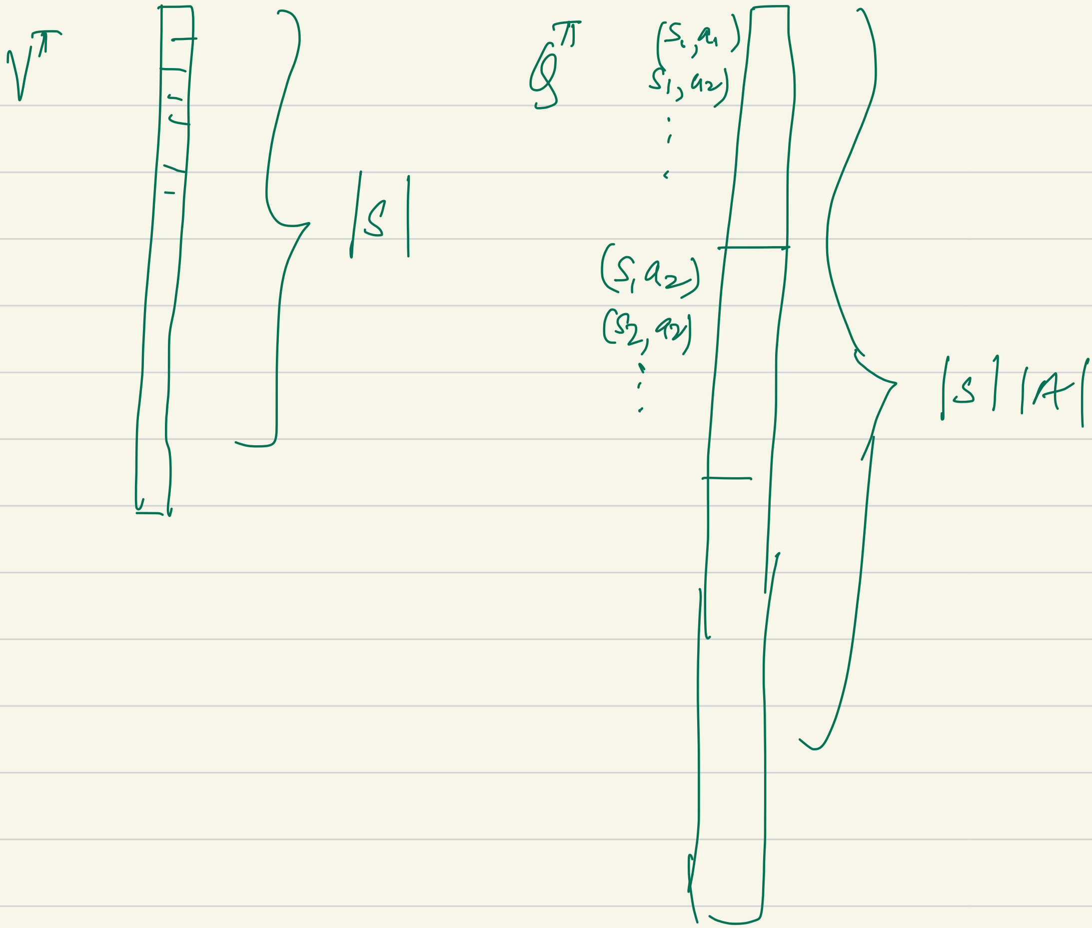
$=$

By definition: If π is deterministic

$$V^{\pi}(s) = Q^{\pi}(s, \pi(s))$$

$$Q^{\pi}(s,a) = (1-\gamma) r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot | s,a)} \left[V^{\pi}(s') \right]$$

Called Bellman equations



- Let P be a $|S||A| \times |S|$ matrix with

$$P_{(s,a),s'} = \Pr(s' \mid s, a)$$

Define P^π, $\left(|S|\times|A| \;\times\; |S|\times|A|\right)$ matrix

$$P^\pi_{(s,a)(s',a')} := \begin{cases} \Pr(s'|s,a) & \text{if } a' = \pi(s') \\ 0 & \end{cases}$$

If π is randomized stationary policy, set

$$P^\pi_{(s,a)(s',a')} := \mathbb{P}(s'|s,a)\cdot\pi(a'|s')$$

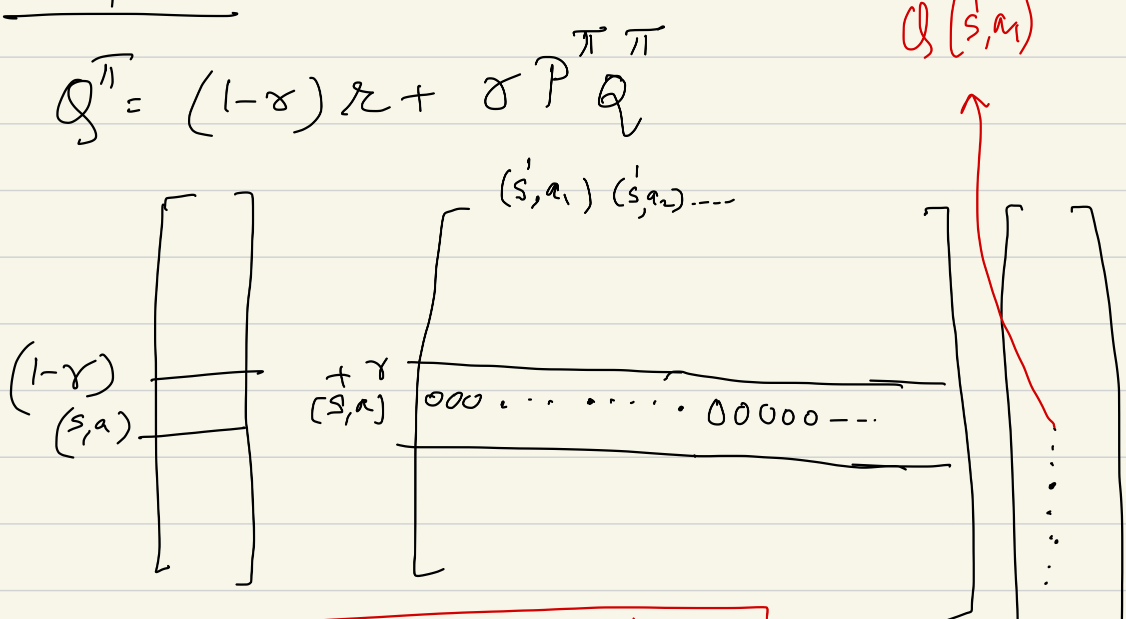
<u>Obs:</u> $$Q^\pi = (1-\gamma)r + \gamma P V^\pi$$

<u>Pf:</u>

$$\begin{array}{c} \\ (s,a) \\ \\ \end{array}\begin{bmatrix} \\ \hline \; \\ \hline \\ \end{bmatrix} = (1-\gamma)\begin{array}{c} \\ (s,a) \\ \\ \end{array}\begin{bmatrix} \\ \hline \; \\ \hline \\ \end{bmatrix} + \gamma\;\begin{array}{c} \\ \\ (s,a) \\ \\ \end{array}\begin{array}{c} \begin{array}{ccccc} & & s' & & \end{array} \\ \begin{bmatrix} & & & & \\ \hline 0 & \cdots & \underset{\Pr(s'|s,a)}{\bullet} & 0 & \cdots \\ \hline & & & & \end{bmatrix}\end{array}\begin{bmatrix} V^\pi(s_1) \\ \vdots \\ V^\pi(s') \\ \vdots \end{bmatrix}$$

· <u>In fact:</u>

$$Q^{\pi} = (1-\gamma) r + \gamma P^{\pi} Q^{\pi}$$



$$\boxed{\therefore \quad Q^{\pi} = (1-\gamma)\left[I - \gamma P^{\pi}\right]^{-1} r}$$

Invertible?

<u>Pf:</u> $\|(I - \gamma P^{\pi}) x\|_\infty = \|x - \gamma P^{\pi} x\|_\infty$

· $x = (x - \gamma P^{\pi} x) + \gamma P^{\pi} x$

$\therefore \quad \|x\|_\infty \leq \|x - \gamma P^{\pi} x\|_\infty + \gamma \|P^{\pi} x\|_\infty$

$$\therefore \|x - \gamma P^{\pi} x\|_\infty \geq \|x\|_\infty - \gamma \|P^{\pi} x\|_\infty$$

$$\geq \|x\|_\infty - \gamma \|x\|_\infty$$

$$\geq (1-\gamma) \|x\|_\infty$$

- Bellman optimality equations:

- <u>Thm</u>: For an MDP there is a single stationary and deterministic policy π simultaneously maximizing $V^{\pi}(s)$ for all $s \in S$ and $Q^{\pi}(s,a)$, $\forall\, s \in S,\ a \in A$;

Denote optimal policy by π^*;

- V^* corresponding value function
- Q^* " Q-value

$$V^*(s) = \max_{a \in A} Q^*(s,a)$$

$$Q^*(s,a) = (1-\gamma)\, r(s,a) + \gamma \underset{s' \sim P(\cdot \mid s,a)}{\mathbb{E}} \left[V^*(s') \right] \quad *$$

- Given Q

<u>define</u>: $\pi_Q(s) := \arg\max_{a \in A} Q(s,a)$

- So, $\pi^* = \pi_{Q^*}$
- Given Q, define $V_Q(s) := \max_{a \in A} Q(s,a)$
- <u>Bellman operator</u>: $T: \mathbb{R}^{S \times A} \longrightarrow \mathbb{R}^{S \times A}$

$$Q \longmapsto (1-\gamma) r + \gamma P V_Q$$

$$* \Longrightarrow \quad T Q^* = Q^* .$$

- Q^* - fixed point for Bellman operator
-

<u>THM</u>: Let $Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$. Then

- $\exists$ a stationary & deterministic policy π s.t $Q^{\pi} = Q^*$
- $Q \in \mathbb{R}^{S \times A}$ is equal to Q^* iff $TQ = Q$.

= <u>Pf</u>:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

$$= \max_{\pi} \left\{ (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[V^{\pi}(s') \right] \right\}$$

$$= (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s'} \left[\max_{\pi} V^{\pi}(s') \right]$$

WHY?

$$= (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s'} \left[\max_{\pi} Q^{\pi}(s', \pi(s')) \right]$$

$$= (1-\gamma) r(s,a) + \gamma \mathop{\mathbb{E}}_{s'} \left[\max_{\pi, a'} Q^{\pi}(s', a') \right]$$

$$\| \quad Q^*(s', a')$$

$\Leftarrow$ Let $Q = \mathcal{T}Q$.

Let $\pi = \pi_Q$; $\therefore Q = (1-\gamma)r + \gamma P^{\pi_Q} Q$

$\therefore \quad (1-\gamma)r = \left(1 - \gamma P^{\pi_Q}\right) Q^{\pi}$.

Let π' be a policy, $Q^{\pi'}$ — the Q-value

$$Q^{\pi'} = \left(1-\gamma P^{\pi'}\right)^{-1} (1-\gamma) r$$

$$= \left(1-\gamma P^{\pi'}\right)^{-1} \left(1 - \gamma P^{\pi_Q}\right) Q^{\pi}$$

$$\therefore Q^{\pi'} - Q^{\pi} = \left(1-\gamma P^{\pi'}\right)^{-1} \left[\left(1-\gamma P^{\pi_Q}\right) - \left(1-\gamma P^{\pi'}\right)\right] Q^{\pi}$$

$$= \gamma \left(1-\gamma P^{\pi'}\right)^{-1} \left[P^{\pi'} - P^{\pi_Q}\right] Q^{\pi}$$

$$\mathbb{E}_{s' \sim P(\cdot \mid s, a)} \left[Q^{\pi}(s', \pi'(s')) - Q^{\pi}(s', \pi(s'))\right] \leq 0$$

$\therefore \quad Q^{\pi'} \ll Q^{\pi}$ coordinate wise.

PART-2:

- Q-value iteration algorithm

- Start with random Q.
- Apply $Q \longleftarrow TQ$

<u>THM</u>: The above algorithm converges.

<u>Pf</u>: (a) Show that TQ is contracting

(b) Bound the distance b/w V^{π_Q} & V^* for any Q.

Lemma:

$$\|TQ - TQ'\|_\infty \le \gamma \|Q - Q'\|_\infty$$

Pf:

$\forall s$, $\left| V_Q(s) - V_{Q'}(s) \right| \le \max_{a \in A} \left| Q(s,a) - Q'(s,a) \right|$

w.l.o.g $V_Q(s) > V_{Q'}(s)$;

Suppose $V_Q(s) = Q(s,a)$

$$\text{LHS} = Q(s,a) - \max_{a' \in A} Q'(s,a')$$

$$\le Q(s,a) - Q'(s,a) \le \max_a \left| Q(s,a) - Q'(s,a) \right|$$

=

$$\|TQ - TQ'\|_\infty = \| \cancel{(1-\gamma)} r + \gamma P V_Q - (\cancel{(1-\gamma)} r + \gamma P V_{Q'}) \|_\infty$$

$$= \gamma \| P V_Q - P V_{Q'} \|_\infty$$

$$\leq \gamma \| V_Q - V_{Q'} \|_\infty$$

$$= \gamma \max_s | V_Q(s) - V_{Q'}(s) |$$

$$\leq \gamma \max_s \max_a | Q(s,a) - Q'(s,a) |$$

Defn

$$= \gamma \| Q - Q' \|_\infty$$

(2) $\forall Q \in \mathbb{R}^{S \times A}$

$$V^{\pi_Q} \geq V^* - \frac{2 \| Q - Q^* \|_\infty}{1-\gamma} \mathbb{1}$$

Pf: $V^*(s) - V^{\pi}(s)$

$$= Q^*(s, \pi^*(s)) - Q^{\pi_Q}(s, \pi_Q(s))$$

$$= Q^*(s,\pi^*(s)) - Q^*(s,\pi_Q(s)) + \underbrace{Q^*(s,\pi_Q(s)) - Q^{\pi_Q}(s,\pi_Q(s))}$$

$$= Q^*(s,\pi^*(s)) - Q^*(s,\pi_Q(s)) +$$

$$\gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \left[V^*(s') - V^{\pi_Q}(s') \right]$$

$$\le Q^*(s,\pi^*(s)) - Q(s,\pi^*(s)) + \underbrace{\qquad}$$

$$+ Q(s,\pi_Q(s)) - Q^*(s,\pi_Q(s))$$

$$\left(\because Q(s,\pi_Q(s)) \ge Q(s,\pi^*(s)) \right)$$

$$\le 2\|Q - Q^*\|_\infty + \gamma \|V^* - V^{\pi_Q}\|_\infty$$

$$\therefore \; V^* - V^{\pi_Q} \le 2\|Q - Q^*\|_\infty + \gamma V^* - \gamma V^{\pi_Q}$$

$$\therefore \quad V^{\pi_Q} \geq V^* - \frac{2\|Q - Q^*\|_\infty}{1-\gamma} \cdot \mathbb{1}$$

THM: If $k \geq \frac{1}{1-\gamma} \log\left(\frac{2}{\varepsilon(1-\gamma)}\right)$ the Q-value update algorithm converges to a Value vector $\tilde{V} \geq V^* - \varepsilon \cdot \mathbb{1}$.

Proof: Start with $Q^0 = 0$, $\quad Q^{k+1} = T Q^k$

$$= T^k Q^0$$

Note: $Q^* = T Q^*$

$$\|Q^{(k)} - Q^*\|_\infty = \|T^k Q_0 - T^k Q^*\|_\infty$$

$$\leq \gamma^k \|Q_0 - Q^*\|_\infty \leq (1-(1-\gamma))^k \underbrace{\|Q^*\|_\infty}_{\leq 1}$$

$$\leq \exp\left(-(1-\gamma)k\right)$$

NEXT TIME : POLICY UPDATE