DP Operators

- DP operators

$$(T_H V_{t+1})(s) = \max_{a \in \mathcal{A}} \left(\overline{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} P_{t,a}(s'|s) V_{t+1}(s') \right)$$

$$(F_t Q_{t+1})(s,a) = \overline{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} P_{t,a}(s'|s) \max_{a' \in \mathcal{A}} Q_{t+1}(s', a')$$

- Value iteration

$$V^*_H(s) = 0 \qquad V^*_t = T_t V^*_{t+1}$$

$$Q^*_{H-1}(s) = R_{H-1,a}(s) \qquad Q^*_t = F_t Q^*_{t+1}$$

- Monotonicity

$$V_{t+1} \geq V'_{t+1} \qquad \Longrightarrow \qquad T_t V_{t+1} \geq T_t V'_{t+1}$$

$$Q_{t+1} \geq Q'_{t+1} \qquad \Longrightarrow \qquad F_t Q_{t+1} \geq F_t Q'_{t+1}$$

Asynchronous Value Iteration

- Start with some Q^0
- For k = 0, 1, 2, …
  - Select (t^k, s^k, a^k)
  - Obtain Q^{k+1} from Q^k
    - Apply value iteration update at (t^k, s^k, a^k)
- Note that
  - Q^{k+1} only differ at Q^k
  - For an appropriately selected sequence of updates, this is the same as regular value iteration
- Theorem: if the sequence samples each triple infinitely often then $Q^k \to Q^*$

Real-Time Value Iteration

- Asynchronous value iteration for a particular sequence of updates

- Episodic "learning"
  - Updates require knowledge of

- Start with some Q^0

- For $\ell = 0, 1, 2$
  - Sample initial state $s_0^\ell \sim \rho(\cdot)$
  - For $t = 0, 1, \ldots, H$
    - Select "greedy" action $a_t^\ell \in \arg\max_{a \in \mathcal{A}} Q_t^\ell(s_t^\ell, a)$
    - Update value at (t, s_t^ℓ, a_t^ℓ)
    - Sample next state $s_{t+1}^\ell \sim P_{t,a_t^\ell}(\cdot | s_t^\ell)$

- "greedy policy"
  - Does not necessarily sample all triples

- Theorem: if $Q^0 \geq Q^*$ then actions are optimal after some (random) finite time

Optimism of Values

- Lemma: $Q^\ell \geq Q^* \qquad \forall \ell$

- Proof:

$$Q_t^\ell(x,a) \geq Q_t^*(x,a)$$

$$\Downarrow$$

$$Q_t^{\ell+1}(x,a) = (F_t Q_{t+1}^\ell)(x,a) \geq (F_t Q_{t+1}^*)(x,a) = Q_t^*(x,a)$$

Eventual Optimality

- Let $\mathbb{X}_\infty = \{(t, s, a) : \text{sampled i. o.}\}$

- Let τ from which we remain in $\mathbb{X}_\infty$

- Updates from τ equivalent to those for an auxiliary MDP where

$$R_{t,a}(s) \approx -\infty \qquad \forall (t, s, a) \in \mathbb{X}_\infty$$

- Updates converge to optimal value function for this auxiliary problem, hence,

$$Q_t^\infty(s, a) \leq Q_t^*(s, a) \qquad \forall (t, s, a) \in \mathbb{X}_\infty$$

- It follow that

$$Q_t^\infty(s, a) = Q_t^*(s, a) \qquad \forall (t, s, a) \in \mathbb{X}_\infty$$

- Hence, actions are eventually optimal for both auxiliary and original MDP

Q-Learning

- Given an observed transition (t, s, a, s'), update value function:
  - If $t = H - 1$

$$Q_{H-1}^{k+1}(s,a) \leftarrow (1-\gamma_k) Q_{H-1}^{k}(s,a) + \gamma_k r$$

  - If $t < H - 1$

$$Q_t^{k+1}(s,a) \leftarrow (1-\gamma_k) Q_t^{k}(s,a) + \gamma_k \left(r + \max_{a' \in \mathcal{A}} Q_{t+1}^{k}(s', a') \right)$$

- For $(\bar{t}, \overline{s}, \overline{a}) \neq (t, s, a)$

$$Q_{\bar{t}}^{k+1}(\overline{s}, \overline{a}) = Q_{\bar{t}}^{k}(\overline{s}, \overline{a})$$

Stochastic Approximation

- Consider an IID sequence $x_0, x_1, \ldots$
- Law of large numbers

$$\overline{x}_K = \frac{1}{K} \sum_{k=1}^{K} x_k \to \mathbb{E}[x_0]$$

- Another way to compute the average

$$\overline{x}_0 = x_0$$

$$\overline{x}_{k+1} = \left(1 - \frac{1}{k+1}\right) \overline{x}_k + \gamma_k x_{k+1}$$

- Generalization

$$\overline{x}_{k+1} = (1 - \gamma_k) \overline{x}_k + \gamma_k x_{k+1}$$

- Law of large numbers applies if

$$\sum_k \gamma_k = \infty \qquad \sum_k \gamma_k^2 < \infty$$

Convergence

- Consider updating based on (t, s, a, s')
- If $t = H - 1$

$$Q_{H-1}^{k+1}(s,a) \leftarrow (1-\gamma_k) Q_{H-1}^{k}(s,a) + \gamma_k r$$

$$\mathbb{E}[r|s,a] = \overline{R}_{t,a}(s)$$

$$Q_{H-1}^{k}(s,a) \rightarrow \overline{R}_{t,a}(s)$$

- If $t < H - 1$ and $Q_{t+1}^{k} = Q_{t+1}^{*}$

$$Q_{t}^{k+1}(s,a) \leftarrow (1-\gamma_k) Q_{t}^{k}(s,a) + \gamma_k \left(r + \max_{a' \in \mathcal{A}} Q_{t+1}^{k}(s',a') \right)$$

$$\begin{aligned} \mathbb{E}\left[r + \max_{a' \in \mathcal{A}} Q_{t+1}^{k}(s',a') \,\middle|\, Q_{t+1}^{k}, s, a \right] &= (T_t Q_{t+1}^{k})(s,a) \\ &= Q_t^{*}(s,a) \end{aligned}$$

$$Q_t^{k}(s,a) \rightarrow Q_t^{*}(s,a)$$

- Convergence if each (t,s,a) updated infinitely often with appropriate step sizes