# $Q$-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

# “Real-Time” Q-Learning

- Episodic “learning”
- 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 $(t, s_t^\ell, a_t^\ell)$
    - 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
- Does this converge

# Dithering

- epsilon-greedy exploration

$$a_t^\ell \in \arg\max_{a \in \mathcal{A}} Q_t^\ell(s_t^\ell, a) \qquad \text{w. p. } \epsilon$$

$$a_t^\ell \sim \text{unif}(\mathcal{A}) \qquad \text{w. p. } 1-\epsilon$$

- Boltzmann exploration

$$a_t^\ell \sim \exp\left(\beta Q_t^\ell(s_t^\ell, \cdot)\right)$$

- Theorem: For all reachable $(t, s, a)$,

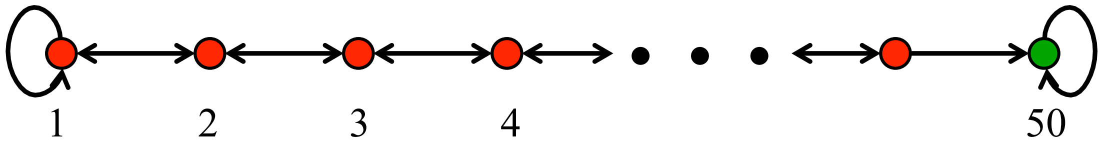
$$Q_t^\ell(s, a) \to Q_t^*(s, a)$$

# Regret

$$\text{Regret}(T) = \sum_{\ell=1}^{T} \left( V_\ell^*(s_0^\ell) - \sum_{t=0}^{H-1} r_t^\ell \right)$$

$$\mathbb{E}\left[\text{Regret}(T)\right] = \sum_{\ell=1}^{T} \sum_{s \in \mathcal{S}} \rho(s) \left( V_\ell^*(s) - V_\ell^{\mu^\ell}(s) \right)$$

# Worst-Case Regret of Dithering




- Deterministic MDP
  - Start at state 1
  - Actions: $A$ = {left, right}
  - Horizon: $H = 50$
  - Receive reward 1 only at state 50

- What is the optimal strategy?

- Regret of dithering?
  - Initialize with $Q = 0$
  - How many episodes are required to learn?

$$\approx 2^{50}$$

# Exploration Via Optimism