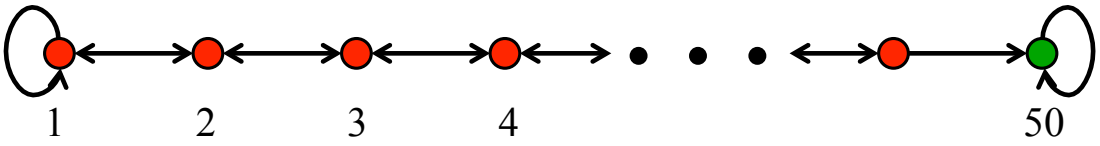
Worst-Case Regret of Dithering



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

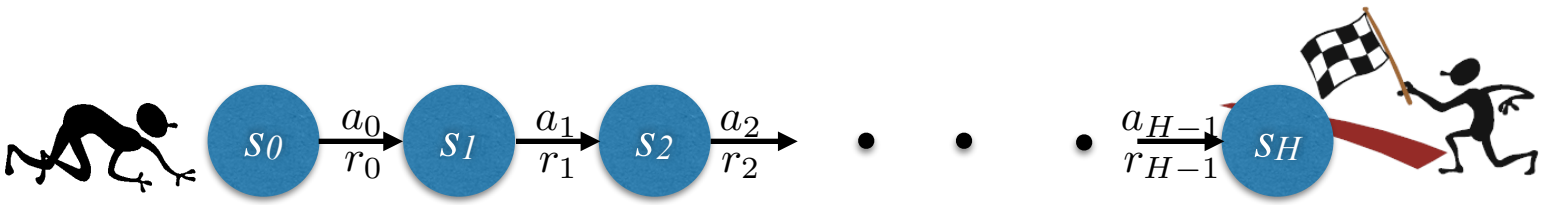
- What is the optimal strategy?

- Regret of dithering?
  - Initialize with $Q = 0$
  - How many episodes are required to learn? $\approx 2^{50}$
  - Expected regret

$$\mathbb{E}[\text{Regret}(L)] \approx L \qquad \text{for } L \ll 2^{50}$$

- Efficient learning requires planning to learn

Model Learning



- Reinforcement learning algorithm
  - Given all past observations
  - Select policy for next episode
  - Apply this policy throughout the next episode

- Model learning
  - Data for learning about one triple (*t*,*s*,*a*)

$$\Lambda = \{\ell : (t, s_t^\ell, a_t^\ell) = (t, s, a)\}$$

  - Learn expected reward

$$\hat{R}_{t,a}(s) = \frac{1}{|\Lambda|} \sum_{\ell \in \Lambda} r_t^\ell$$

  - Learn transition probabilities

$$\hat{P}_{t,a}(s'|s) = \frac{1}{|\Lambda|} \sum_{\ell \in \Lambda} \mathbf{1}(s' = s_{t+1}^\ell)$$

- Given a model, how to select next policy?

Confidence Sets and UCB

- UCB Concept
  - Maintain confidence set Θ_t
    - Set of statistically plausible models
  - Optimize optimistically $\max_{x \in \mathbb{X}} \max_{\theta \in \Theta_t} f_\theta(x)$

- Confidence sets for MDPs
  - Expected rewards

$$\left\{ \tilde{R}_{t,a}(s) : \left| \tilde{R}_{t,a}(s) - \hat{R}_{t,a}(s) \right| \leq \sqrt{\frac{c_1 + c_2 \log(\ell/\delta)}{\max\{1, n_\ell(t,s,a)\}}} \right\}$$

  - Transition probabilities

$$\left\{ \tilde{P}_{t,a}(\cdot|s) : \left| \tilde{P}_{t,a}(\cdot|s) - \hat{P}_{t,a}(\cdot|s) \right| \leq \sqrt{\frac{c_3 + c_4 \log(\ell/\delta)}{\max(1, n_\ell(t,s,a))}} \right\}$$

  - Observation count $n_\ell(t,s,a)$
  - Tolerance parameter δ

A Simple UCB Algorithm

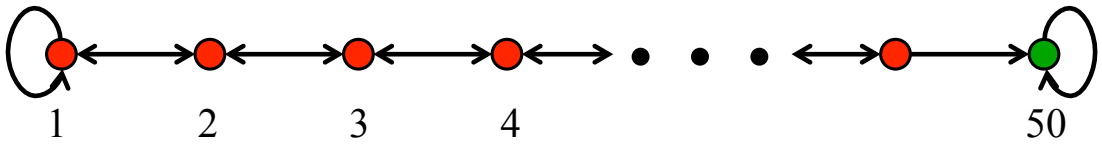
- At the start of episode ℓ
  - Solve sample MDP

$$(\mathcal{S}, \mathcal{A}, \hat{R}, \hat{P}, \rho, H) \Rightarrow \hat{V}$$

  - Optimistic optimization

$$\mu_t^\ell \in \arg\max_{a \in \mathcal{A}} \max_{\tilde{R}_{t,a}(s), \tilde{P}_{t,a} \cdot | s)} \left(\tilde{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} \tilde{P}_{t,a}(s'|s) \hat{V}(s') \right)$$

- Does this plan to learn?



UCRL

- Optimistic optimization

$$\mu^{\ell} \in \arg\max_{\mu} \max_{\tilde{R},\tilde{P}} \mathbb{E}\left[\sum_{t=0}^{H-1} \tilde{R}_{t,a_t}(s_t) \;\middle|\; a_t = \mu_t(s_t), s_{t+1} \sim \tilde{P}_{t,a_t}(\cdot|s_t)\right]$$

- Optimistic value iteration

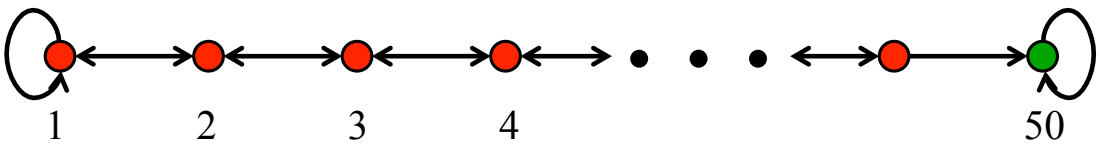
$$\tilde{Q}_{H-1}(s,a) = \max_{\tilde{R}_{H-1,a}(s)} \tilde{R}_{H-1,a}(s)$$

$$\tilde{Q}_t(s,a) = \max_{\tilde{R}_{t,a}(s), \tilde{P}_{t,a}(\cdot|s)} \left(\tilde{R}_{t,a}(s) + \sum_{s' \in \mathcal{S}} \tilde{P}_{t,a}(s'|s) \max_{a' \in \mathcal{A}} \tilde{Q}_{t+1}(s',a')\right)$$

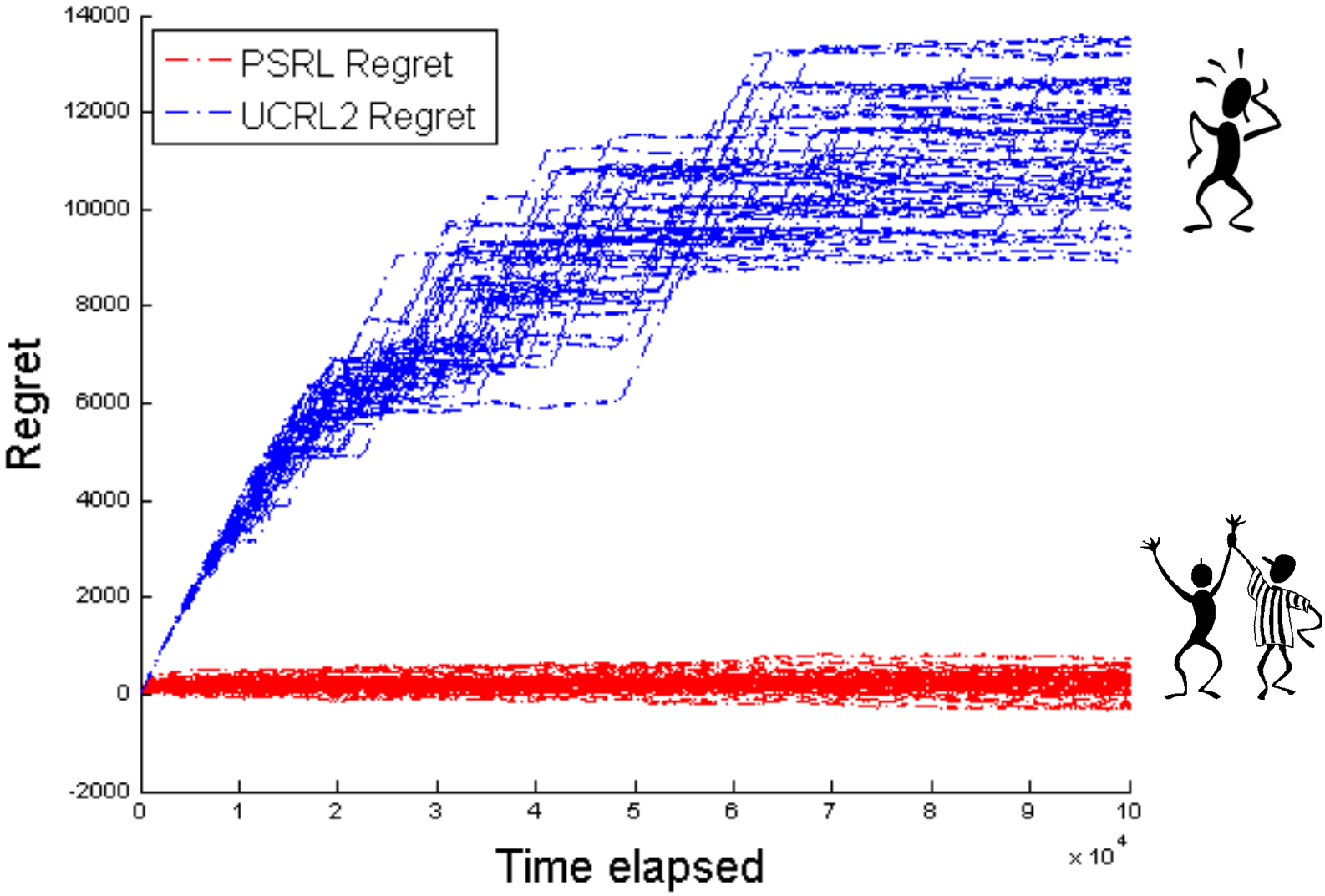
- Use greedy policy

$$\mu_t^{\ell}(s) \in \arg\max_{a \in \mathcal{A}} \tilde{Q}_t(s,a)$$

- Does this plan to learn?



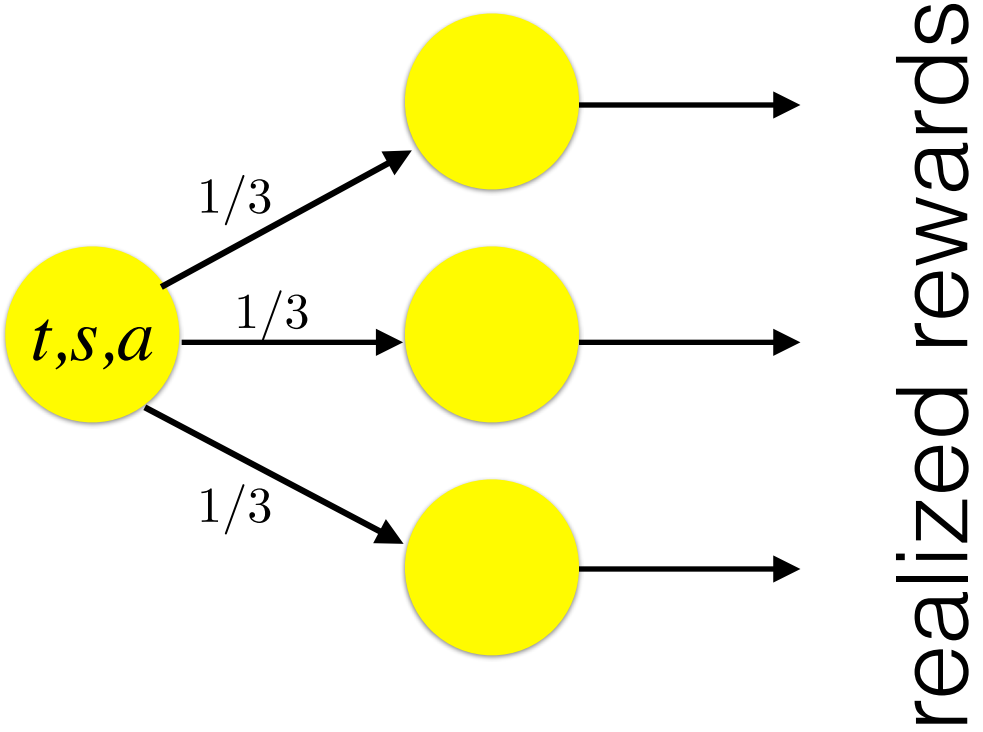
UCRL versus PSRL



$$\sum_{\ell=1}^{L} \left(V_0^*(s_0^\ell) - \sum_{t=0}^{H-1} r_t^\ell \right)$$

Conservatism of Confidence Sets

- Consider a single (*H*-2, *s*, *a*)
  - No immediate rewards
  - Uniform transitions $P_{H-2,a}(\cdot|s) \sim \text{unif}(\mathcal{S})$
  - Actions taken at time *H*-1 are indistinguishable



- Observations
  - *n* samples at each (*H*-1, *s'*, *a'*)
  - Rewards for each (*H*-1, *s'*, *a'*) exhibit sample mean 0 and sample variance σ^2

- What is a reasonable optimistic $\tilde{Q}_{H-2}(s,a)$?

$$\sigma \qquad \text{or} \qquad \frac{\sigma}{\sqrt{|\mathcal{S}}} \; ?$$

Coupling Confidence Sets

- The "right" confidence sets couple estimates across states, actions, and time
- Similar issue with UCB for online LP
  - Box-shaped confidence sets statistically inefficient
  - Ellipsoidal confidence set statistically efficient but computationally inefficient
- One solution: Thompson sampling
  - Randomized approximation to UCB with "right" confidence sets