CMU 15-889e

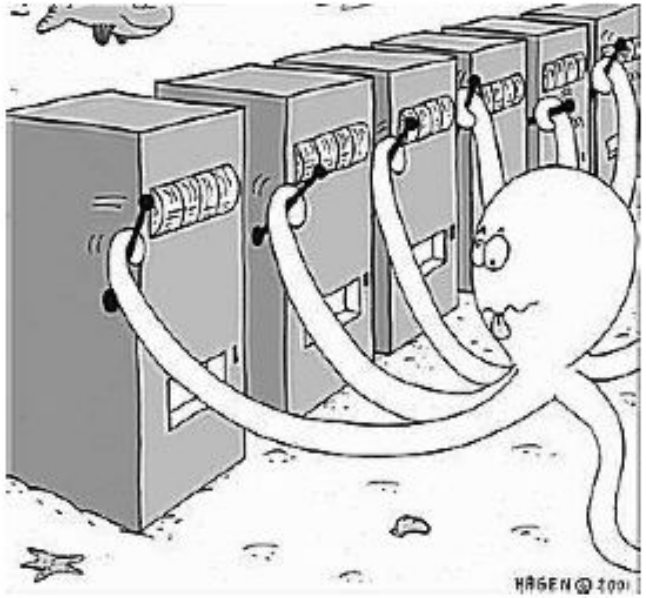
Bandits, Regret & MDPs

Emma Brunskill

Fall 2015

The Multi-Armed Bandit

- A multi-armed bandit is a tuple $\langle \mathcal{A}, \mathcal{R} \rangle$
- $\mathcal{A}$ is a known set of m actions (or "arms")
- $\mathcal{R}^a(r) = \mathbb{P}[r|a]$ is an unknown probability distribution over rewards
- At each step t the agent selects an action $a_t \in \mathcal{A}$
- The environment generates a reward $r_t \sim \mathcal{R}^{a_t}$
- The goal is to maximise cumulative reward $\sum_{\tau=1}^{t} r_\tau$



slide from David Silver

Regret

- The *action-value* is the mean reward for action a,

$$Q(a) = \mathbb{E}[r|a]$$

- The *optimal value* V^* is

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

- The *regret* is the opportunity loss for one step

$$I_t = \mathbb{E}[V^* - Q(a_t)]$$

- The *total regret* is the total opportunity loss

$$L_t = \mathbb{E}\left[\sum_{\tau=1}^{t} V^* - Q(a_\tau)\right]$$

- Maximise cumulative reward $\equiv$ minimise total regret

slide from David Silver

Counting Regret

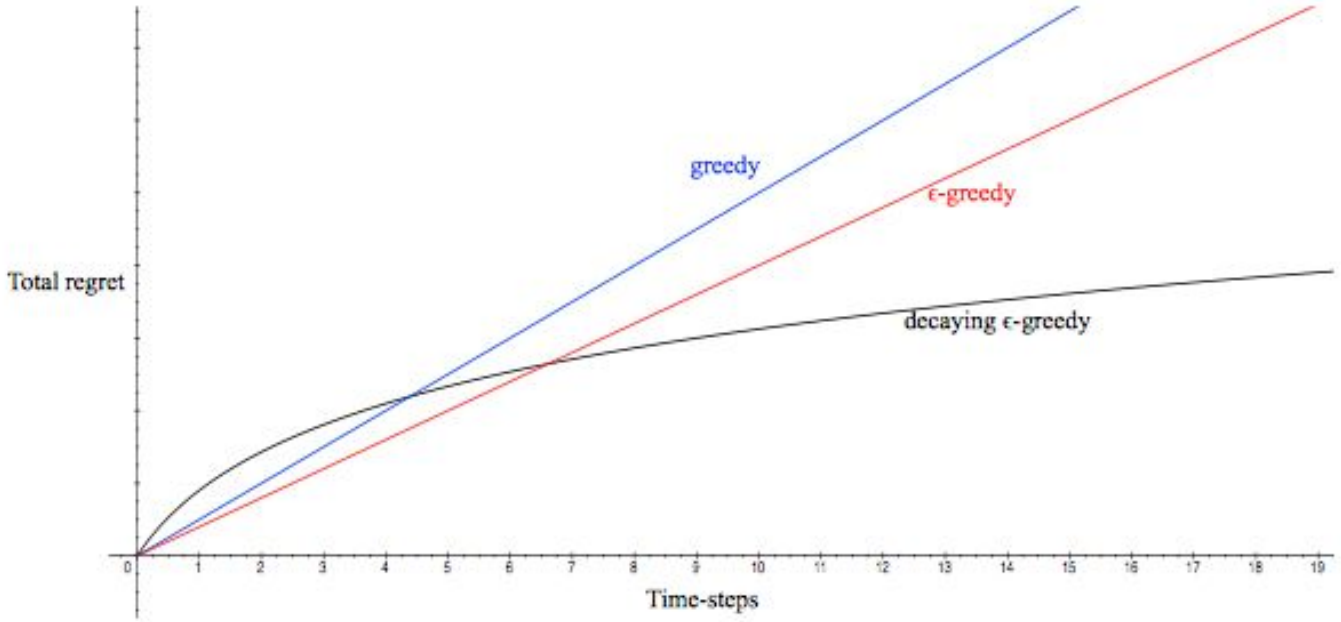
- The *count* $N_t(a)$ is expected number of selections for action a
- The *gap* Δ_a is the difference in value between action a and optimal action a^*, $\Delta_a = V^* - Q(a)$
- Regret is a function of gaps and the counts

$$
\begin{aligned}
L_t &= \mathbb{E}\left[\sum_{\tau=1}^{t} V^* - Q(a_\tau)\right] \\
&= \sum_{a \in \mathcal{A}} \mathbb{E}\left[N_t(a)\right] \left(V^* - Q(a)\right) \\
&= \sum_{a \in \mathcal{A}} \mathbb{E}\left[N_t(a)\right] \Delta_a
\end{aligned}
$$

- A good algorithm ensures small counts for large gaps
- Problem: gaps are not known!

slide from David Silver

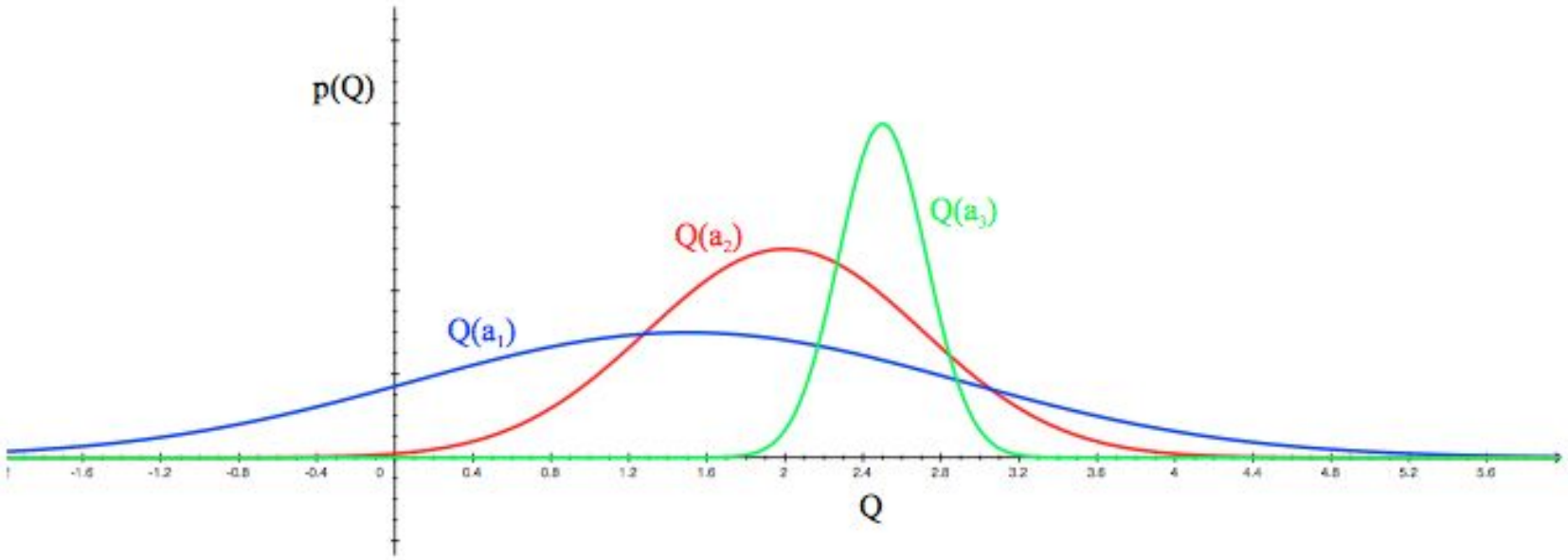
Linear or Sublinear Regret



- If an algorithm forever explores it will have linear total regret
- If an algorithm never explores it will have linear total regret
- Is it possible to achieve sublinear total regret?

slide from David Silver

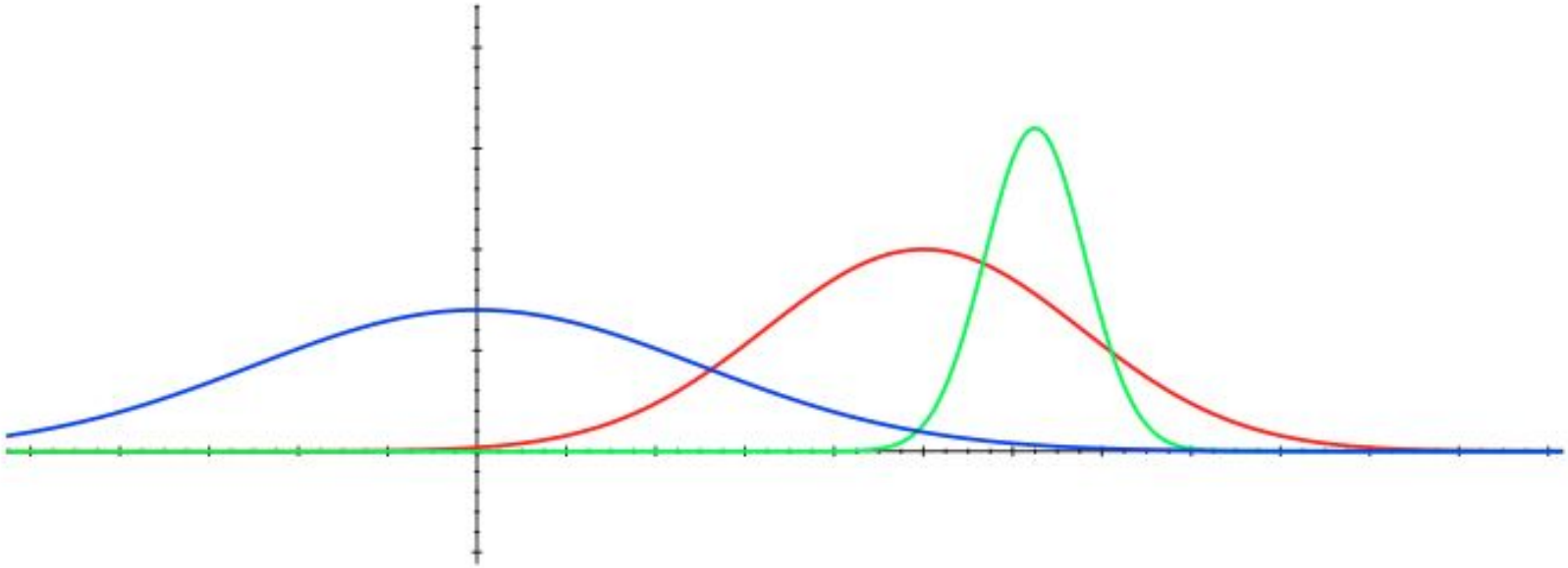
Optimism in the Face of Uncertainty



- Which action should we pick?
- The more uncertain we are about an action-value
- The more important it is to explore that action
- It could turn out to be the best action

slide from David Silver

Optimism in the Face of Uncertainty (2)



- After picking blue action
- We are less uncertain about the value
- And more likely to pick another action
- Until we home in on best action

slide from David Silver

Upper Confidence Bounds

- Estimate an upper confidence $\hat{U}_t(a)$ for each action value
- Such that $Q(a) \leq \hat{Q}_t(a) + \hat{U}_t(a)$ with high probability
- This depends on the number of times $N(a)$ has been selected
  - Small $N_t(a) \Rightarrow$ large $\hat{U}_t(a)$ (estimated value is uncertain)
  - Large $N_t(a) \Rightarrow$ small $\hat{U}_t(a)$ (estimated value is accurate)
- Select action maximising Upper Confidence Bound (UCB)

$$a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\ \hat{Q}_t(a) + \hat{U}_t(a)$$

slide from David Silver

Hoeffding's Inequality

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_t$ be i.i.d. random variables in [0,1], and let $\overline{X}_t = \frac{1}{\tau}\sum_{\tau=1}^{t} X_\tau$ be the sample mean. Then

$$\mathbb{P}\left[\mathbb{E}\left[X\right] > \overline{X}_t + u\right] \leq e^{-2tu^2}$$

- We will apply Hoeffding's Inequality to rewards of the bandit
- conditioned on selecting action a

$$\mathbb{P}\left[Q(a) > \hat{Q}_t(a) + U_t(a)\right] \leq e^{-2N_t(a)U_t(a)^2}$$

slide from David Silver

Calculating Upper Confidence Bounds

- Pick a probability p that true value exceeds UCB
- Now solve for $U_t(a)$

$$e^{-2N_t(a)U_t(a)^2} = p$$

$$U_t(a) = \sqrt{\frac{-\log p}{2N_t(a)}}$$

- Reduce p as we observe more rewards, e.g. $p = t^{-4}$
- Ensures we select optimal action as $t \to \infty$

$$U_t(a) = \sqrt{\frac{2 \log t}{N_t(a)}}$$

slide from David Silver

UCB1

- This leads to the UCB1 algorithm

$$a_t = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q(a) + \sqrt{\frac{2 \log t}{N_t(a)}}$$

Theorem

The UCB algorithm achieves logarithmic asymptotic total regret

$$\lim_{t \to \infty} L_t \leq 8 \log t \sum_{a | \Delta_a > 0} \Delta_a$$

slide from David Silver

UCB bandits with independent arms
Proof on board (and/or see lecture notes)

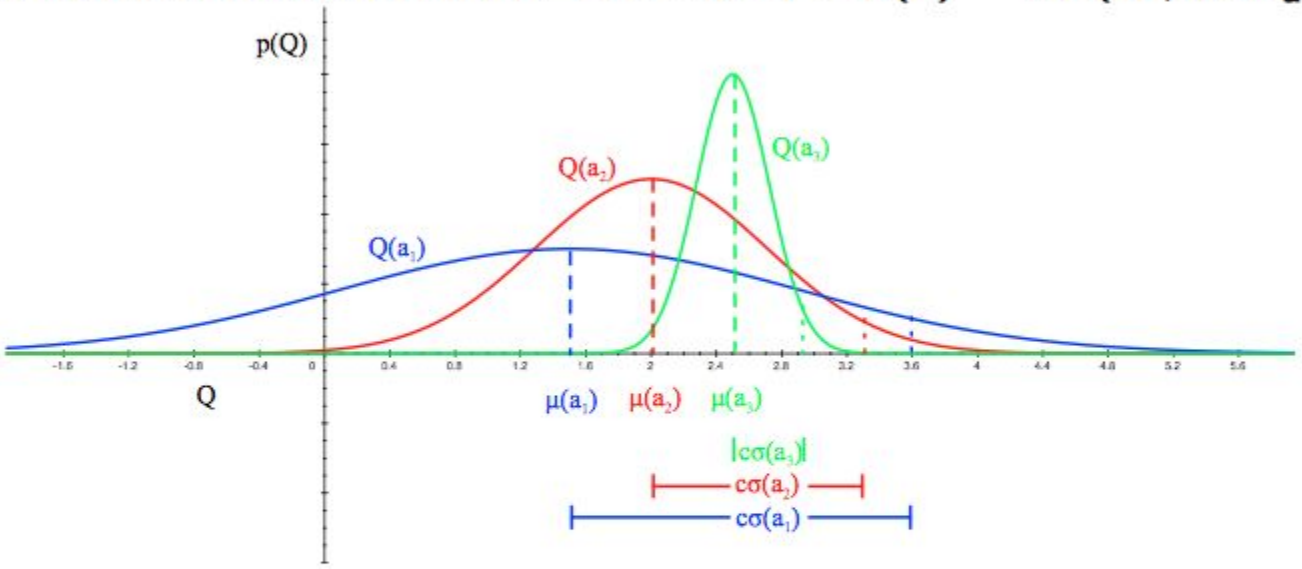
Bayesian Bandits

- So far we have made no assumptions about the reward distribution $\mathcal{R}$
  - Except bounds on rewards
- Bayesian bandits exploit prior knowledge of rewards, $p[\mathcal{R}]$
- They compute posterior distribution of rewards $p[\mathcal{R} \mid h_t]$
  - where $h_t = a_1, r_1, ..., a_{t-1}, r_{t-1}$ is the history
- Use posterior to guide exploration
  - Upper confidence bounds (Bayesian UCB)
  - Probability matching (Thompson sampling)
- Better performance if prior knowledge is accurate

slide from David Silver

Bayesian UCB Example: Independent Gaussians

- Assume reward distribution is Gaussian, $\mathcal{R}_a(r) = \mathcal{N}(r; \mu_a, \sigma_a^2)$



- Compute Gaussian posterior over μ_a and σ_a^2 (by Bayes law)

$$p\left[\mu_a, \sigma_a^2 \mid h_t\right] \propto p\left[\mu_a, \sigma_a^2\right] \prod_{t \mid a_t = a} \mathcal{N}(r_t; \mu_a, \sigma_a^2)$$

- Pick action that maximises standard deviation of $Q(a)$

$$a_t = \text{argmax}\ \mu_a + c\sigma_a / \sqrt{N(a)}$$

slide from David Silver

Probability Matching

- Probability matching selects action a according to probability that a is the optimal action

$$\pi(a \mid h_t) = \mathbb{P}\left[Q(a) > Q(a'), \forall a' \neq a \mid h_t\right]$$

- Probability matching is optimistic in the face of uncertainty
  - Uncertain actions have higher probability of being max
- Can be difficult to compute analytically from posterior

slide from David Silver

Thompson Sampling

- Thompson sampling implements probability matching

$$\pi(a \mid h_t) = \mathbb{P}\left[Q(a) > Q(a'), \forall a' \neq a \mid h_t\right]$$
$$= \mathbb{E}_{\mathcal{R}|h_t}\left[\mathbf{1}(a = \underset{a \in \mathcal{A}}{\text{argmax}}\ Q(a))\right]$$

- Use Bayes law to compute posterior distribution $p\left[\mathcal{R} \mid h_t\right]$
- Sample a reward distribution $\mathcal{R}$ from posterior
- Compute action-value function $Q(a) = \mathbb{E}\left[\mathcal{R}_a\right]$
- Select action maximising value on sample, $a_t = \underset{a \in \mathcal{A}}{\text{argmax}}\ Q(a)$

slide from David Silver

Empirically Thompson Sampling Does Well, Sometimes Better than UCB!
Example: Selecting Advertisements to Maximize Click Through Rates

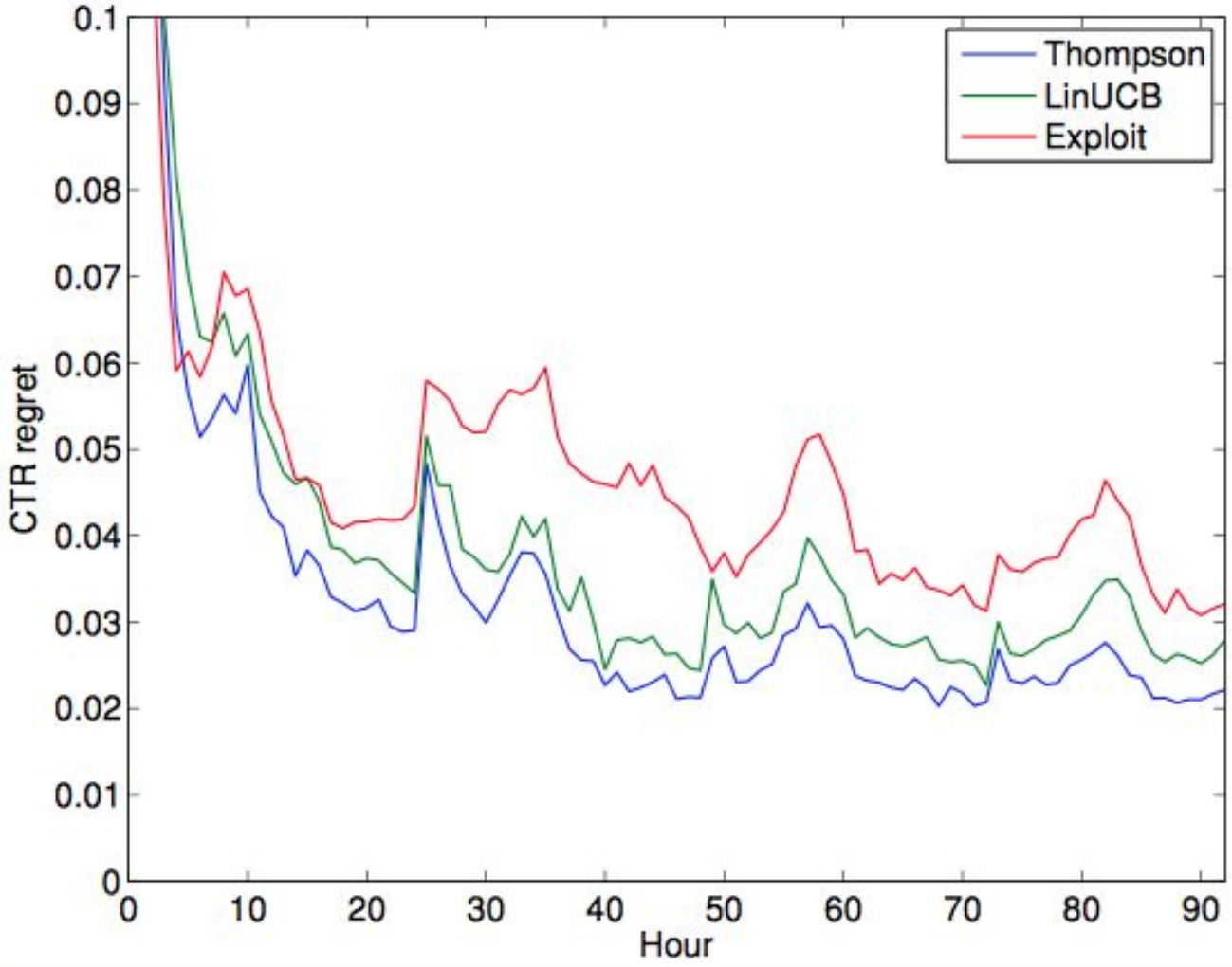


image from Chapelle & Li
NIPS 2011

Posterior sampling with independent arms
Proof on board (and/or see lecture notes)

Exploration/Exploitation Principles to MDPs

The same principles for exploration/exploitation apply to MDPs

- Naive Exploration
- Optimistic Initialisation
- Optimism in the Face of Uncertainty
- Probability Matching
- Information State Search

slide from David Silver

Optimistic Initialisation: Model-Based RL

- Construct an optimistic model of the MDP
- Initialise transitions to go to heaven
  - (i.e. transition to terminal state with r_{max} reward)
- Solve optimistic MDP by favourite planning algorithm
  - policy iteration
  - value iteration
  - tree search
  - ...
- Encourages systematic exploration of states and actions
- e.g. RMax algorithm (Brafman and Tennenholtz)

slide from David Silver

Upper Confidence Bounds for RL

Ex. Model Based Interval Estimation

On each time step t:

- Construct confidence intervals/sets over transition and reward models given data

$$CI = \{\tilde{T}(s,a,\cdot) \mid ||\tilde{T}(s,a,\cdot) - \hat{T}(s,a,\cdot)||_1 \leq \epsilon^T_{n(s,a)}\}$$

$$\epsilon^T_{n(s,a)} = \sqrt{\frac{2[\ln(2^{|S|} - 2) - \ln(\delta_T)]}{n(s,a)}}$$

$$\tilde{R}(s,a) = \hat{R}(s,a) + \sqrt{\frac{\ln(2/\delta_R) R_{\max}{}^2}{2n(s,a)}}$$

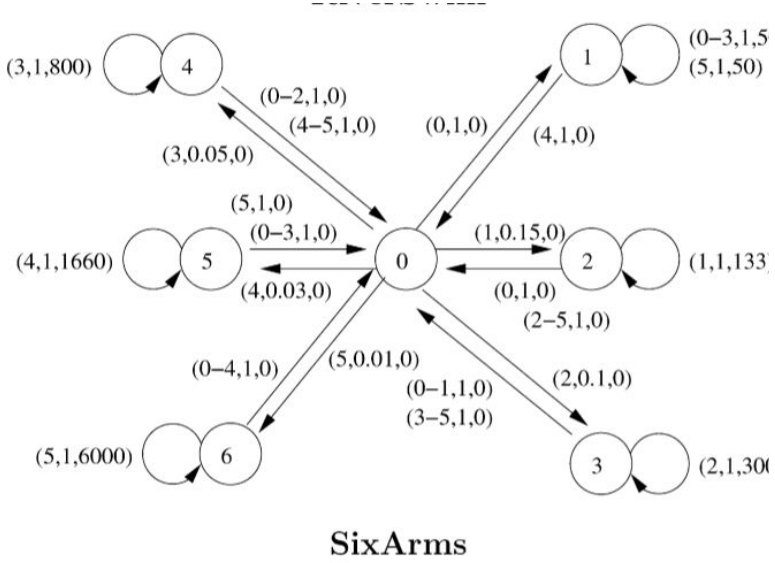
- Use to compute an optimistic Q

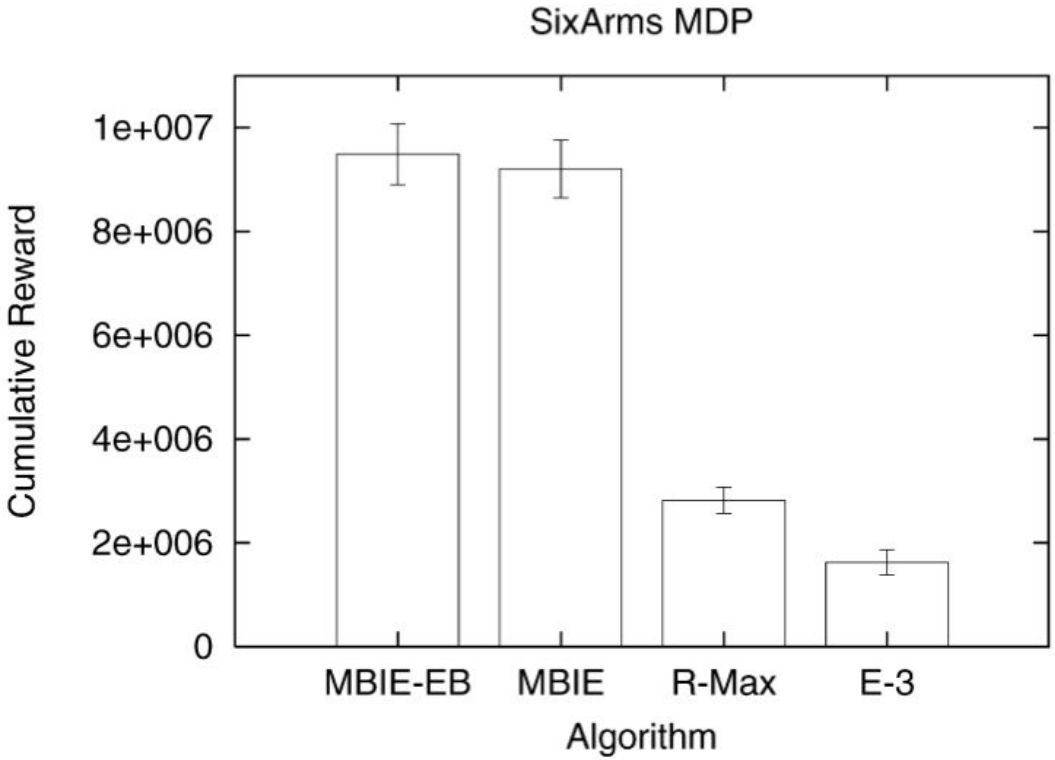
$$\tilde{Q}(s,a) = \tilde{R}(s,a) + \max_{\tilde{T}(s,a,\cdot) \in CI} \gamma \sum_{s'} \tilde{T}(s,a,s') \max_{a'} \tilde{Q}(s',a')$$

- a = argmax Q(s,a)

see Strehl & Littman 2008
IE idea dates at least to early 1990s, see Kaelbling

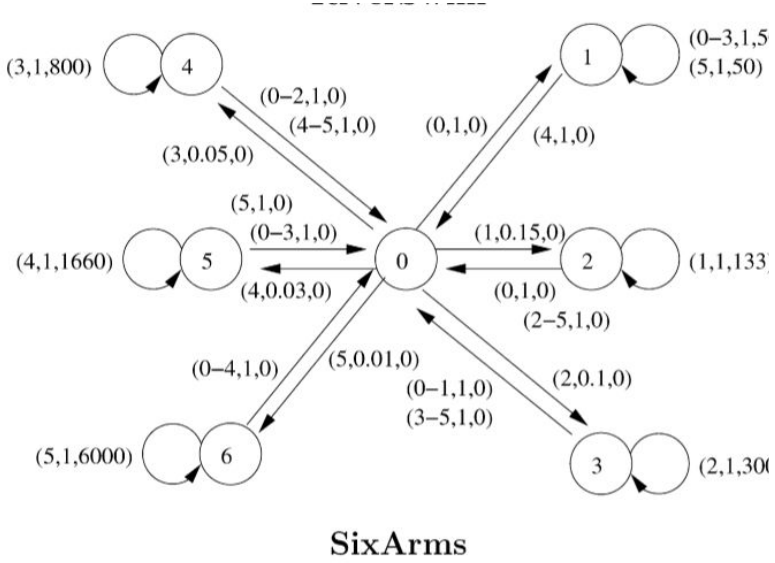
MBIE is PAC & can prove regret bounds (see UCRL2 algorithm)
Empirically can do much better than R-max

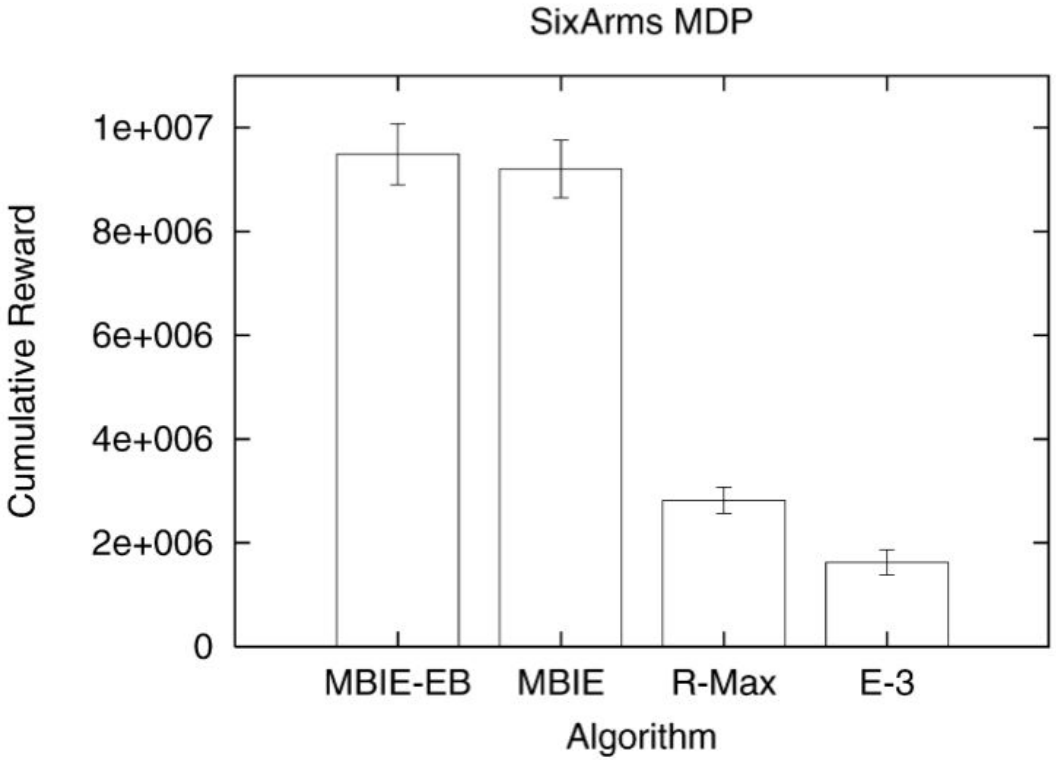




images from Strehl & Littman 2008

MBIE is PAC & can prove regret bounds (see UCRL2 algorithm)
Empirically can do much better than R-max





Intuition? MBIE starts leveraging knowledge immediately.
If things don't look promising, may never learn optimal pi for all states

images from Strehl & Littman 2008

Thompson Sampling: Model-Based RL

- Thompson sampling implements probability matching

$$\pi(s, a \mid h_t) = \mathbb{P}\left[Q^*(s, a) > Q^*(s, a'), \forall a' \neq a \mid h_t\right]$$
$$= \mathbb{E}_{\mathcal{P},\mathcal{R}|h_t}\left[\mathbf{1}(a = \underset{a \in \mathcal{A}}{\text{argmax}}\ Q^*(s, a))\right]$$

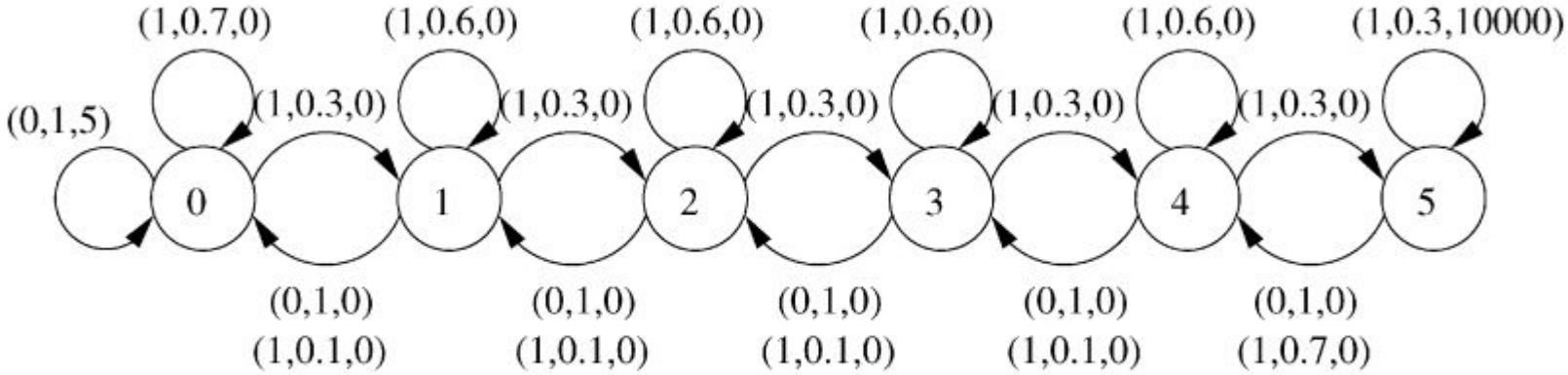
- Use Bayes law to compute posterior distribution $p[\mathcal{P}, \mathcal{R} \mid h_t]$
- Sample an MDP $\mathcal{P}, \mathcal{R}$ from posterior
- Solve MDP using favourite planning algorithm to get $Q^*(s, a)$
- Select optimal action for sample MDP, $a_t = \underset{a \in \mathcal{A}}{\text{argmax}}\ Q^*(s_t, a)$

slide from David Silver

Posterior Sampling for (episodic) RL
Osband, Van Roy & Russo (NIPS 2013)

Expected regret: $\tilde{O}(\tau S\sqrt{AT})$

A = # of actions
S = # of states
T = # number of time steps
τ = # number of time steps per episode (<<T)



RiverSwim

Algorithm	*Random MDP* τ*-episodes*	*Random MDP* ∞*-horizon*	*RiverSwim* τ*-episodes*	*RiverSwim* ∞*-horizon*
PSRL	1.04×10^4	7.30×10^3	6.88×10^1	1.06×10^2
UCRL2	5.92×10^4	1.13×10^5	1.26×10^3	3.64×10^3

PSRL can do much better than optimism under uncertainty based approach

riverswim images from Strehl & Littman 2008; results from Osband, Van Roy & Russo 2013

Key Points

- Understand the definitions of regret and sample complexity
- Know (or be able to derive) bounds & rates
- Know what algorithms yield these and what features of those algorithms make it possible (e.g. optimism under uncertainty, sampling from a posterior)
- Be able to implement these style of algorithms & understand where different approaches may be best empirically