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

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

UCB and TS

- 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

- Thompson sampling: randomized approximation of UCB with “right” confidence sets

- UCB: select most optimistic plausible model

$$\max_{x \in \mathcal{X}} \max_{\tilde{\theta} \in \Theta_t} f_{\tilde{\theta}}(x)$$

- Thompson Sampling: sample from posterior

$$\tilde{\theta} \sim p_{t-1} \qquad \max_{x \in \mathcal{X}} f_{\tilde{\theta}}(x)$$

TS for MDPs

- Prior over rewards and transition probs
- This prior can encode what we believe about the target MDP
- Posterior computed by conditioning on data observed over preceding episodes
- Sample from posterior an MDP, and apply for one episode a policy that is optimal for this sampled MDP
- Computational challenges
    - Computing posterior
    - Sampling from posterior

An “Uninformative” Prior

- A prior for efficient *tabula rasa* RL
- Independent prior for each (t,s,a)
- Rewards
  - Gaussian rewards but with unknown mean and variance
  - Normal-Gamma prior on (mean,variance)
  - 4-parameter distribution
- Transition probabilities
  - Uniform prior over simplex

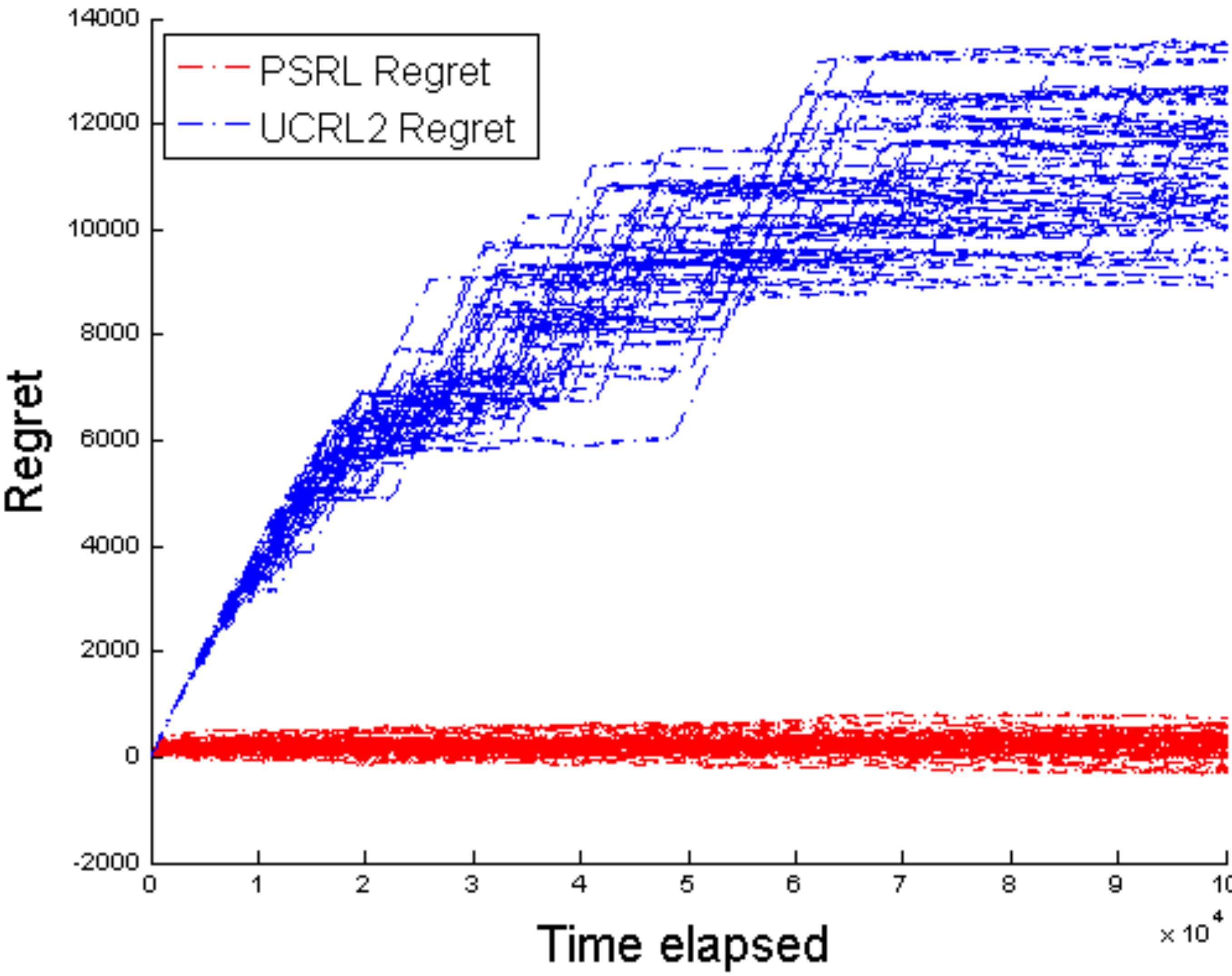
Posterior Computation

- Conjugate priors
- Uniform prior over simplex is a special case of a Dirichlet prior which is efficiently updated given observed state transitions
- Normal-Gaussian prior is efficiently updated given observed rewards

PSRL

- For ℓ = 1, 2, 3,…
  - Compute posterior distribution over (P, R)
  - Sample $(\hat{P}, \hat{R})$ from posterior distribution
  - μ = optimal policy for sampled MDP $(\mathcal{S}, \mathcal{A}, \hat{R}, \hat{P}, \rho, H)$
  - Observe initial state $s_0^\ell \sim \rho(\cdot)$
  - For t = 0, 1, 2,…, *H*-1
    - Apply action $a_t = \mu_t(s_t^\ell)$
    - Observe reward $r_t^\ell \sim R_{t,a}(\cdot|s_t^\ell)$
    - Observe state transition $s_{t+1}^\ell \sim P_{t,a}(\cdot|s_t^\ell)$

UCRL versus PSRL



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

Regret Bounds

- Regret for PSRL

$$\mathbb{E}\left[\mathrm{Regret}(L)\right] = O\left(H\mathcal{S}\sqrt{\mathcal{A}HL\log(\mathcal{S}\mathcal{A}HL)}\right)$$

- Regret for UCRL

$$\mathrm{Regret}(L) = O\left(H\mathcal{S}\sqrt{\mathcal{A}HL\log(HL/\delta)}\right)$$

$$\text{with probability } 1-\delta$$

- What does this imply about learning time?