Theoretical Guarantees in Reinforcement Learning : Regret analysis

Shipra Agrawal

IEOR 8100 (Reinforcement Learning)

Quantifying the performance of RL algorithms

- Computational complexity
  - the amount of per-time-step computation the algorithm uses during learning;
- Space complexity
  - the amount of memory used by the algorithm;
- Learning complexity
  - a measure of how much experience the algorithm needs to learn in a given task.

Learning complexity analysis

- Using a single thread of experience
  - No resets or generative model

Vs.

Generative models (simulator)

- http://hunch.net/~jl/projects/RL/EE.html

Learning complexity analysis

- Optimal learning complexity?
  - "optimally explore" meaning to obtain maximum expected discounted reward $\mathrm{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_t]$ over *a known prior* over the unknown MDPs
  - Tractable in special cases ($\gamma < 1, MAB$, Gittins, 1989).

PAC analysis: single thread of experience

- Bound number of steps on which suboptimal policy is played
- near-optimally on all but a polynomial number of steps (Kakade, 2003; Strehl and Littman, 2008b, Strehl et al 2006, 2009)
- Example: [Strehl et al 2006] (may been improved in recent work)

Theorem 1 *Let M be any MDP and let ϵ and δ be two positive real numbers. If Delayed Q-learning is executed on MDP M, then it will follow an ϵ-optimal policy on all but* $O\left(\frac{SA}{(1-\gamma)^8\epsilon^4}\ln\frac{1}{\delta}\ln\frac{1}{\epsilon(1-\gamma)}\ln\frac{SA}{\delta\epsilon(1-\gamma)}\right)$ *timesteps, with probability at least* $1-\delta$.

Generative model: Sample complexity for ϵ optimal

Some examples:

- An O(SA) analysis of Q-learning.
  - Michael Kearns and Satinder Singh, Finite-Sample Convergence Rates for Q-learning and Indirect Algorithms NIPS 1999.
- An analysis of TD-lambda.
  - Michael Kearns and Satinder Singh, Bias-Variance Error Bounds for Temporal Difference Updates, COLT 2000.

Regret analysis

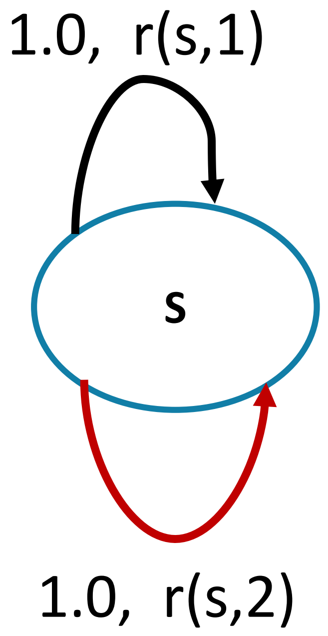
- Bound difference in total reward of algorithm compared to a benchmark policy
  - $Reg(M,T) = T\,\rho^*(s_0) \; - E[\sum_{t=1}^{T} r_t]$
  - $\rho^*(s_0)$ is infinite horizon average reward achieved by the best stationary policy π^*
- Episodic/single thread
  - Episodic: starting state is reset every H steps
  - Single thread: assume Communicating with diameter D
- Worst-case regret bounds; example
  - Theorem [Agrawal, Jia, 2017]: For **any communicating MDP** M with (unknown) diameter D, and for $\mathrm{T} \geq S^5 A$, our algorithm achieves: $\mathrm{Regret}(M,T) \leq \tilde{O}(D\sqrt{SAT})$
- Bayesian regret bounds: (expectation over known prior on MDP M); example
  - Theorem [Osband, Russo, Van Roy, 2013]: For **any known prior distribution** f on MDP M, our algorithm achieves in episodic setting: $\mathrm{E}_{\mathrm{M}\sim f}[\mathrm{Regret}(M,T)] \leq \tilde{O}(HS\sqrt{AT})$

Lecture overview

- Exploration-exploitation and regret minimization for multi-armed bandit
  - UCB
  - Thompson Sampling
- Regret minimization for RL (tabular)
  - UCRL
  - Posterior sampling based algorithms

Multi-armed bandit == single state MDP

Two armed bandit



$P(s,a,s),\ r(s,a)$

Stochastic multi-armed bandit problem

- Online decisions
  - At every time step $t = 1, \dots, T$, pull one arm out of N arms
- Stochastic feedback
  - For each arm i, reward is generated i.i.d. from a **fixed but unknown distribution** ν_i support [0,1], mean μ_i
- Bandit feedback
  - Only the reward of the pulled arm can be observed

The multi-armed bandit problem (Thompson 1933; Robbins 1952)

Multiple rigged slot machines in a casino.

Which one to put money on?

- Try each one out

Arm == actions

WHEN TO STOP TRYING (EXPLORATION) AND START PLAYING (EXPLOITATION)?

Regret minimization problem

- Minimize expected regret in time T
  - $Regret(T) = T\mu^* - E[\sum_{t=1}^{T} r_t]$ where $\mu^* = \max_j \mu_j$
- Equivalent formulation
  - Expected regret for playing arm $\mathrm{I_t} = i$ at time t: $\Delta_i = \mu^* - \mu_i$
  - $n_{i,T}$ be the number times arm i is played till time T
  - $Regret(T) = \sum_t (\mu^* - \mu_{I_t}) = \sum_{i:\mu_j \neq \mu^*} \Delta_i \, E[n_{i,T}]$
- If we can bound $n_{i,T}$ by $\frac{C \ \log(T)}{\Delta_i^2}$, then regret bound: $\sum_{i:\mu_i \neq \mu^*} \frac{C \ \log(T)}{\Delta_i}$
  - Problem-dependent bound, close to lower bound [Lai and Robbins 1985]
- For problem independent bound: $\mathrm{C}\sqrt{NTlog(T)}$
  - Separately bound total regret for playing an arm with $\Delta_i \leq \sqrt{\frac{N\log(\mathrm{T})}{T}}$ and $\Delta_i > \sqrt{\frac{N\log(\mathrm{T})}{T}}$
  - Lower bound $\Omega\left(\sqrt{NT}\right)$,

The need for exploration

- Two arms **black** and **red**
  - Random rewards with unknown mean $\boldsymbol{\mu_1 = 1.1}, \boldsymbol{\mu_2 = 1}$
  - Optimal expected reward in T time steps is $1.1 \times T$

- Exploit only strategy: use the current best estimate (MLE/empirical mean) of unknown mean to pick arms
- Initial few trials can mislead into playing red action forever

  1.1, 1, 0.2,

  1, 1, 1, 1, 1, 1, 1, ……

- Expected regret in T steps is close to $0.1 \times T$

Exploration-Exploitation tradeoff

- Exploitation: play the empirical mean reward maximizer
- Exploration: play less explored actions to ensure empirical estimates converge

Upper confidence bound algorithms for MAB

Optimism under face of uncertainty

- Empirical mean at time t for arm i

$$\hat{\mu}_{i,t} = \frac{\sum_{s=1:\, I_s=i}^{t} r_s}{n_{i,t}}$$

- Upper confidence bound

$$\mathrm{UCB}_{i,t} := \hat{\mu}_{i,t} + 2\sqrt{\frac{\ln t}{n_{i,t}}}$$

Algorithm 1: UCB algorithm for the stochastic N-armed bandit problem

```
foreach t = 1, ..., N do
    Play arm t
end
foreach t = N+1, N+2 ..., T do
    Play arm I_t = argmax_{i∈{1...N}} UCB_{i,t-1}.
    Observe r_t, compute UCB_{i,t}
end
```

Regret analysis

- Using Azuma-Hoeffding (since $E[r_t|I_t = i\,] = \mu_i$), with probability $1 - \frac{2}{t^2}$,

$$|\hat{\mu}_{i,t} - \mu_i| < \sqrt{\frac{4\ln t}{n_{i,t}}}$$

- Two implications: for every arm i
  - $UCB_{i,t} > \mu_i$ with probability $1 - \frac{2}{t^2}$
  - If $n_{i,t} > \frac{16\ln(T)}{\Delta_i^2}$, $\hat{\mu}_{i,t} < \mu_i + \frac{\Delta_i}{2}$

Recall: $\mathrm{UCB}_{i,t} := \hat{\mu}_{i,t} + 2\sqrt{\frac{\ln t}{n_{i,t}}}$

- Each suboptimal arm is played at most $\frac{16\ln(T)}{\Delta_i^2}$ times, since after that many plays:

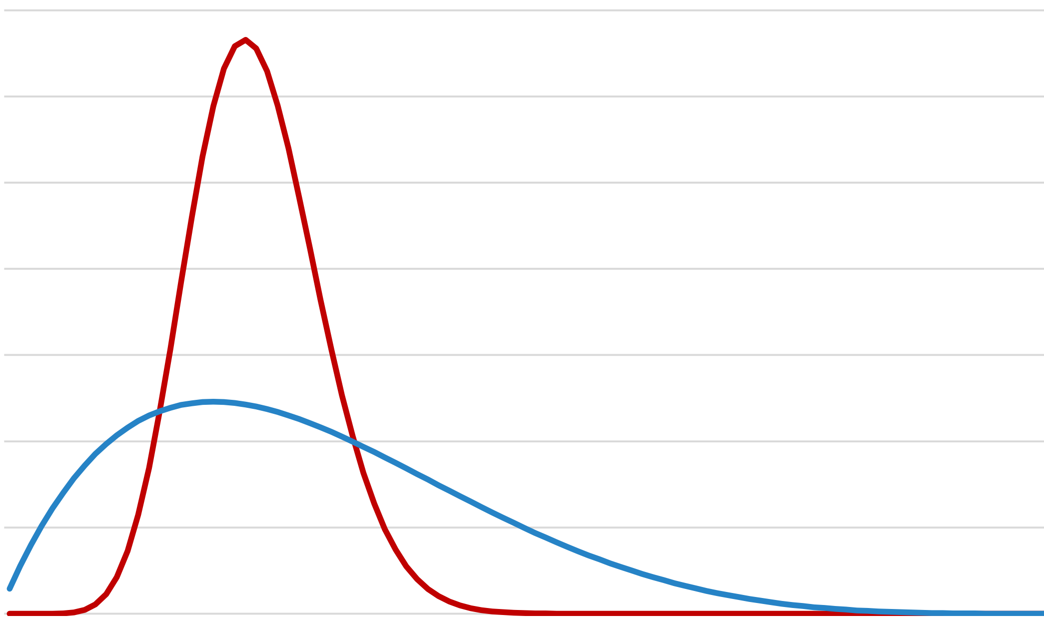
$$\begin{aligned}\mathrm{UCB}_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{\ln t}{n_{i,t}}} &\leq \hat{\mu}_{i,t} + \frac{\Delta_i}{2} \\ &< \left(\mu_i + \frac{\Delta_i}{2}\right) + \frac{\Delta_i}{2} \\ &= \mu^* \\ &< \mathrm{UCB}_{i^*,t}\end{aligned}$$

Thompson Sampling [Thompson, 1933]

- Natural and Efficient heuristic
- Maintain belief about parameters (e.g., mean reward) of each arm
- Observe feedback, update belief of pulled arm *i* in Bayesian manner
- Pull arm with posterior probability of being best arm
  - NOT same as choosing the arm that is most likely to be best

Posterior Sampling: main idea [Thompson 1933]

- Maintain Bayesian posteriors for unknown parameters
- With more trials posteriors concentrate on the true parameters
  - Mode captures MLE: enables exploitation
- Less trials means more uncertainty in estimates
  - Spread/variance captures uncertainty: enables exploration
- A sample from the posterior is used as an estimate for unknown parameters to make decisions

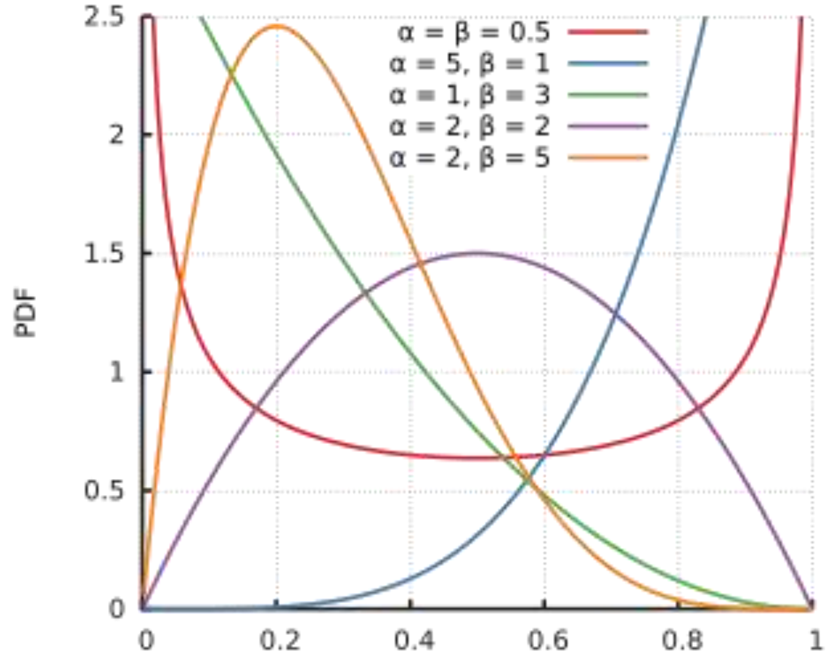


Bernoulli rewards, Beta priors

Uniform distribution $Beta(1,1)$

$Beta(\alpha, \beta)$ prior $\Rightarrow$ Posterior

- $Beta(\alpha + 1, \beta)$ if you observe 1
- $Beta(\alpha, \beta + 1)$ if you observe 0



Start with $Beta(\alpha_0, \beta_0)$ prior belief for every arm

In round t,

- For every arm i, sample $\theta_{i,t}$ independently from posterior $Beta\left(\alpha_0 + S_{i,t}, \beta_0 + F_{i,t}\right)$
- Play arm $i_t = \max_i \theta_{i,t}$
- Observe reward and update the Beta posterior for arm i_t

Bayesian regret bounds for Thompson Sampling

- Theorem [Russo and Van Roy 2014]: Given any prior f over MAB instance M described by reward distributions with $[0,1]$ bounded support (e.g., prior distribution over $\mu_1, \mu_2, \dots, \mu_N$ in case of Bernoulli rewards)

$$\text{BayesianRegret}(T) = E_{M\sim f}[\text{Regret}(T, M)] \leq O(\sqrt{NT \ln T})$$

Note;

- $\text{Regret}(T, M)$ notation is used to explicitly indicate dependence on regret on the MAB instance M.
- In comparison, worst case regret,

$$\text{Regret}(T) = max_M[\text{Regret}(T, M)]$$

Worst-case regret bounds

[A. and Goyal COLT 2012, AISTATS 2013, JACM 2017, Kaufmann et al. 2013]

Instance-dependent bounds for {0,1} rewards

- Regret$(T) \leq \boldsymbol{ln}(\boldsymbol{T})(1+\epsilon)\sum_i \frac{\Delta_i}{KL(\mu^*||\mu_i)} + O(\frac{N}{\epsilon^2})$
  - Matches *asymptotic instance wise lower bound* [Lai Robbins 1985]
  - UCB algorithm achieves this only after careful tuning [Kaufmann et al. 2012]

Arbitrary bounded [0,1] rewards (using Beta and Gaussian priors)

- Regret$(T) \leq O(\boldsymbol{ln}(\boldsymbol{T})\sum_i \frac{1}{\Delta_i})$
  - Matches the best available for UCB for general reward distributions

Instance-independent bounds (Beta and Gaussian priors)

- Regret$(T) \leq O(\sqrt{NT\ln T})$
- Prior and likelihood mismatch allowed!

Why does it work? Two arms example

- Two arms, $\mu_1 \geq \mu_2$, $\Delta = \mu_1 - \mu_2$
- Every time arm 2 is pulled, Δ regret
- Bound the number of pulls of arm 2 by $\frac{\log(\mathrm{T})}{\Delta^2}$ to get $\frac{\log(\mathrm{T})}{\Delta}$ regret bound
- How many pulls of arm 2 are actually needed?

Easy situation

After n=O($\frac{\log(T)}{\Delta^2}$) pulls of arm 2 **and arm 1**

- Empirical means are well separated

  Error $|\widehat{\mu_i} - \mu_i| \quad \leq \sqrt{\frac{\log(T)}{n}} \leq \frac{\Delta}{4}$ whp

  (Using Azuma Hoeffding inequality)

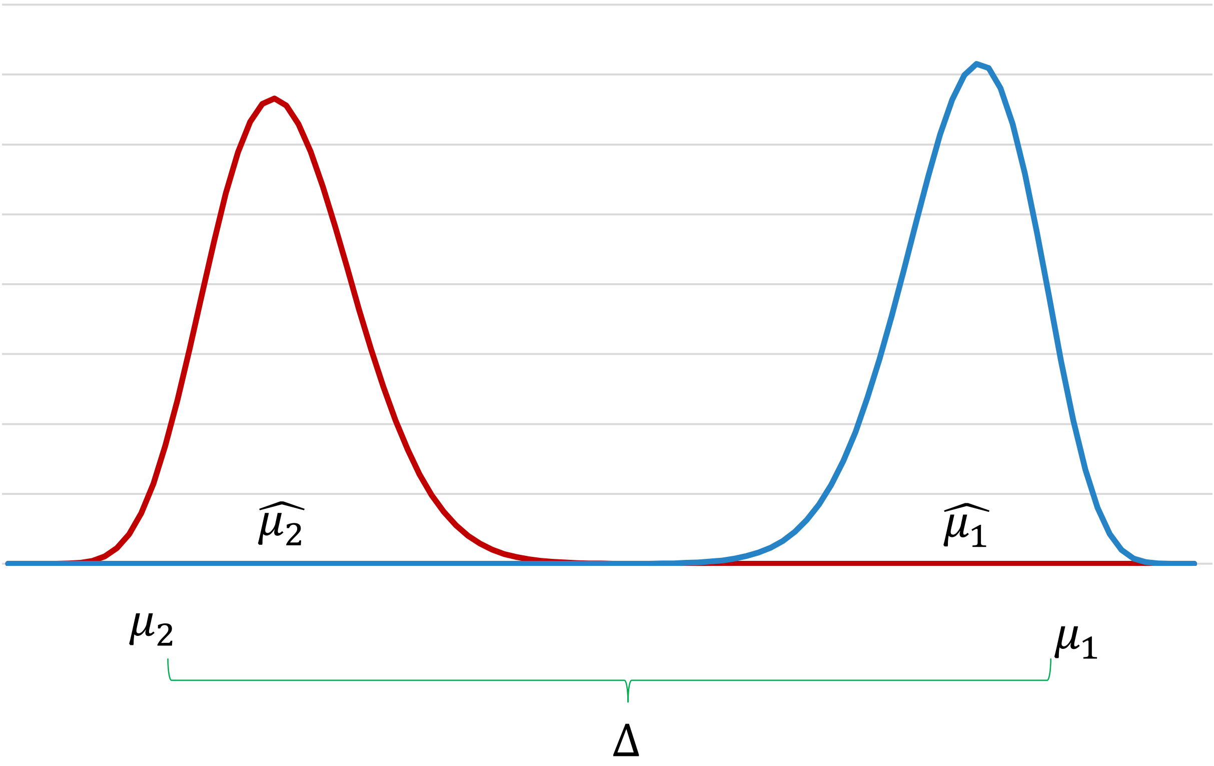
- Beta Posteriors are well separated

  Mean = $\frac{\alpha_i}{\alpha_i+\beta_i} = \widehat{\mu_i}$

  standard deviation $\simeq \frac{1}{\sqrt{\alpha+\beta}} = \frac{1}{\sqrt{\mathrm{n}}} \leq \frac{\Delta}{4}$

The two arms can be distinguished!

No more arm 2 pulls.

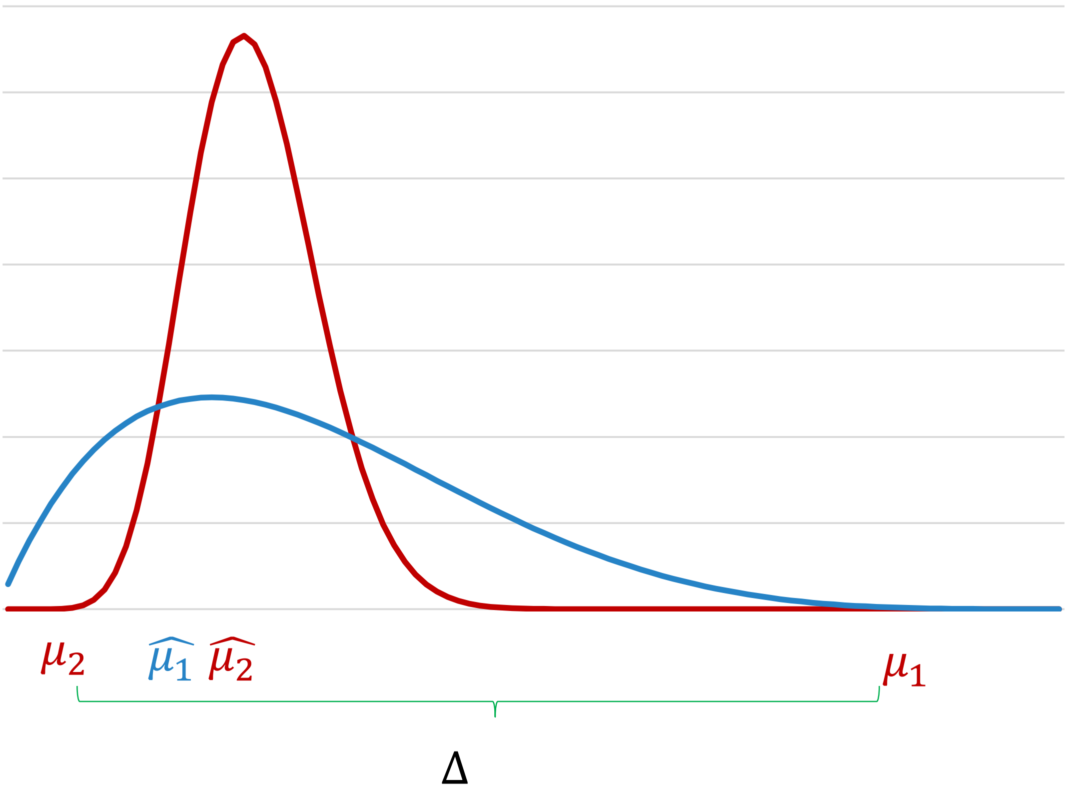


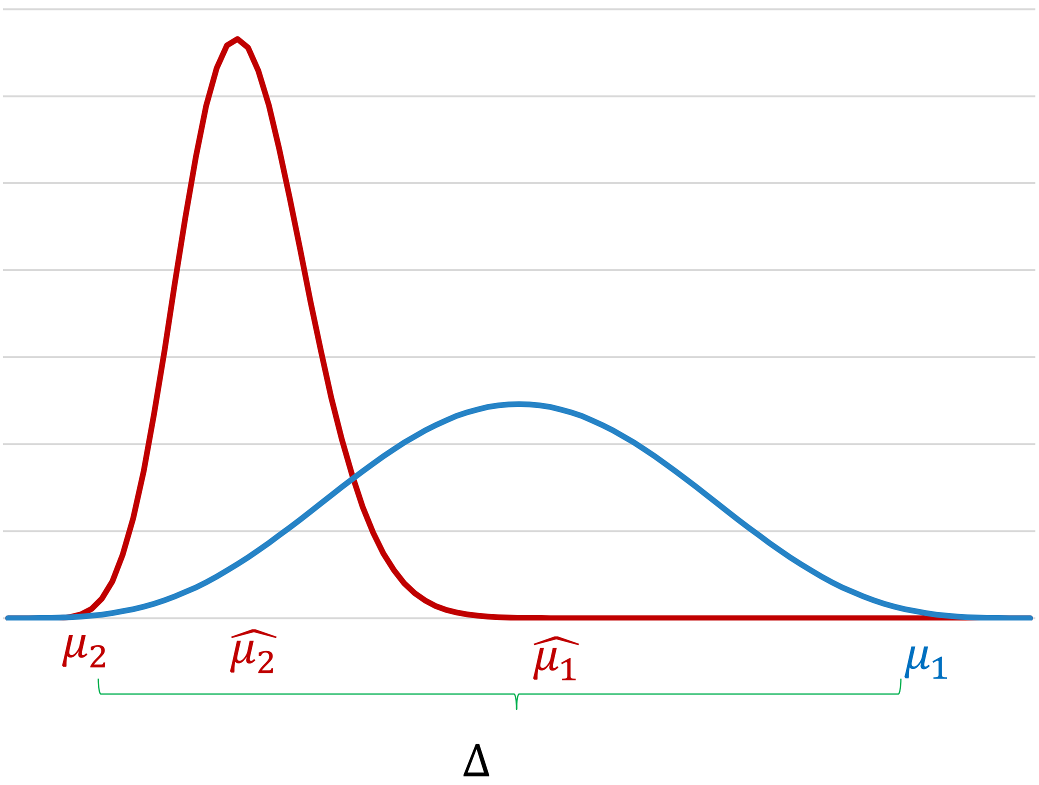
Easy situation

- If arm 2 is pulled less than n=$O(\frac{\log(T)}{\Delta^2})$ times?
  - Regret is at most $n\Delta = \frac{\log(T)}{\Delta}$

Difficult situation

- At least $\frac{\log(T)}{\Delta^2}$ pulls of arm 2, but few pulls of arm 1





Main insight

- Arm 1 will be played roughly every constant number of steps in this situation
- It will take at most *constant* $\times \frac{\log T}{\Delta^2}$ steps (extra pulls of arm 2) to get out of this situation
- Total number of pulls of arm 2 is at most $O(\frac{\log T}{\Delta^2})$

- Summary: variance of posterior enables exploration
- Optimal bounds (and for multiple arms) require more careful use of posterior structure

Next…

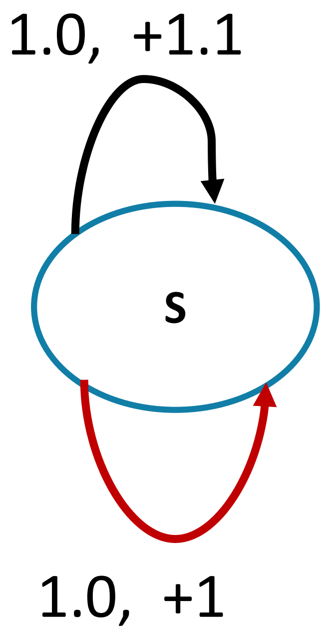
Using UCB and TS ideas for exploration-exploitation in RL

The need for exploration

- Single state MDP
  - Solution concept: optimal action
  - **Multi-armed bandit problem**
- Uncertainty in rewards
  - Random rewards with unknown mean μ_1, μ_2
- Exploit only: use the current best estimate (MLE/empirical mean) of unknown mean to pick arms
- Initial few trials can mislead into playing red action forever

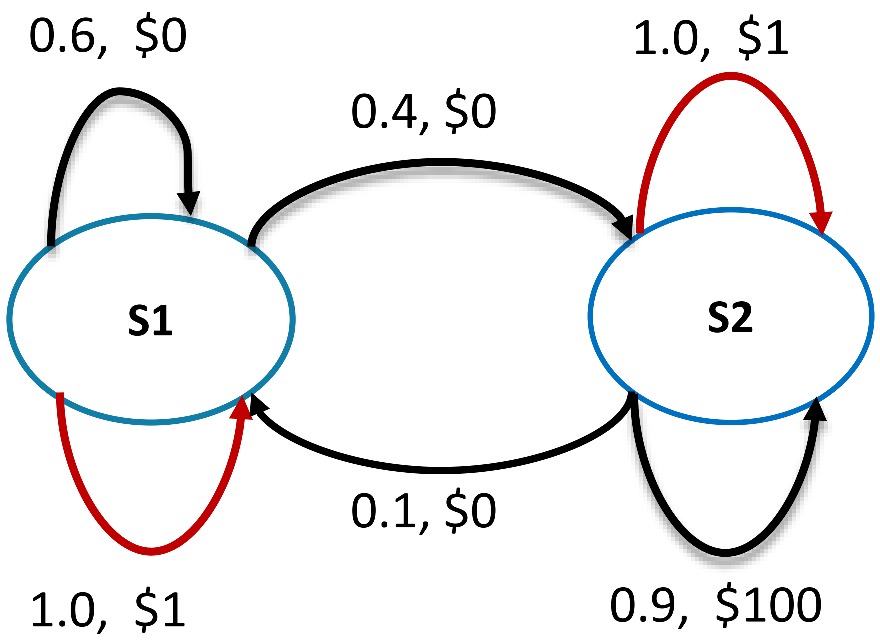
1.1, 1, 0.2,

1, 1, 1, 1, 1, 1, 1, ……



$P(s, a, s),\ r(s, a)$

The need for exploration



- Uncertainty in rewards, state transitions
- Unknown reward distribution, transition probabilities
- Exploration-exploitation:
  - Explore actions/states/policies, learn reward distributions and transition model
  - Exploit the (seemingly) best policy

Summary of recent work

Upper confidence bound based algorithms [Jaksch, Ortner, Auer, 2010] [Bartlett, Tewari, 2012]

- Worst-case regret bound $\tilde{O}(DS\sqrt{AT})$ for communicating MDP
- Lower bound $\Omega\left(\sqrt{DSAT}\right)$

Optimistic Posterior Sampling [A. Jia 2017]

- Worst-case regret bound $\tilde{O}(D\sqrt{SAT})$ for communicating MDP of diameter D
- Improvement by a factor of $\sqrt{S}$

Optimistic Value iteration [Azar, Osband, Munos, 2017]

- Worst-case reget bound $\tilde{O}(\sqrt{HSAT})$ in episodic setting

Posterior sampling known prior setting [Osband, Russo, and Van Roy 2013, Osband and Van Roy, 2016, 2017]b

- Bayesian regret bound of $\tilde{O}(H\sqrt{SAT})$ in episodic setting, length H episodes

Next...

- **UCRL: Upper confidence bound based algorithm for RL**
- Posterior sampling based algorithm for RL
  - Main result
  - Proof techniques

UCRL algorithm [Jacksch, Ortner, Auer 2002]

- Similar principles as UCB

This is a ***Model-based approach***

- Maintain an estimate of model $\hat{P}, \hat{R}$
- Occassionally solve the MDP $(S, A, \hat{P}, \hat{R}, s_1)$ to find a policy
- Run this policy for some time to get samples, and update model estimate

Compare to ``model-free" approach or direct learning approach like Q-learning

- Directly update Q-values or value function or policy using samples.

UCRL algorithm

- Proceed in epochs, an epoch ends when the number of visits of *some* state-action pair doubles.

In the beginning of every epoch k

- Use samples to compute an optimistic MDP (S, A, $\tilde{R}, \tilde{P}, s_1$)
  - MDP with value greater than true MDP (Upper bound!!)
- Solve the optimistic MDP to find optimal policy $\tilde{\pi}$

Execute $\tilde{\pi}$ in epoch k

- observe samples s_t, r_t, s_{t+1}

Go to next epoch If visits of *some* state-action pair doubles

- If $n_k(s,a) \geq 2\, n_{k-1}(s,a)$ for some s, a

UCRL algorithm (computing optimistic MDP)

In the beginning of every epoch k

- For every s, a, compute **empirical** model estimate
  - let $\mathrm{n_k}(s, a)$ be the number of times s, a was visited before this epoch,
  - let $\mathrm{n_k}(s, a, s')$ be the number of transition to s'
  - Set $\hat{R}(s, a)$ as average reward over these $\mathrm{n_k}(s, a)$ steps
  - Set $\hat{P}(s, a, s')$ as $\frac{\mathrm{n_k}(s,a,s\prime)}{\mathrm{n_k}(s,a)}$
- Compute **optimistic** model estimate
  - Use Chernoff bounds to define confidence region around $\hat{R}, \hat{P}$
    - $|\hat{P}(s, a, s') - P(s, a, s')| \leq \frac{\log(t)}{\sqrt{n_k(s,a)}}$ with probability $1 - \frac{1}{t^2}$
    - True R, P lies in this region
  - Find the best combination , $\tilde{R}, \tilde{P}$ in this region
    - MDP (S, A, $\tilde{R}, \tilde{P}$, s_1) with maximum value
    - Will have value more than the true MDP

Main result

- Recall regret:

$$\text{Regret}(M, T) = T\,\rho^* - \sum_{t=1}^{T} r(s_t, a_t)$$

- Theorem: For any **communicating** MDP M with (unknown) diameter D, with high probability:

$$\text{Regret}(M, T) \leq \tilde{O}(DS\sqrt{AT})$$

- $\tilde{O}(\cdot)$ notation hides logarithmic factors in S,A,T beyond constants.

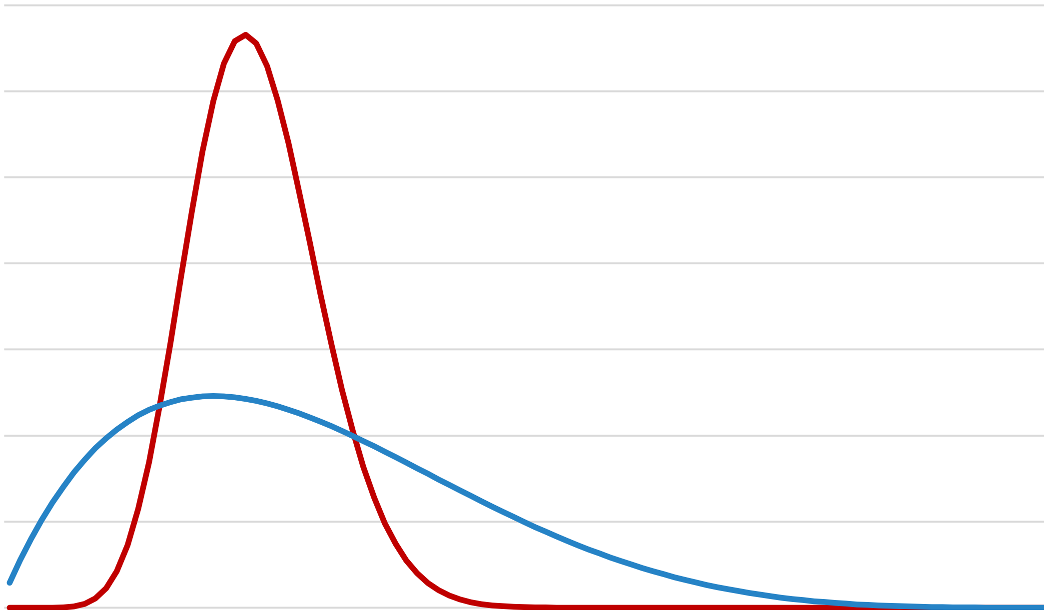
Next…

- Our setting, regret definition

- **Posterior sampling algorithm** for MDPs
  - Main result
  - Proof techniques

Posterior Sampling: main idea [Thompson 1933]

- Maintain Bayesian posteriors for unknown parameters
- With more trials posteriors concentrate on the true parameters
  - Mode captures MLE: enables exploitation
- Less trials means more uncertainty in estimates
  - Spread/variance captures uncertainty: enables exploration
- A sample from the posterior is used as an estimate for unknown parameters to make decisions



Posterior Sampling : Bayesian posteriors

- Assume for simplicity: Known reward distribution
- Needs to learn the unknown transition probability vector $P_{s,a} = (P_{s,a}(1), \ldots, P_{s,a}(S))$ for all s, a
- In any state $s_t = s, a_t = a$, observes new state s_{t+1}
  - outcome of a Multivariate Bernoulli trial with probability vector $P_{s,a}$

Posterior Sampling with Dirichlet priors

- Given prior $\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_S)$ on $P_{s,a}$
- After a Multinoulli trial with outcome (new state) i, Bayesian posterior on $P_{s,a}$

$$\text{Dirichlet}(\alpha_1, \alpha_2, \dots, \alpha_i + 1, \dots, \alpha_S)$$

- After $n_{s,a} = \alpha_1 + \cdots + \alpha_S$ observations for a state-action pair s, a
  - Posterior mean vector is empirical mean

$$\hat{P}_{s,a}(i) = \frac{\alpha_i}{\alpha_1 + \cdots + \alpha_S} = \frac{\alpha_i}{n_{s,a}}$$

  - variance bounded by $\frac{1}{n_{s,a}}$
- With more trials of s, a, the posterior mean concentrates around true mean

Posterior Sampling for RL (Thompson Sampling)

Learning

- Maintain a Dirichlet posterior for $P_{s,a}$ for every s, a
  - After round t, on observing outcome s_{t+1}, update for state s_t and action a_t

To decide action

- Sample a $\tilde{P}_{s,a}$ for every s, a
- Compute the optimal policy $\tilde{\boldsymbol{\pi}}$ for sample MDP $(S, A,\ \tilde{P},\ r, s_0)$
- Choose $a_t = \tilde{\pi}(s_t)$

Exploration-exploitation

- Exploitation: With more observations Dirichlet posterior concentrates, $\tilde{P}_{s,a}$ approaches empirical mean $\hat{P}_{s,a}$
- Exploration: Anti-concentration of Dirichlet ensures exploration for states/actions/policies less explored

Optimistic Posterior Sampling [A., Jia, NIPS 2017]

- Proceed in epochs, an epoch ends when the number of visits $N_{s,a}$ of any state-action pair doubles.

In every epoch

- For every s, a, generate **multiple** $\psi = \tilde{O}(S)$ independent samples from a Dirichlet posterior for $\mathrm{P}_{s,a}$
- Form **extended** sample MDP $(S, \psi A,\ \tilde{P},\ r, s_0)$
- Find optimal policy $\tilde{\pi}$ and use through the epoch

Further, initial exploration:

- For s, a with very small $N_{s,a} < \sqrt{\frac{TS}{A}}$, use a simple optimistic sampling, that provides extra exploration

Main result [A., Jia NIPS 2017]

- An algorithm based on posterior sampling with high probability near-optimal worst-case regret upper bound
- Recall regret:

$$\text{Regret}(M, T) = T\,\rho^* - \sum_{t=1}^{T} r(s_t, a_t)$$

- Theorem: For any **communicating** MDP M with (unknown) diameter D, and for $\mathrm{T} \geq S^5 A$, with high probability:

$$\text{Regret}(M, T) \leq \tilde{O}(D\sqrt{SAT})$$

- Improvement of $\sqrt{S}$ factor above UCB based algorithm