Thompson Sampling based methods for reinforcement learning

Shipra Agrawal

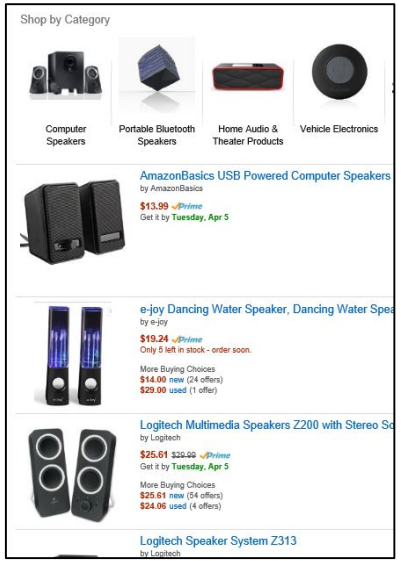
Industrial Engineering and Operations Research

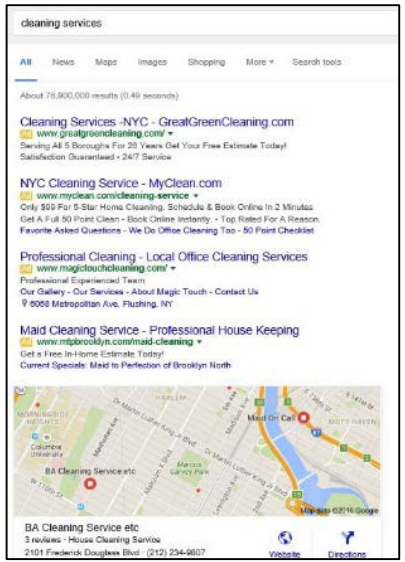
Columbia University

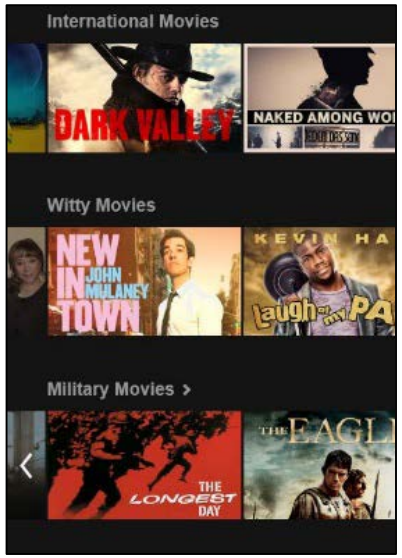
Learning from sequential interactions

Tradeoff between

- learning and optimization
- information and rewards
- Exploration and exploitation







Robotic Manipulation and Mobility

Thompson Sampling: a general Bayesian principle for managing exploration-exploitation tradeoff

Reinforcement learning

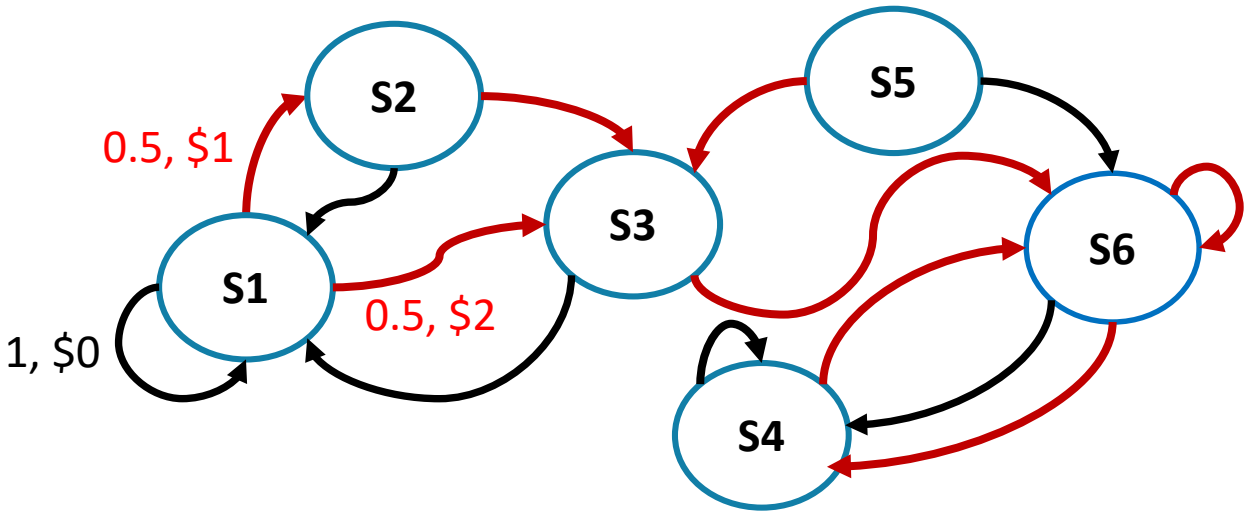
- Markov assumption: Response to an action depends on history only through current state
- Sequential rounds $t = 1, \ldots, T$
  - Observe current state of the system
  - Take an action
  - Observe reward and new state
- Solution concept: policy
  - Mapping from state to action
- Goal: Learn the model while optimizing aggregate reward

The reinforcement learning problem

System dynamics given by an MDP (S, A, P, r, s_0)

Rounds $t = 1, \ldots, T$. In round t,

- observe state s_t, take action a_t,
- observe reward $r_t \in [0,1]$, $\mathrm{E}[r_t] = \boldsymbol{r_{s_t,a_t}}$
- observe the transition to next state s_{t+1} with probability $\mathbf{P}_{s_t,a_t}(\boldsymbol{s_{t+1}})$



The reinforcement learning problem

Solution concept: optimal policy

- Which action to take in which state
- $\pi_t : S \to A$, Action $\pi_t(s)$ in state s

Goal: maximize total reward $\sum_{t=1}^{T} r(s_t, a_t)$

- Learn the MDP model parameters ($\boldsymbol{r}, \boldsymbol{P}$) **from observations** while maximizing reward

Goal: minimize regret

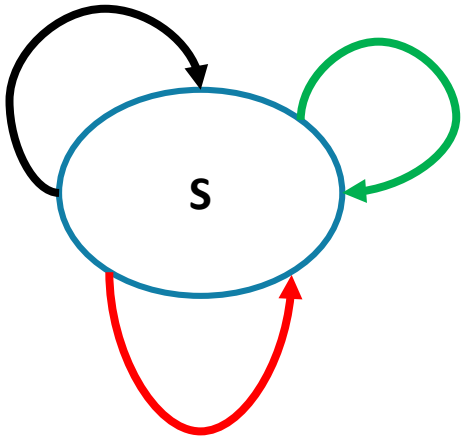
$$\text{Regret}(M, T) = \sum_{t=1}^{T} r(s_t^*, a_t^*) - \sum_{t=1}^{T} r(s_t, a_t)$$

$(s_t^*, a_t^*, t = 1, \ldots T)$ is the trajectory of best single stationary policy π^*

Multi-Armed Bandit (MAB) is a special case

- Single state MDP
- Uncertainty only in rewards : mean $\mathrm{r}(\mathrm{s},\mathrm{a}) = \mu_a$ for action a
- Solution concept: optimal action

$$\mathrm{Regret}(T) = T\,\mu_{a^*} - \sum_{t=1}^{T} \mu_{a_t}$$

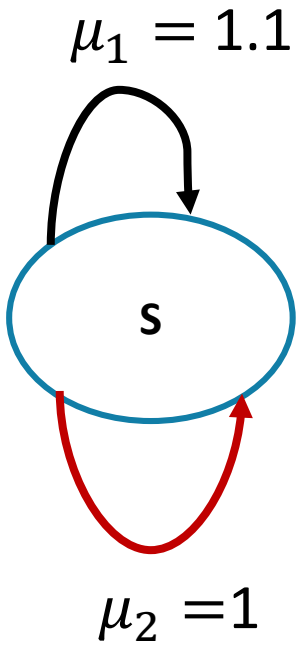


= MAB problem

Actions are called arms

The need for exploration

- Two actions **black** and **red**
  - Unknown mean rewards $\boldsymbol{\mu_1 = 1.1}, \boldsymbol{\mu_2 = 1}$
  - 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

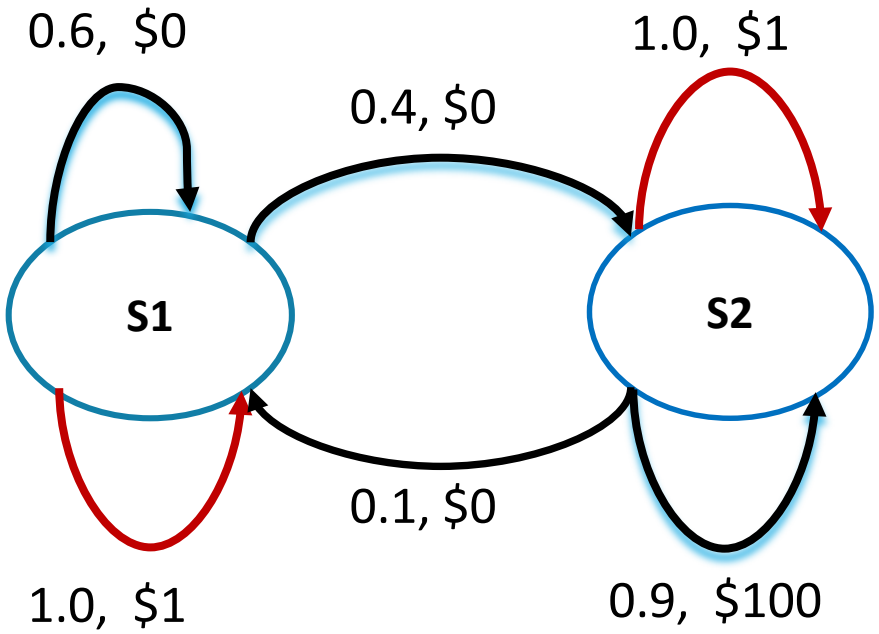
  1.1, 1, 0.2,

  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

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

Thompson Sampling [Thompson, 1933]

aka posterior sampling

- Natural and Efficient Bayesian principle for managing exploration-exploitation in sequential decisions
- Maintain belief distribution(s) about the unknown model parameters
- On taking a decision
  - observe feedback, update belief (posterior) in Bayesian manner
- Take the decision with its posterior probability of being the best decision
  - NOT same as choosing the decision with highest posterior probability of being the best

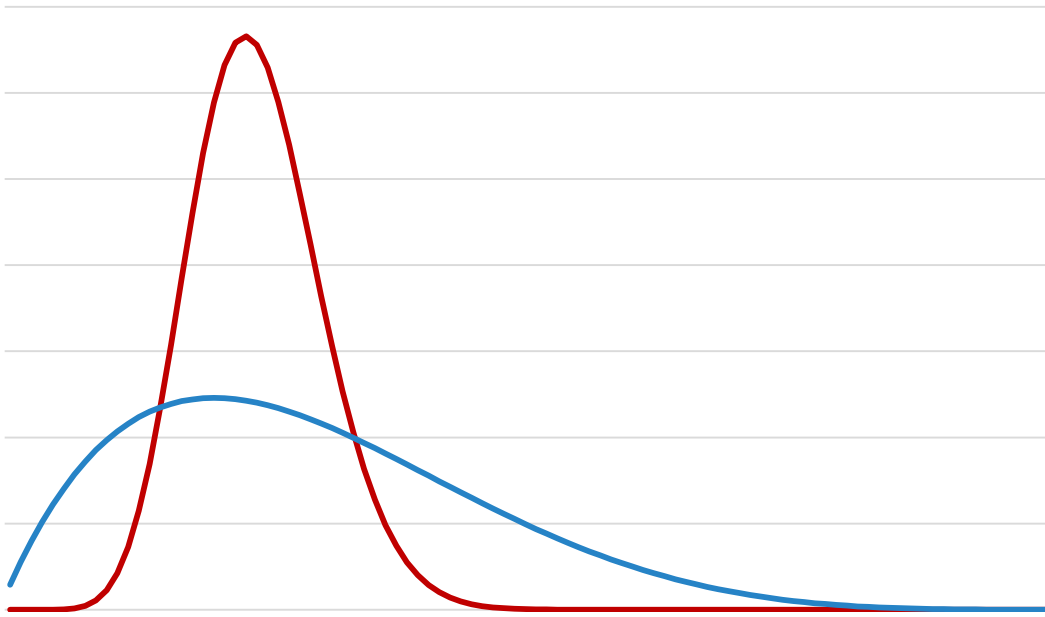
Thompson Sampling for MAB

- Unknown parameters $(\mu_1, \ldots, \mu_N)$
- Maintain N belief distributions: one for mean μ_a of each arm a
- On pulling an arm a
  - Observe reward, update posterior distribution for arm a
- At any time t, given the posterior distributions, pull the arm with its posterior probability of being best arm
  - Sample a parameter θ_a (independently) from each posterior distribution.
  - Pull $a_t = \operatorname{argmax}_a \ \theta_a$

Later: algorithm description for specific forms of distributions, and extension to general RL

Thompson Sampling: exploration-exploitation

- 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**

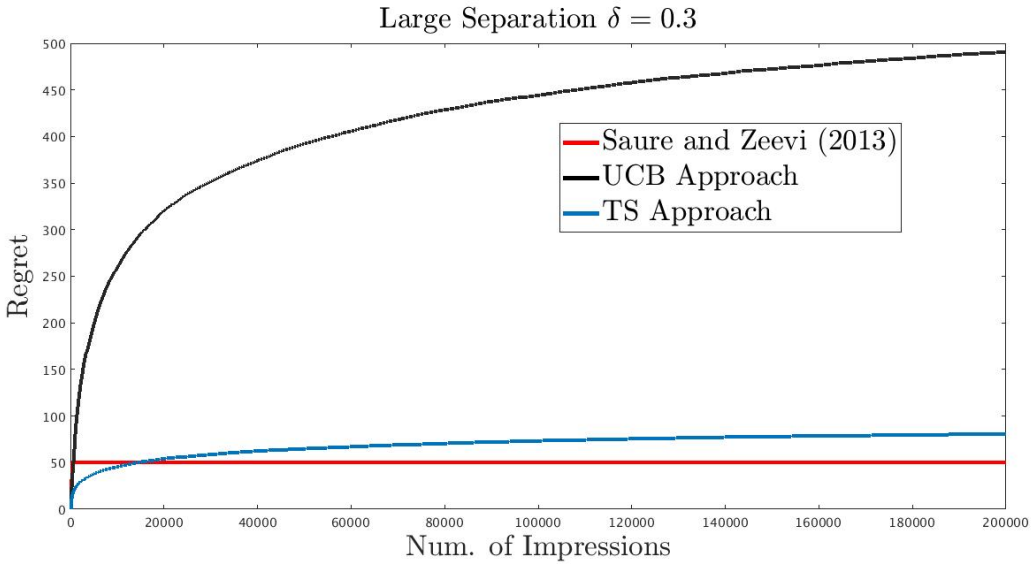


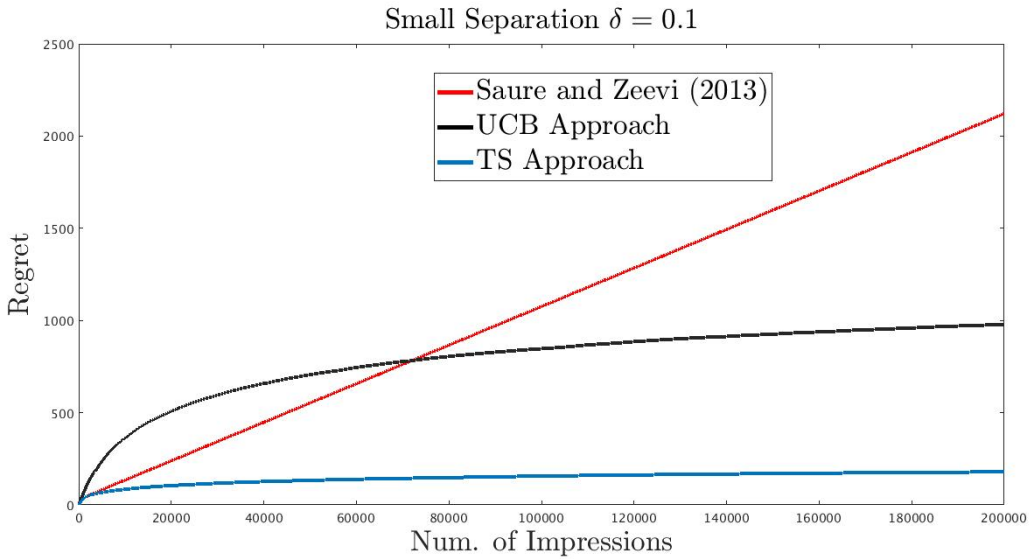
History

- 1933: The general principle was proposed [Thompson 1933]
  - Natural and Easy to implement: Posterior updates are simple and efficient
- 1933-2002: Rediscovered numerous times independently in the context of reinforcement learning [Wyatt 1997; Ortega & Braun 2010; Strens 2000]
- 2002-2011: UCB algorithm dominated the theoretical results on MAB, starting with [Auer 2002]
- 2011-2012: Promising empirical performance [Chapelle, Li, NeurIPS 2011]

Thompson Sampling Empirical performance

[Avadhanula, A., Goyal, Zeevi, COLT 2017]





History

- 2011-2012: Promising empirical performance [Chapelle, Li, NeurIPS 2011]
- However Little was known about theoretical performance
  - Convergence $\frac{Regret(T)}{T} \rightarrow 0$ as $T \rightarrow \infty$ [Granmo, 2009][May et al. 2011]
- 2012: First regret bounds (logarithmic) in [A. and Goyal COLT 2012]
  - Optimal regret bounds [Kaufmann et al. ALT 2012][A. and Goyal AISTATS 2013]
- Last 5 years
  - Regret bounds for many extensions, including RL
    - Still remains harder to analyze than UCB
  - Proliferation in industry: Known implementations in Twitter, Amazon, Microsoft, Google, LinkedIn, Netflix, ….

Outline of the remaining talk

1. Thompson Sampling for MAB
   - Algorithm design using Beta and Gaussian priors
   - Regret bounds
   - Proof for two-armed bandit case
   - Proof overview for N-armed bandits
2. Thompson Sampling for (tabular) Reinforcement Learning
   - Algorithm design using Dirichlet priors
   - Regret bounds
   - Proof techniques

Part I: Thompson Sampling for MAB

Stochastic Multi-armed Bandit problem

- Online decisions
  - At every time step $t = 1, \ldots, T$, pull one arm out of N arms
- Stochastic feedback
  - For each arm a, reward is generated i.i.d. from an arbitrary fixed but unknown distribution support [0,1], mean μ_a
- Bound regret compared to the best arm

$$\mathrm{Reg(T)} = T\mu^* - \sum_{t=1}^{T} \mu_{a_t} \quad \text{where } \mu^* = \max_j \mu_j$$

Thompson Sampling [Thompson, 1933]

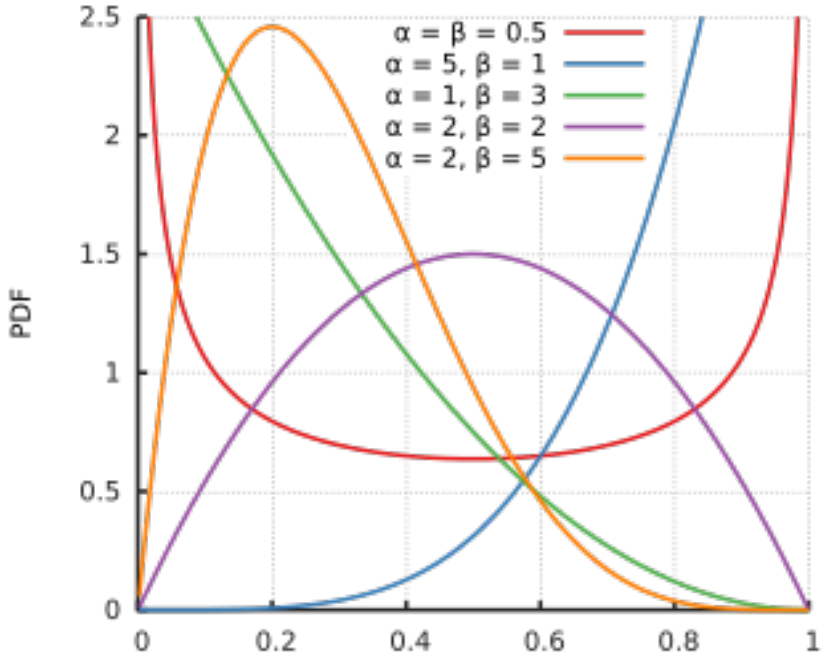
- Maintain posterior belief about mean reward parameter (μ_a) for each arm a
- Pull arm with posterior probability of being best arm
  - Sample θ_a from posterior for each arm a
  - Pull $a_t = \arg\max_a \theta_a$
- Posterior update
  - Observe reward for arm a_t update posterior

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 $Beta(1,1)$ prior belief for every arm

In round t,

- For every arm i, sample $\theta_{i,t}$ independently from posterior $Beta(S_{i,t} + 1, F_{i,t} + 1)$
- Play arm $i_t = \max_i \theta_{i,t}$
- Observe reward and update the Beta posterior for arm i_t

Arbitrary reward distribution mean μ, Gaussian prior

Standard normal prior $N(0,1)$

Gaussian likelihood $N(\mu, 1)$ of reward

Posterior after n independent observations: $N\left(\hat{\mu}, \frac{1}{n+1}\right)$

- $\hat{\mu}$ is the empirical mean

Start with $N(0,1)$ prior belief for every arm

In round t,

- For every arm i, sample $\theta_{i,t}$ independently from posterior $\mathrm{N}\left(\hat{\mu}_i, \frac{1}{n_i+1}\right)$
- Play arm $i_t = \max_i \theta_{i,t}$
- Observe reward and update empirical mean $\hat{\mu}_i$ and number of plays n_i for arm i_t

MAB: worst-case regret bounds
[A. and Goyal COLT 2012, AISTATS 2013, Journal of ACM 2017]

Optimal instance-dependent bounds for Bernoulli rewards

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

Arbitrary bounded reward distribution (Beta and Gaussian priors)

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

Instance-independent bounds (Beta and Gaussian priors)

- $\text{Regret}(T) \leq O(\sqrt{NT\ln T})$
  - Lower bound $\Omega(\sqrt{NT})$
- Prior mismatch (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, total regret $= \Delta\, k_2(T)$
- Bound the number of pulls of arm 2 by $\frac{\log(T)}{\Delta^2}$ to get $\frac{\log(T)}{\Delta}$ regret bound
- How many pulls of arm 2 are actually needed?

Easy situation

After $n \geq \frac{16 \log(T)}{\Delta^2}$ pulls of arm 2 **and arm 1**

- Empirical means are well separated

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

  (Using Azuma Hoeffding inequality)

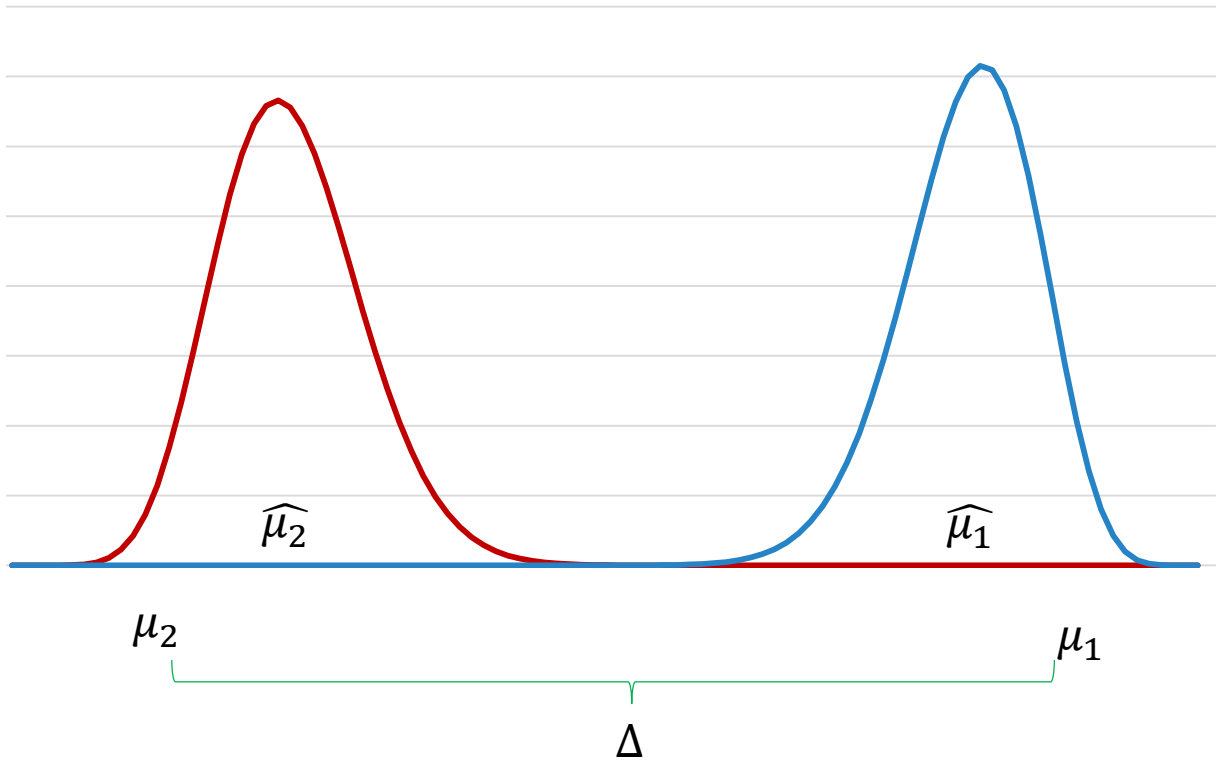
- Beta Posteriors are well separated

  Mean = $\frac{\alpha_a}{\alpha_a + \beta_a} = \widehat{\mu_a}$

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

The two arms can be distinguished!

No more arm 2 pulls.

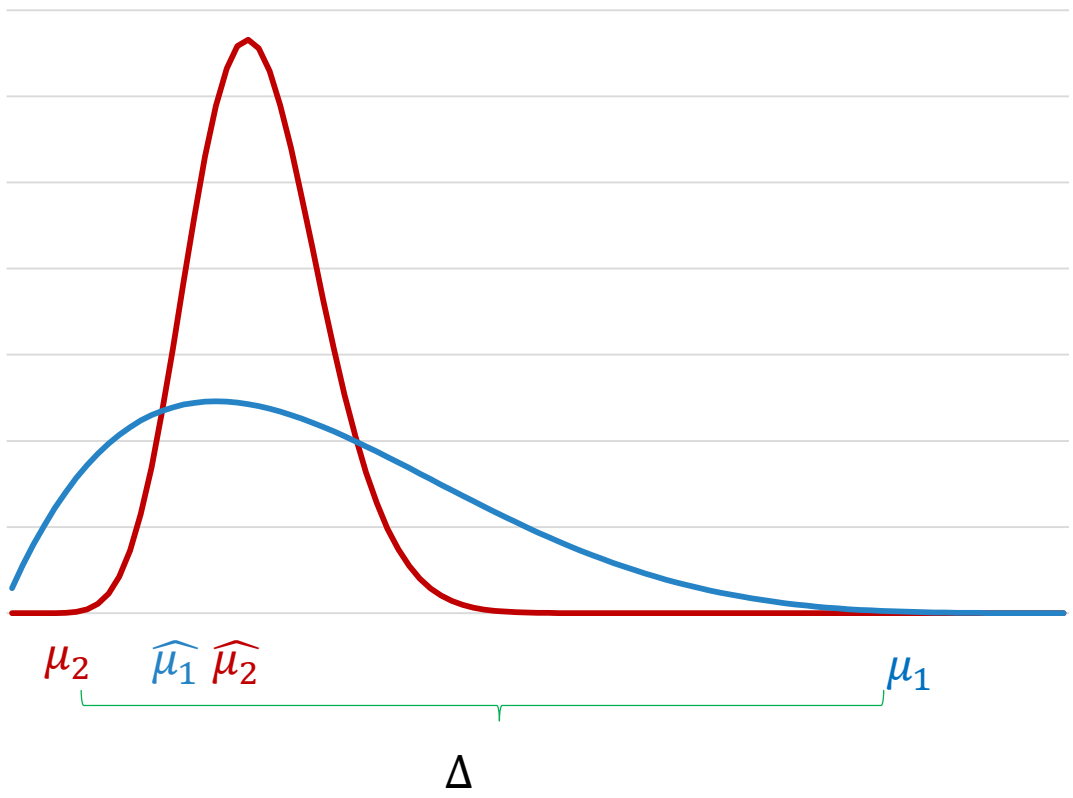


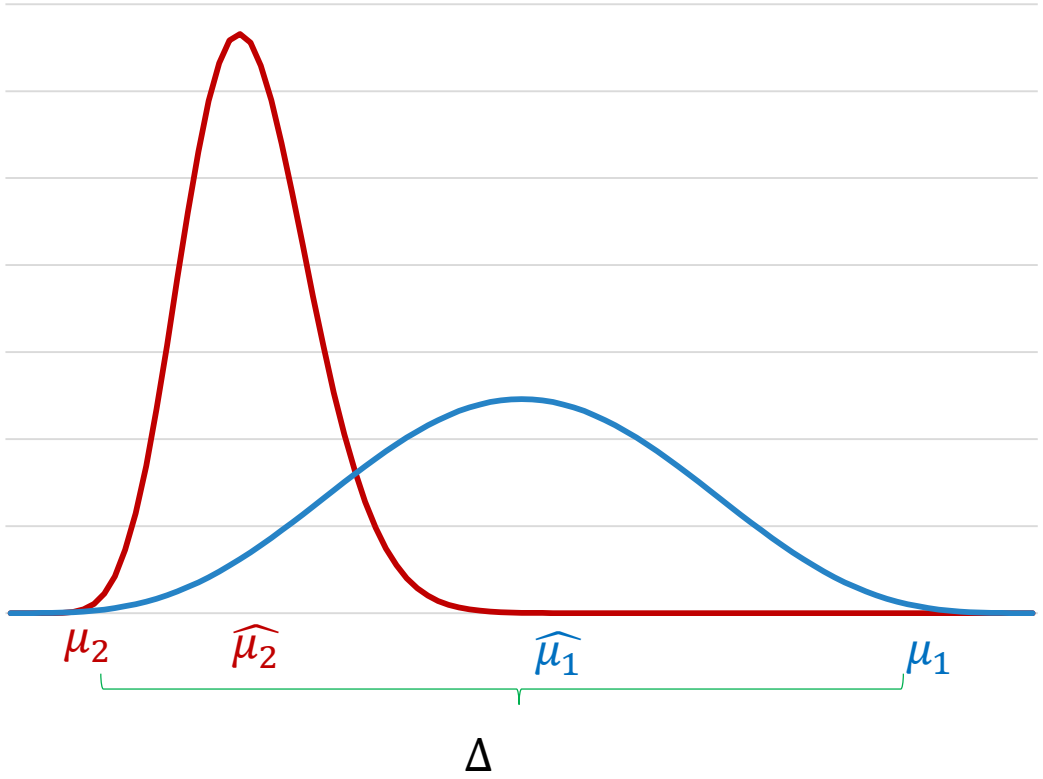
Easy situation

- Before arm 2 is pulled less than n= $\frac{16\log(T)}{\Delta^2}$ times?
  - Regret is at most $n\Delta = \frac{16\log(T)}{\Delta}$

Difficult situation

- ***After*** $\frac{16\log(T)}{\Delta^2}$ pulls of arm 2, but ***before*** arm 1 is pulled enough





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 before enough pulls of arm 1 is at most $\mathrm{O}(\frac{\log T}{\Delta^2})$

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

Multiple arms case

- Main observation: Given some high probability events

$$\Pr(a_t = a^* \mid F_{t-1}) \geq \frac{p}{1-p} \cdot \Pr(a_t = a \mid F_{t-1})$$

- p is the probability of anti-concentration of posterior sample for the best arm
  - E.g., $p := \Pr(\theta_{a^*} \geq \mu_{a^*} - \frac{\Delta_a}{4})$
- Best arm gets played roughly every $\frac{1}{p}$ plays **of arm a**
  - p can be lower bounded by Δ_a in general but it actually goes to 1 exponentially fast with increase in number of trials of best arm.
  - Cannot accumulate much regret from arm a without playing arm a^* sufficiently

Part II: Thompson Sampling for (tabular) RL

The reinforcement learning problem

System dynamics given by an MDP (S, A, r, P, s_0)

Rounds $t = 1, \dots, T$. In round t,

- observe state s_t, take action a_t,
- observe reward $r_t \in [0,1]$, $\mathrm{E}[r_t] = \boldsymbol{r_{s_t,a_t}}$, next state s_{t+1} with probability $\mathbf{P}_{\boldsymbol{s_t,a_t,}}(\boldsymbol{s_{t+1}})$

Solution concept: optimal policy

- $\pi: S \to A$, Action $\pi(s)$ in state s

Goal: minimize regret compared to best stationary policy π^*

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

Learn the MDP model parameters ($\boldsymbol{r, P}$) from observations **while maximizing** reward

Posterior Sampling

- Finite state finite action MDP: S states , A actions
- Assume for simplicity: Known reward distribution
- Needs to learn the unknown transition probability vector $P_{s,a} = (P_{s,a}(1), \dots, P_{s,a}(S))$ for all s, a
  - S^2A parameters
- 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 $\text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_S)$ on $P_{s,a}$
- On observing a categorical outcome i (new state), Bayesian posterior on $P_{s,a}$

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

- 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 probability

Posterior Sampling for RL (Thompson Sampling)

Learning

- Maintain a Dirichlet posterior for $P_{s,a}$ for every s, a
  - Start with an uninformative prior e.g., $\text{Dirichlet}(1,1,\ldots,1)$
  - 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 $\tilde{\pi}$ for sample MDP $(S, A, \tilde{P}, r, s_0)$
- Choose $a_t = \tilde{\pi}(s_t)$

Our Algorithm (Optimistic Posterior Sampling)

- 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

Bounding regret for communicating MDPs

Non-episodic setting, no restarts

- Can get stuck on a bad state for a long time

Communicating MDPs:

- There is always a way to get out of a bad state in finite time
- Definition: For every pair of states s, s', there exists a policy π such that using this policy starting from s, expected time to reach s' is finite and bounded by D, called **the diameter** of the MDP

Useful properties of communicating MDPs

- Optimal asymptotic average reward doesn't depend on the starting state.
- Asymptotic average reward (**Gain**) of policy π

$$\lambda^{\pi}(s) = E\left[\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} r\left(s_t, \pi(s_t)\right) \mid s_1 = s\right]$$

- There exists a single policy π^* such that

$$\max_{\pi} \lambda^{\pi}(s) = \lambda^{\pi^*}(s), \forall s \quad =: \lambda^* \text{ (**Optimal gain**)}$$

- Regret compared to asymptotic optimal

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

λ^* is optimal infinite horizon average reward (gain), achieved by best single policy π^*

Main result [A., Jia NeurIPS 2017]

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

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

λ^* is optimal infinite horizon average reward (gain), achieved by best single policy π^*

- Theorem: For any **communicating** MDP M with (unknown) diameter D, and for $\mathrm{T} \geq S^4 A^3$, with high probability:

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

Related work

[Our result] Worst-case regret bound $\tilde{O}(DS\sqrt{AT})$ for communicating MDP of diameter D

[Jaksch, Ortner, Auer, 2010] [Bartlett, Tewari, 2012]

- Matches previously best known worst-case regret bound $\tilde{O}(DS\sqrt{AT})$ for communicating MDP
  - With simpler algorithm
- Lower bound $\Omega\left(\sqrt{DSAT}\right)$

Other related results

- (episodic MDP) Posterior sampling with *Bayesian regret* bound of $\tilde{O}(H\sqrt{SAT})$ [Osband and Van Roy, 2016, 2017]
- (episodic MDP) Worst-case rerget bound $\tilde{O}\left(\sqrt{HSAT}\right)$ [Azar, Osband, Munos, 2017] [Kakade, Wang, Yang 2018]
- (Prior knowledge) $\tilde{O}(\sqrt{t_{mix}SAT})$ for uniformly ergodic MDP with known mixing time t_{mix}

Recap: Optimistic Posterior Sampling algorithm

- 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

Regret Analysis (outline)

- Lemma : With high probability, sample extended MDP is a communicating MDP with diameter at most 2D
- Useful property of communicating MDP: Optimal asymptotic average reward doesn't depend on the starting state.
- Average regret in an epoch k

$$\lambda^* - \frac{1}{T_k}\sum_{t \in T_k} r_{s_t,a_t} = \boxed{(\lambda^* - \tilde{\lambda})} + \left(\tilde{\lambda} - \frac{1}{T_k}\sum_{t \in T_k} r_{s_t,a_t}\right)$$

  - Optimal Gain (Asymptotic average reward) for true MDP M (policy π^*)
  - Optimal Gain for sample extended MDP (policy $\tilde{\pi}_k$)
- **First term:** we show optimism $\tilde{\lambda} \geq \lambda^*$

Gain and bias for communicating MDP

- Bellman equations relating optimal average reward to per state reward: $\forall\ s$,

$$\lambda^* = r_{s,\pi^*(s)} + P_{s,\pi^*(s)} \cdot h^* - h_s^*$$

- $h^* \in R^S$ is called **bias vector** and satisfies $\left|h_s^* - h_{s'}^*\right| \leq D$ for all s, $s' \in S$

Optimism

Two main results

Optimism of transition matrix on a projection is sufficient

$$\tilde{\lambda} \geq \lambda^* \text{ if for every } s, a,\ \tilde{P}_{s,a} \cdot h^* \geq P_{s,a} \cdot h^*$$

If a set of samples satisfy optimism on projection to *unknown* bias vector h^*

Anti-concentration of Dirichlet posterior:

For any fixed bounded vector h, a sample satisfies above with probability $1/S$

- No need to know h^* **!**
- (need multiple $O(S \log(\frac{SA}{\rho}))$ samples for high probability)

Regret Analysis (main insights)

- Average regret in an epoch k

$$\lambda^* - \frac{1}{T_k}\sum_{t \in T_k} r_{s_t,a_t} = \overbrace{(\lambda^* - \tilde{\lambda})}^{\leq 0} + \left(\tilde{\lambda} - \frac{1}{T_k}\sum_{t \in T_k} r_{s_t,a_t}\right)$$

  - Optimal Gain for true MDP M (policy π^*)
  - Optimal Gain for sample extended MDP (policy $\tilde{\pi}_k$)

- **Second term**
  - Same policy but different MDP
  - $\tilde{\lambda}$: follows sampled transition probability vector
  - $\frac{1}{T_k}\sum_{t \in T_k} r_{s_t,a_t}$: follows true transition probability vector
  - Bounded using **concentration** of transition probability vector samples from posterior

Concentration

- Recall Lemma: With high probability, sample extended MDP is a communicating MDP with diameter at most 2D
- Recall Bellman equations for communicating MDP: for all s

$$\tilde{\lambda} - r_{s,\pi(s)} = \tilde{P}_{s,\pi(s)} \cdot \tilde{\mathrm{h}} - \tilde{h}_s$$

- $\tilde{h} \in R^S$ is the bias vector of sample and satisfies $\left|\tilde{h}_i - \tilde{h}_j\right| \leq 2D$ for all $i, j \in S$

$$\left(\tilde{\lambda} - \frac{1}{T_k} \sum_{t \in T_k} r_{s_t,a_t}\right) = \frac{1}{T_k} \sum_{t \in T_k} (\tilde{P}_{s_t,a_t} \cdot \tilde{\mathrm{h}} - \tilde{h}_{s_t})$$

Bounding the second term

$$\left(\tilde{\lambda} - \frac{1}{T_k}\sum_{t\in T_k} r_{s_t,a_t}\right) = \frac{1}{T_k}\sum_{t\in T_k}(\tilde{P}_{s_t,a_t}\cdot \tilde{\mathrm{h}} - P_{s_t,a_t}\cdot \tilde{\mathrm{h}} + P_{s_t,a_t}\cdot \tilde{\mathrm{h}} - \tilde{h}_{s_t})$$

- Martingale: $P_{s_t,a_t}\cdot \tilde{\mathrm{h}} = \mathrm{E}[\tilde{\mathrm{h}}\cdot 1_{\mathrm{s_{t+1}}}] = \mathrm{E}[\tilde{h}_{s_{t+1}}]$
- Bias $\tilde{\mathrm{h}}$ is bounded
- Use Azuma-Hoeffding to bound by $\tilde{O}(\mathrm{D}\sqrt{T})$

Bounding the second term

$$\left(\tilde{\lambda} - \frac{1}{T_k}\sum_{t\in T_k} r_{s_t,a_t}\right) = \frac{1}{T_k}\sum_{t\in T_k} (\tilde{P}_{s_t,a_t}\cdot \tilde{\mathrm{h}} - P_{s_t,a_t}\cdot \tilde{\mathrm{h}} + P_{s_t,a_t}\cdot \tilde{\mathrm{h}} - \tilde{h}_{s_t})$$

- Bound deviation of posterior sample from true model $(\tilde{P}_{s,a} - P_{s,a})\cdot \tilde{\mathrm{h}}$
  - Bound posterior variance
  - Bound sample error
- Challenge: $\tilde{\mathrm{h}}$ is not fixed, need union bound
- $\tilde{O}(\frac{D\sqrt{S}}{\sqrt{N_{s,a}}})$ bound

Summary

- Regret in an epoch k

$$\mathrm{T_k}\,\lambda^* - \sum_{t\in T_k} r_{s_t,a_t} = \left(T_k\lambda^* - T_k\tilde{\lambda}\right) + \left(T_k\tilde{\lambda} - \sum_{t\in T_k} r_{s_t,a_t}\right)$$

- At most $SA\log(T)$ Epochs
- Overall bound of $\tilde{O}(DS\sqrt{AT})$

Some further directions

- Extension to contexts
  - Transition probability depends on time varying context in addition to state
- How to integrate posterior sampling with Q-learning or policy gradient methods?
  - Worst-case regret bounds