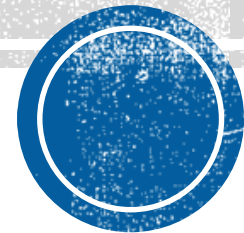# Theoretical Guarantees in Reinforcement Learning : Regret analysis

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

- <u>Using a single thread of experience</u>
  - 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 $E[\sum_{t=1}^{\infty} \gamma^{t-1} r_t]$ 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
  - <u>Theorem</u> [Agrawal, Jia, 2017]<u>:</u> For **any communicating MDP** M with (unknown) diameter D, and for $T \geq S^5 A$, our algorithm achieves: $\mathrm{Regret}(M,T) \leq \tilde{O}(D\sqrt{SAT})$

- Bayesian regret bounds: (expectation over known prior on MDP M); example
  - <u>Theorem</u> [Osband, Russo, Van Roy, 2013]<u>:</u> For **any known <u>prior distribution</u>** f on MDP M, our algorithm achieves in episodic setting: $\mathrm{E}_{M \sim f}[\mathrm{Regret}(M,T)] \leq \tilde{O}(HS\sqrt{AT})$

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

1.0, r(s,1)

s

1.0, r(s,2)

$P(s, a, s), \ r(s, a)$
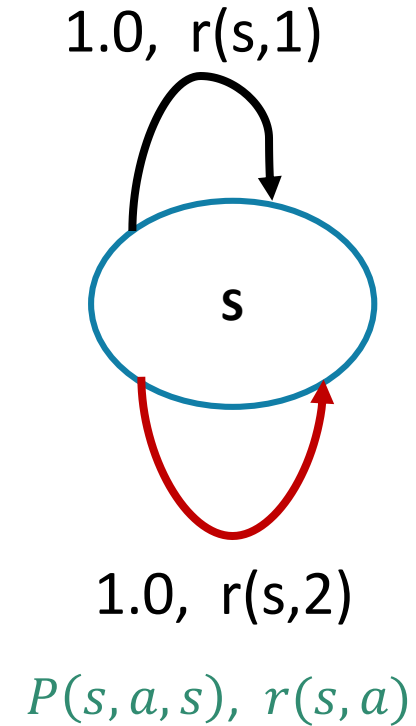
Stochastic multi-armed bandit problem

- Online decisions
  - At every time step $t = 1, \ldots, T$, pull one arm out of N arms

- Stochastic feedback
  - For each arm 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 $I_t = i$ 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 $\dfrac{C \, \log(T)}{\Delta_i^2}$, then regret bound: $\sum_{i:\mu_i \neq \mu^*} \dfrac{C \, \log(T)}{\Delta_i}$
  - Problem-dependent bound, close to lower bound [Lai and Robbins 1985]

- For problem independent bound: $C\sqrt{NT log(T)}$

  - Separately bound total regret for playing an arm with $\Delta_i \leq \sqrt{\dfrac{N\log(T)}{T}}$ and $\Delta_i > \sqrt{\dfrac{N\log(T)}{T}}$
  - Lower bound $\Omega(\sqrt{NT})$,

The need for exploration

- Two arms **black** and **red**
  - Random rewards with unknown mean $\boldsymbol{\mu_1 = 1.1}$, $\textcolor{red}{\boldsymbol{\mu_2 = 1}}$
  - 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 $t = 1, \ldots, N$ **do**
|Play arm t
end
foreach $t = N+1, N+2 \ldots, T$ **do**
|Play arm $I_t = \arg\max_{i \in \{1...N\}} \mathrm{UCB}_{i,t-1}$.
|Observe r_t, compute $\mathrm{UCB}_{i,t}$
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 $n_{i,t} > \frac{16\ln(T)}{\Delta_i^2}$, $\hat\mu_{i,t} < \mu_i + \frac{\Delta_i}{2}$

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

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

$$\text{UCB}_{i,t} = \hat\mu_{i,t} + \sqrt{\frac{\ln t}{n_{i,t}}} \leq \hat\mu_{i,t} + \frac{\Delta_i}{2}$$

$$< \left(\mu_i + \frac{\Delta_i}{2}\right) + \frac{\Delta_i}{2}$$

$$= \mu^*$$

$$< \text{UCB}_{i^*,t}$$

16

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
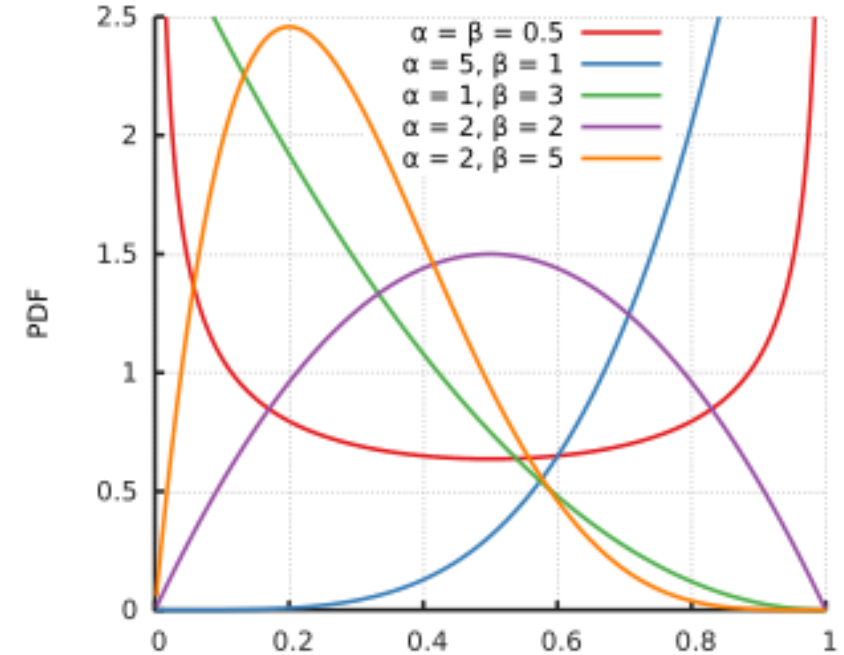
18
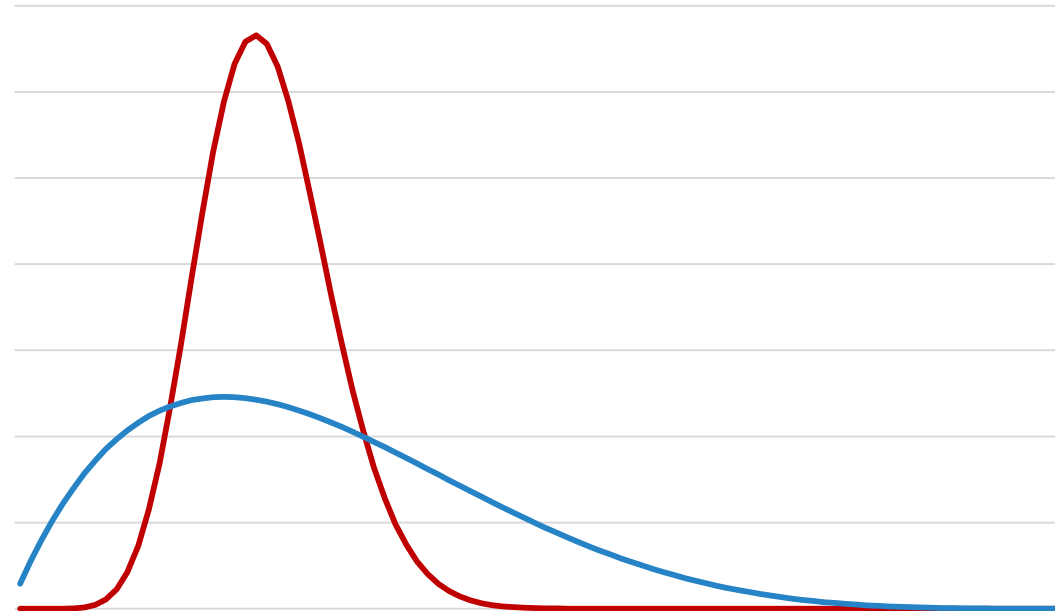
Bernoulli rewards, Beta priors

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 $Beta(\alpha_0 + S_{i,t}, \beta_0 + F_{i,t})$
- 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

- <u>Theorem</u> [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 $\mu_1, \mu_2, \ldots, \mu_N$ 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

- $\text{Regret}(T) \leq \boldsymbol{ln(T)}(1+\epsilon)\sum_i \frac{\Delta_i}{KL(\mu^*||\mu_i)} + O(\frac{N}{\epsilon^2})$
  - 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)

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

Instance-independent bounds (Beta and Gaussian priors)

- $\text{Regret}(T) \leq O(\sqrt{NT\ln T}\,)$

- 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 $\dfrac{\log(\text{T})}{\Delta^2}$ to get $\dfrac{\log(\text{T})}{\Delta}$ 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 $|\widehat{\mu_i} - \mu_i| \leq \sqrt{\frac{\log(T)}{n}} \leq \frac{\Delta}{4}$ whp
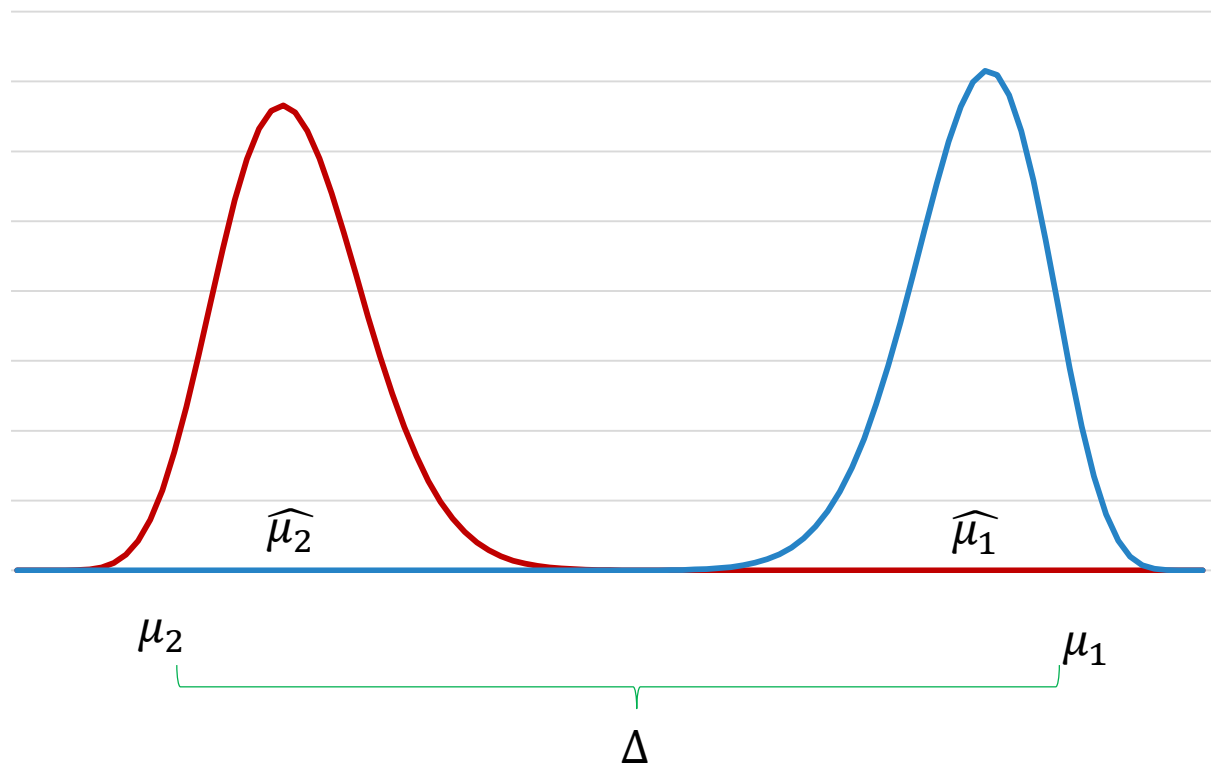
  (Using Azuma Hoeffding inequality)

- Beta Posteriors are well separated

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

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

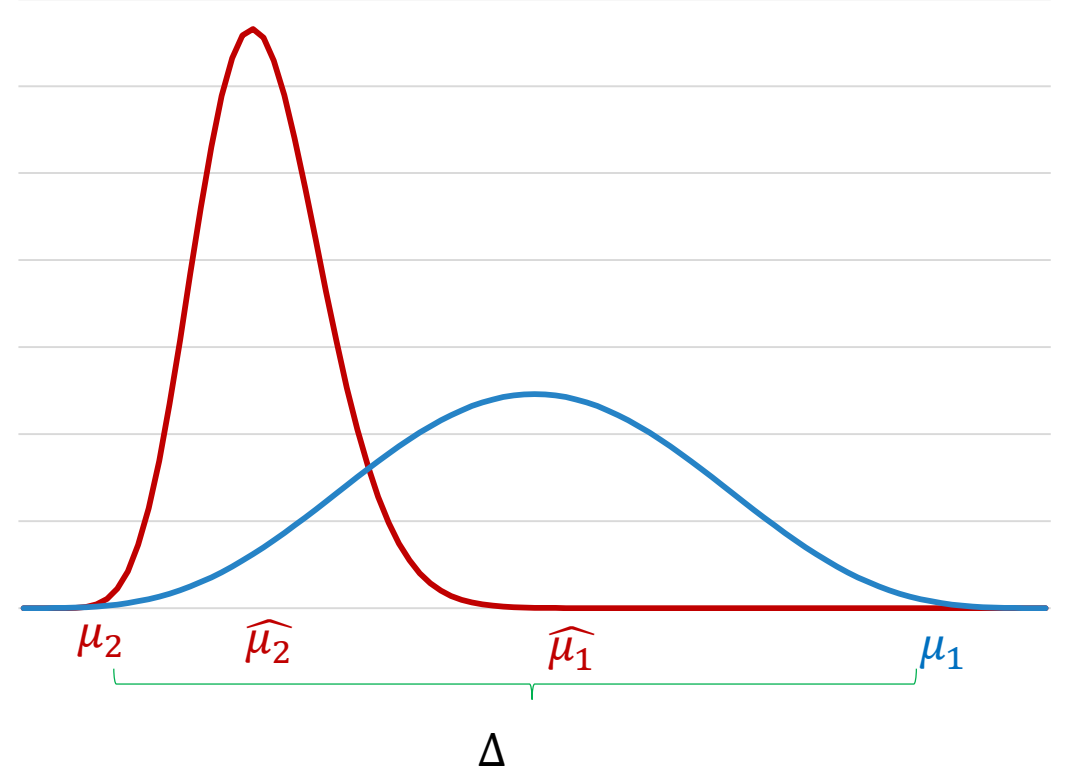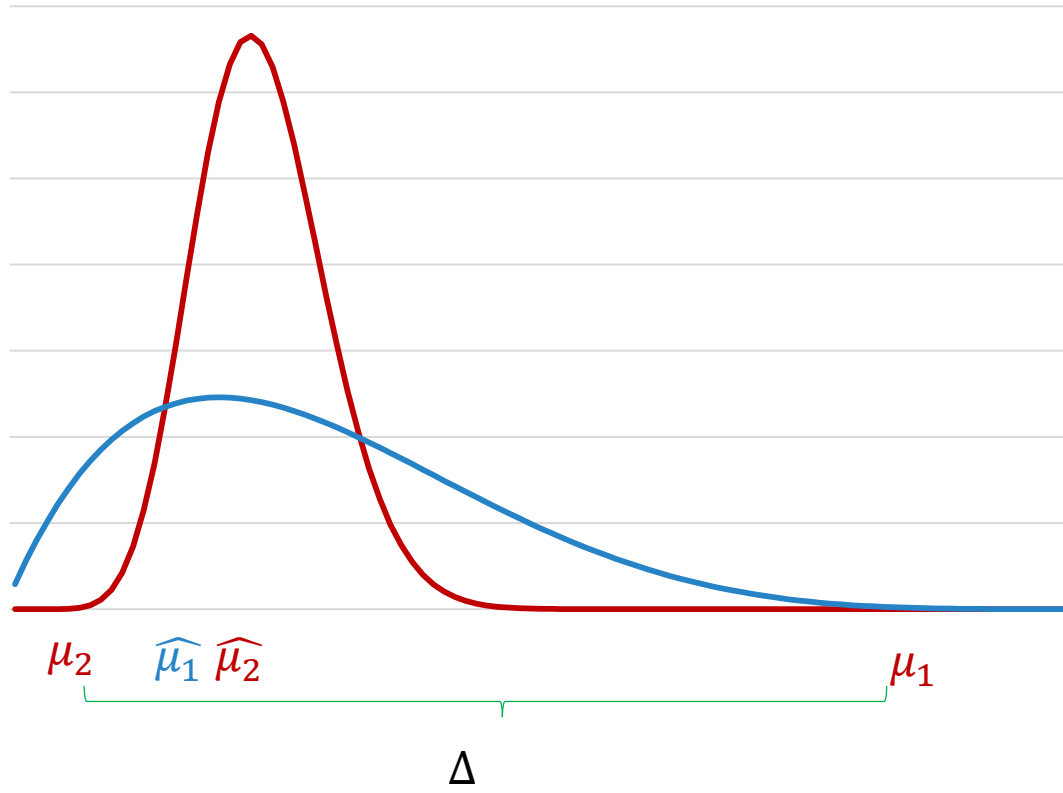The two arms can be distinguished!

No more arm 2 pulls.

Easy situation

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

Difficult situation

- At least $\dfrac{\log(\text{T})}{\Delta^2}$ 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