Thompson Sampling overview

Shipra Agrawal

Thompson Sampling [Thompson 1933]

A simple heuristic based on Bayesian philosophy of learning

Simple Bayesian heuristic

- Aim is to quickly identify the Bernoulli arm with highest mean
- Need to learn the unknown mean for each arm to sufficient accuracy

Heuristic:

For every arm, maintain a belief about where the mean is in [0,1]

- Initially you are very unsure about where the mean is, start with uniform
- Shift the distribution to left if you observe a 0, right if you observe a 1
- Play the arm with highest probability of "being the best arm"

This is Thompson Sampling!

Thompson Sampling [Thompson, 1933]

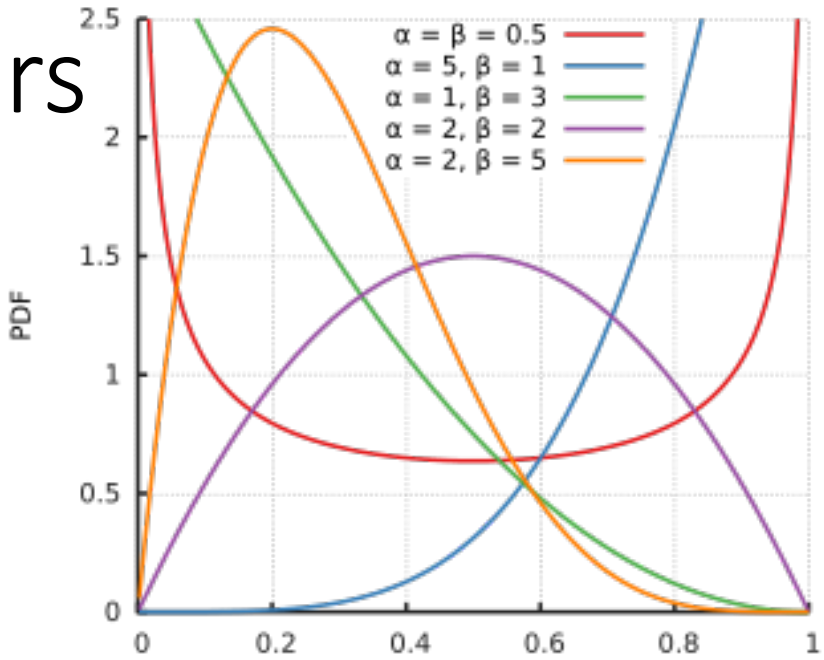
- Maintain belief about effectiveness (mean reward) of each arm
- Observe feedback, update belief of pulled arm *i* in Bayesian manner

  Bayes rule $\Pr(\mu_i \,|r) \propto \Pr(r|\mu_i) \cdot \Pr(\mu_i)$
- Pull arm with posterior probability of being best arm
  - NOT choose the one most likely to be effective
  - Gives benefit of doubt those less explored

- Thompson Sampling gives "optimal" benefit of doubt

  [Agrawal and Goyal, COLT 2012, AISTATS 2013]

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

Regret bounds

[A. Goyal, COLT 2012, AISTATS 2013]

Optimal instance-dependent bounds for Bernoulli rewards

- $\frac{\text{Regret(T)}}{\log(T)} \rightarrow constant\ (instance\ specific)$
  - Matches *asymptotic lower bound*

Near-optimal worst-case-instance bounds

- $\text{Regret}(T) \leq\ O(\sqrt{NT \ln T}\ )$
  - Lower bound $\Omega(\sqrt{NT})$

- Only assumption: Bernoulli likelihood

Thompson Sampling with Gaussian Priors

- Suppose reward for arm i is i.i.d. $N(\mu_i, 1)$
- Starting prior N(0,1)
- Gaussian Prior, Gaussian likelihood → Gaussian posterior $N(\hat{\mu}_{i,t}, \frac{1}{n_{i,t}+1})$
  - $\hat{\mu}_{i,t}$ is empirical mean of $n_{i,t}$ observations for arm i

- Algorithm
  - Sample $\theta_{i,t}$ from posterior $N\left(\hat{\mu}_{i,t}, \frac{1}{n_{i,t}+1}\right)$ for arm i
  - Play arm $I_t = \arg\max_i \theta_{i,t}$
  - Observe reward, update empirical mean for arm i_t

- Now apply this algorithm for any reward distribution!

Regret bounds
[A., Goyal, AISTATS 2013]

Near-optimal instance-dependent bounds

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

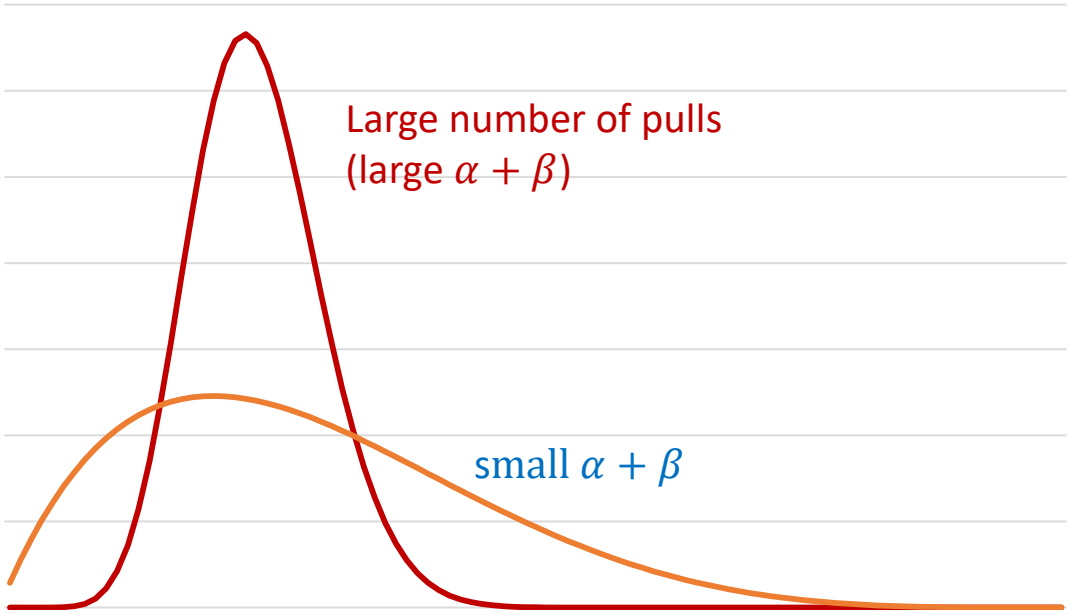
Near-optimal worst-case-instance bounds

- Regret$(T) \leq O(\sqrt{NT \ln N})$
  - Matches lower bound within logarithmic factors

- Only assumption: Bounded [0,1] or subGaussian noise $\eta_t = r_t - \mu_{i_t}$

Why does it work?

- The Bayesian approach takes into account the level of uncertainty about the means.
- Higher uncertainty ensures that an less explored arm is pulled
- The uncertainty reduces as the arm is explored more and more



http://eurekastatistics.com/beta-distribution-pdf-grapher/

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 $\frac{\log(T)}{\Delta^2}$ to get $\frac{\log(T)}{\Delta}$ regret bound

Easy situation

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

- Empirical means are well separated

$$\text{Error} \leq \sqrt{\frac{\log(T)}{n}} \leq \frac{\Delta}{4} \text{ whp}$$

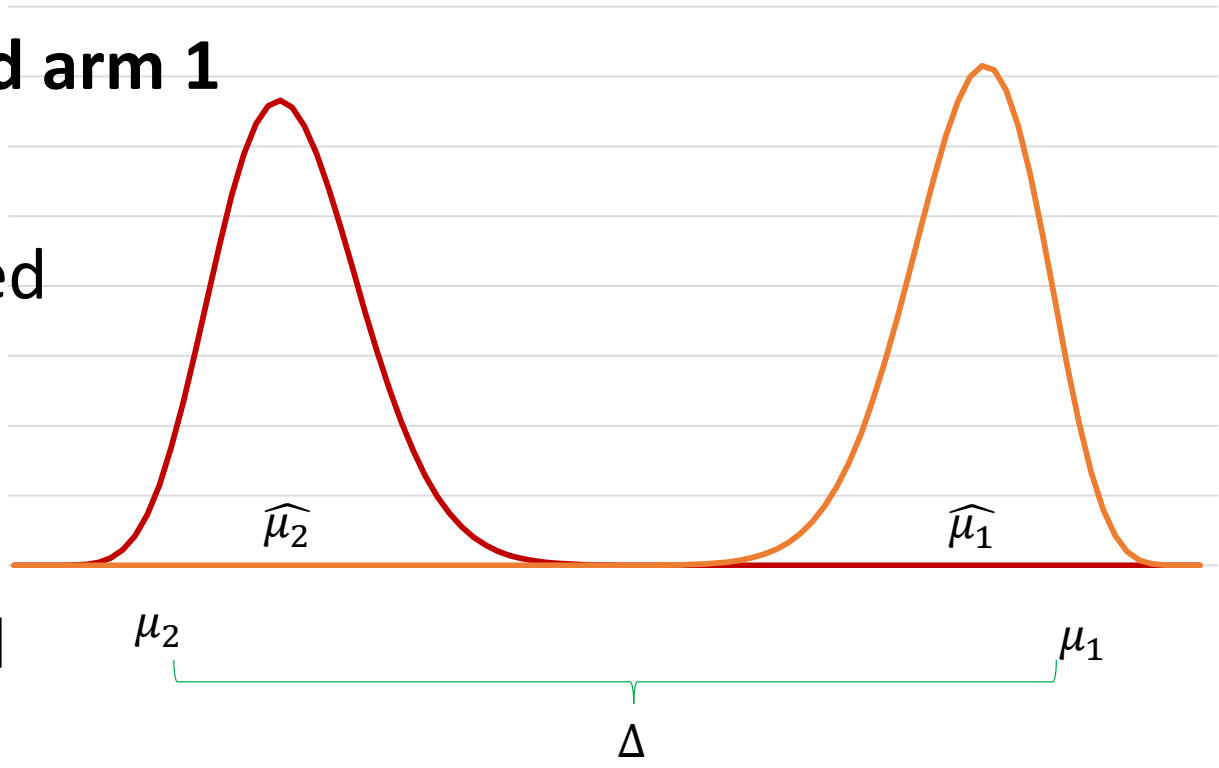
(Using Azuma Hoeffding inequality)

- Beta Posteriors are well separated

$$\text{standard deviation} \simeq \frac{1}{\sqrt{n}} \leq \frac{\Delta}{4}$$

The two arms can be distinguished!

No more arm 2 pulls.

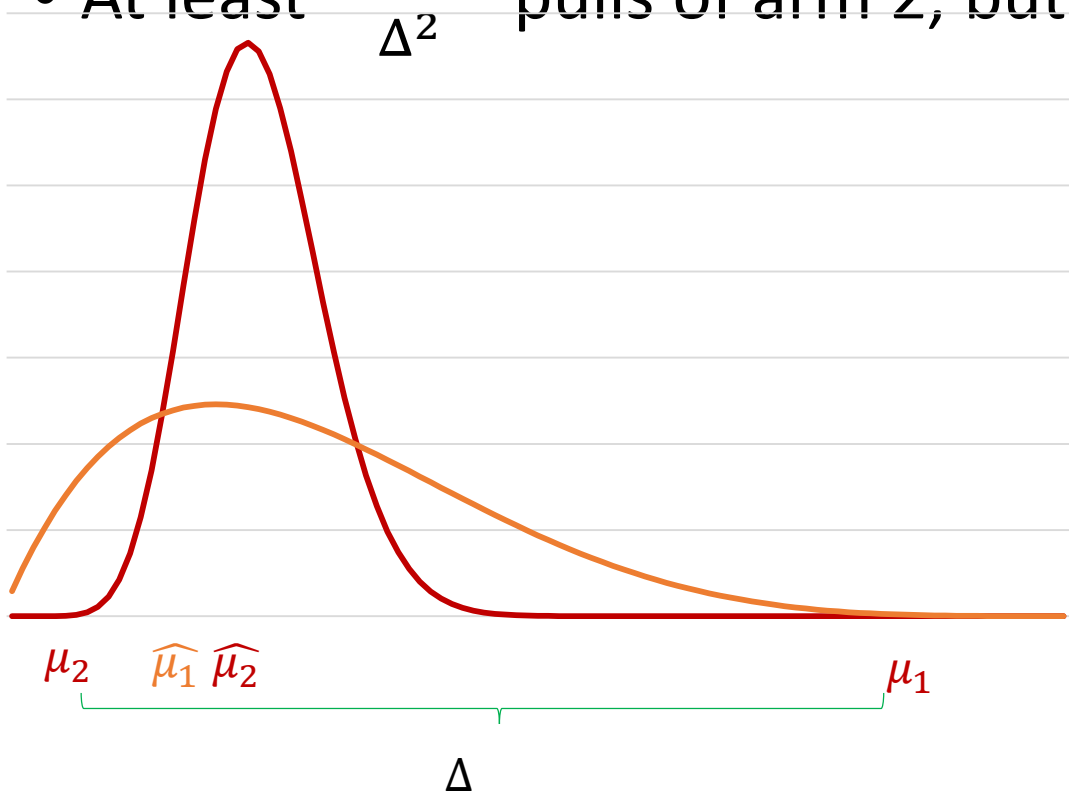


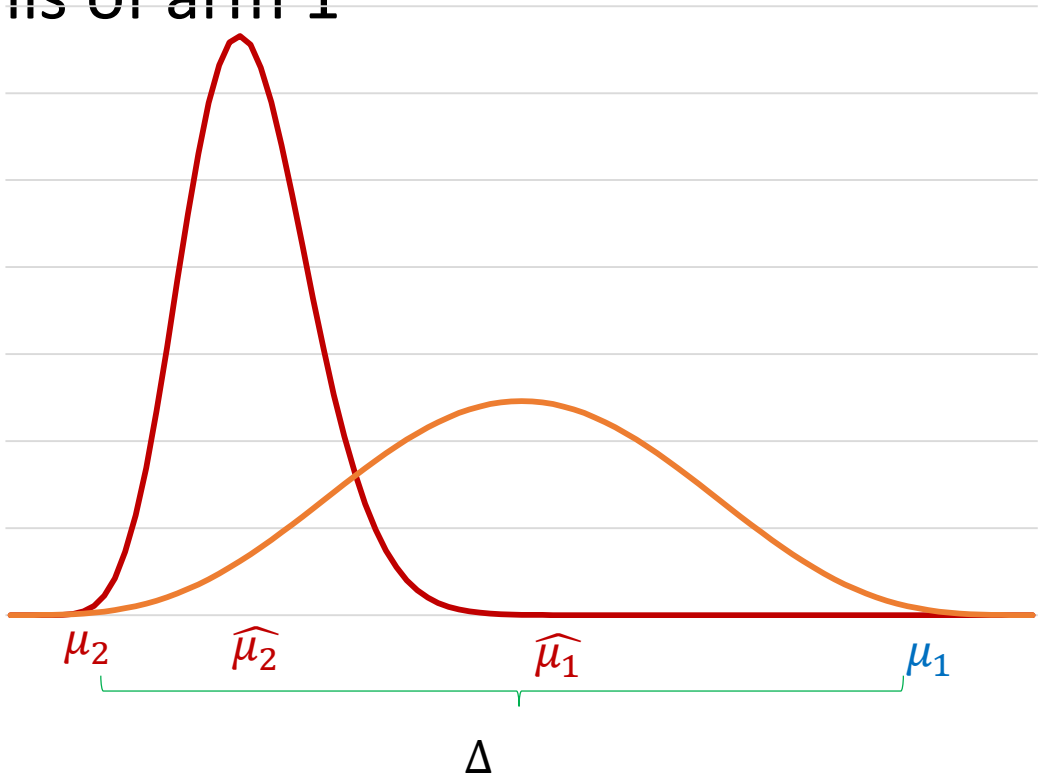
Easy situation

- If arm 2 is pulled less than $n=O(\frac{\log(T)}{\Delta^2})$ 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 (up to optimal constants) require more careful use of posterior structure