Stochastic multi-armed bandits

Shipra Agrawal

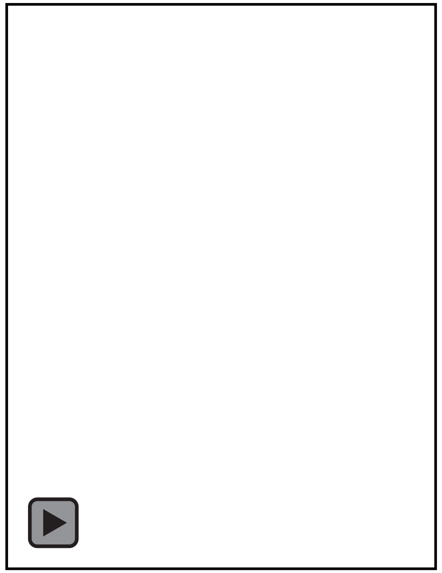
Industrial Engineering and Operations Research

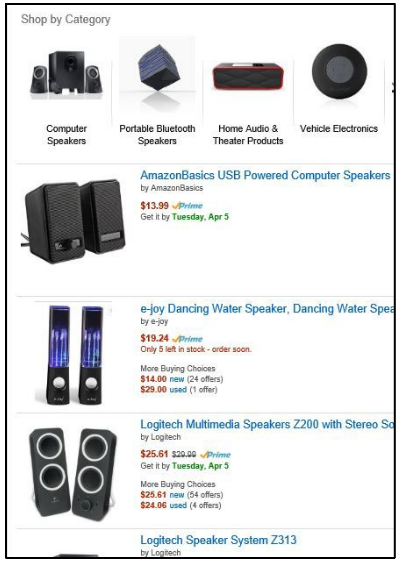
Columbia University

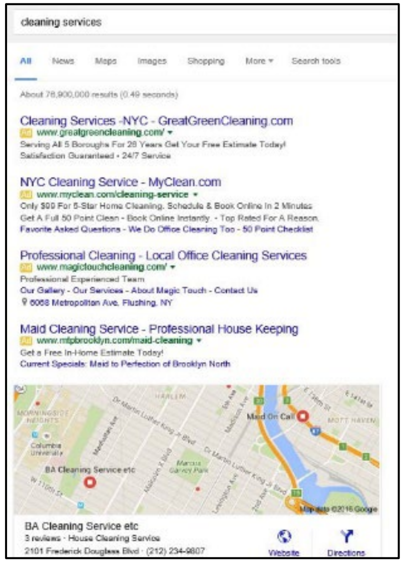
Learning from sequential interactions

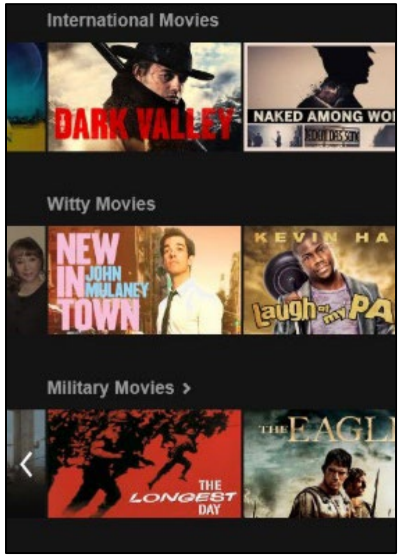
Tradeoff between

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









Robotic Manipulation and Mobility

Managing exploitation-exploitation tradeoff

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

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

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** support [0,1], mean μ_i
- Bandit feedback
  - Only the reward of the pulled arm can be observed
- Minimize **regret** compared to the best arm

$$E\left[\sum_{t=1}^{T}(\mu^* - \mu_{i_t})\right] \quad \text{where } \mu^* = \max_j \mu_j$$

Other formulations: Bayesian bandits and Gittins index

- Prior distribution over parameters of each arm's reward distribution
  - E.g. if arm i has reward distribution Bernoulli(μ_i), there is a prior on distribution of μ_i
  - On observing a reward we have a posterior
- Expected reward/regret:
  - *Expectation over prior distribution*, in addition to reward distribution of arms
- Gittins Index [Gittins, 1979]
  - Optimal policy when maximizing *expected total discounted reward*
- Bayesian Regret minimization
  - e.g., see [Osband, Russo and Van Roy 2013, Russo and Van Roy 2014, 2015, 2016], [Bubeck and Liu 2013]

Outline

- Basic algorithmic techniques for the stochastic MAB problem
  - UCB
  - Thompson Sampling
- Useful Generalizations
  - Contextual bandits
  - Assortment optimization
  - Bandits with constraints
- Later: Bandit techniques for MDP/RL

Recall: 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** support [0,1], mean μ_i
- Bandit feedback
  - Only the reward of the pulled arm can be observed
- Minimize regret in time T

$$E\left[\sum_{t=1}^{T}(\mu^* - \mu_{i_t})\right]$$

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

Lower bounds

- Expected regret in any time T,

$$\text{Regret}(T) = \sum_i (\mu^* - \mu_{i_t}) = \sum_i \Delta_i \, E[k_i(T)]$$

Lower bounds

- Lai and Robbins 1985 [Informal] For *any* given instance of the MAB problem, any "reasonable algorithm" will play a suboptimal arm at least $\Omega(\log(T))$ times for large T
- Worst case bound: For every algorithm, there exists an instance with $\Omega\left(\sqrt{NT}\right)$ regret

UCB algorithm [Auer 2002]

- 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 (UCB) $$UCB_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{4 \ln t}{n_{i,t}}}$$
- Optimism: $UCB_{i,t} > \mu_i$ w.h.p.
- Optimistic Algorithm
  - At each time step t, play the with best optimistic estimates $$i_t = \arg\max_i \mathrm{UCB}_{t,i}$$

UCB algorithm [Auer 2002]

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 $I_t = \arg\max_{i \in \{1,\ldots,N\}} \mathrm{UCB}_{i,t-1}$.
  Observe r_t, compute $\mathrm{UCB}_{i,t}$
end

Regret analysis

- Recall Regret in any time T,

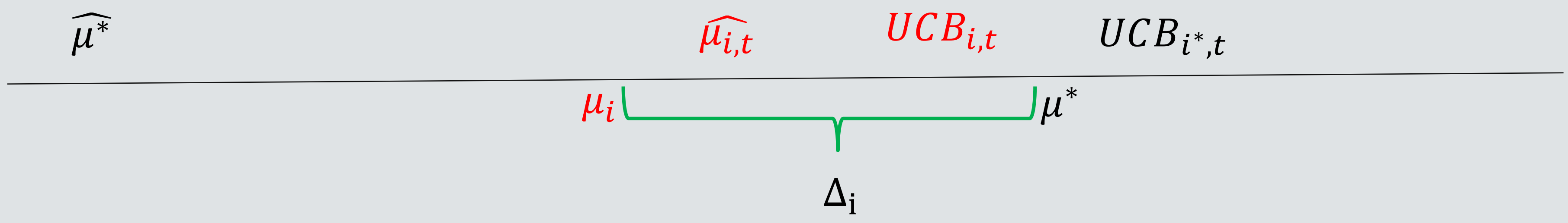
$$Regret(T) = \sum_i \Delta_{i_t} = \sum_i \Delta_i \, E[k_i(T)]$$

where $\Delta_i = \mu^* - \mu_i$

- Bound the number of mistakes $E[k_i(T)]$ for all suboptimal arms $i \neq i^*$
  - A bound of $E[k_i(T)] \leq \frac{C \ \ln T}{\Delta_i^2}$ for i each implies $\sum_{i \neq i^*} \frac{C \ \ln T}{\Delta_i}$ regret bound

Regret analysis

- Arm i will be played at time t only if $UCB_{i,t} > UCB_{i^*,t}$
- If $n_{i,t} > \frac{16\ln(T)}{\Delta_i^2}$: $\hat{\mu}_{i,t} < \mu_i + \frac{\Delta_i}{2}$, $UCB_{i,t} \leq \hat{\mu}_{i,t} + \frac{\Delta_i}{2}$



- No more plays of arm i (with high probability)
  - $\frac{16\ln(T)}{\Delta_i^2}$ bound on expected number of mistakes
  - $Regret(T) = \sum_i \Delta_{i_t} = \sum_i \Delta_i \, E[k_i(T)] \leq \sum_{i \neq i^*} \frac{16\ln(T)}{\Delta_i}$

Thompson Sampling [Thompson, 1933]

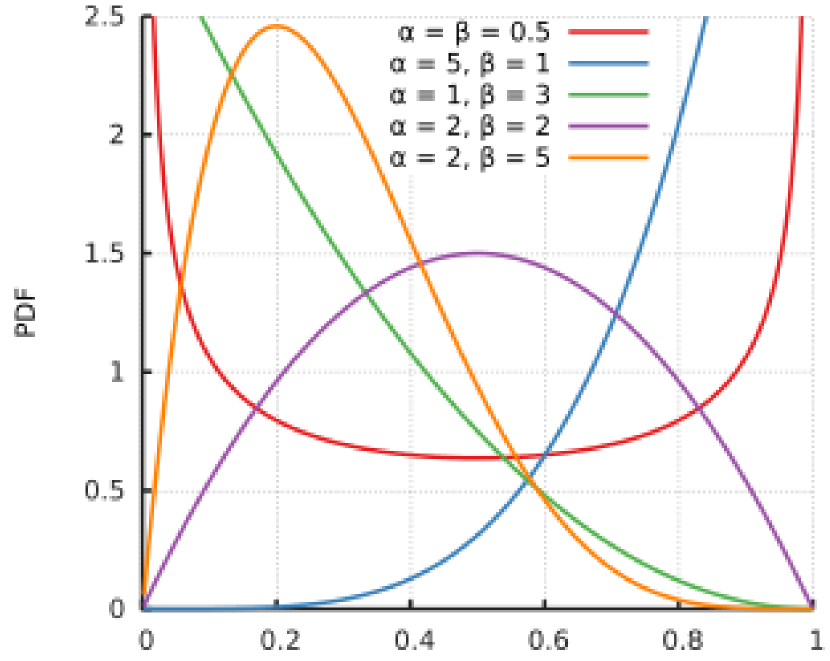
- Natural and Efficient heuristic
- Maintain belief about parameters (e.g., mean reward) of each arm
- Observe feedback, update belief of pulled arm *i* in Bayesian manner
- Pull arm with posterior probability of being best arm
  - NOT same as choosing the arm that is most likely to be best

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

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 $N(0, \nu^2)$ prior belief for every arm

In round t,

- For every arm i, sample $\theta_{i,t}$ independently from posterior $\mathrm{N}\left(\hat{\mu}_i, \frac{\nu^2}{n_i+1}\right)$
- 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

Regret bounds

Optimal instance-dependent bounds for Bernoulli rewards

- Regret$(T) \leq$ $\boldsymbol{ln}(\boldsymbol{T})(1+\epsilon)\sum_i \frac{\Delta_i}{KL(\mu^*||\mu_i)} + O(\frac{N}{\epsilon^2})$ [A. and Goyal 2012, 2013]
  - Matches *asymptotic instance wise lower bound* [Lai Robbins 1985]
  - Closely related bounds by [Kaufmann et al. 2013]
  - Bayesian UCB algorithm also achieves this [Kaufmann et al. 2012]

Arbitrary bounded reward distribution (Beta and Gaussian priors)

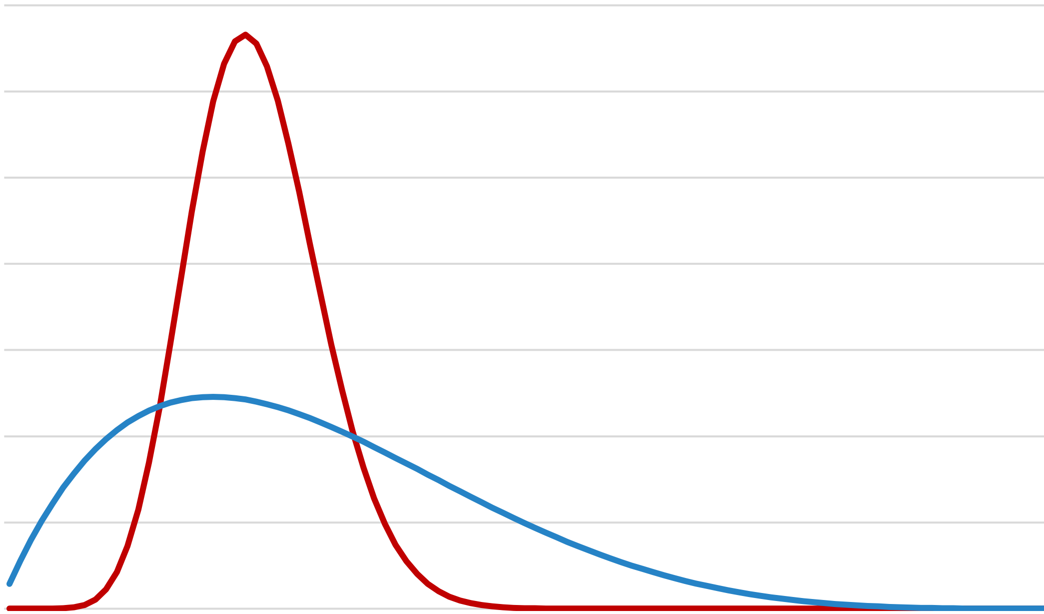
- Regret$(T) \leq$ $O(\boldsymbol{ln}(\boldsymbol{T})\sum_i \frac{1}{\Delta_i})$ [A. and Goyal 2013]
  - Matches the best available for UCB for general reward distributions

Instance-independent bounds (Beta and Gaussian priors)

- Regret$(T) \leq$ $O(\sqrt{NT\ln T})$ [A. and Goyal 2013]
  - Lower bound $\Omega(\sqrt{NT})$
- Prior and likelihood mismatch allowed! – worst case regret bounds

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



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
- How many pulls of arm 2 are actually needed?

Easy situation

After $\mathrm{n} \geq \frac{16\log(\mathrm{T})}{\Delta^2}$ 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)

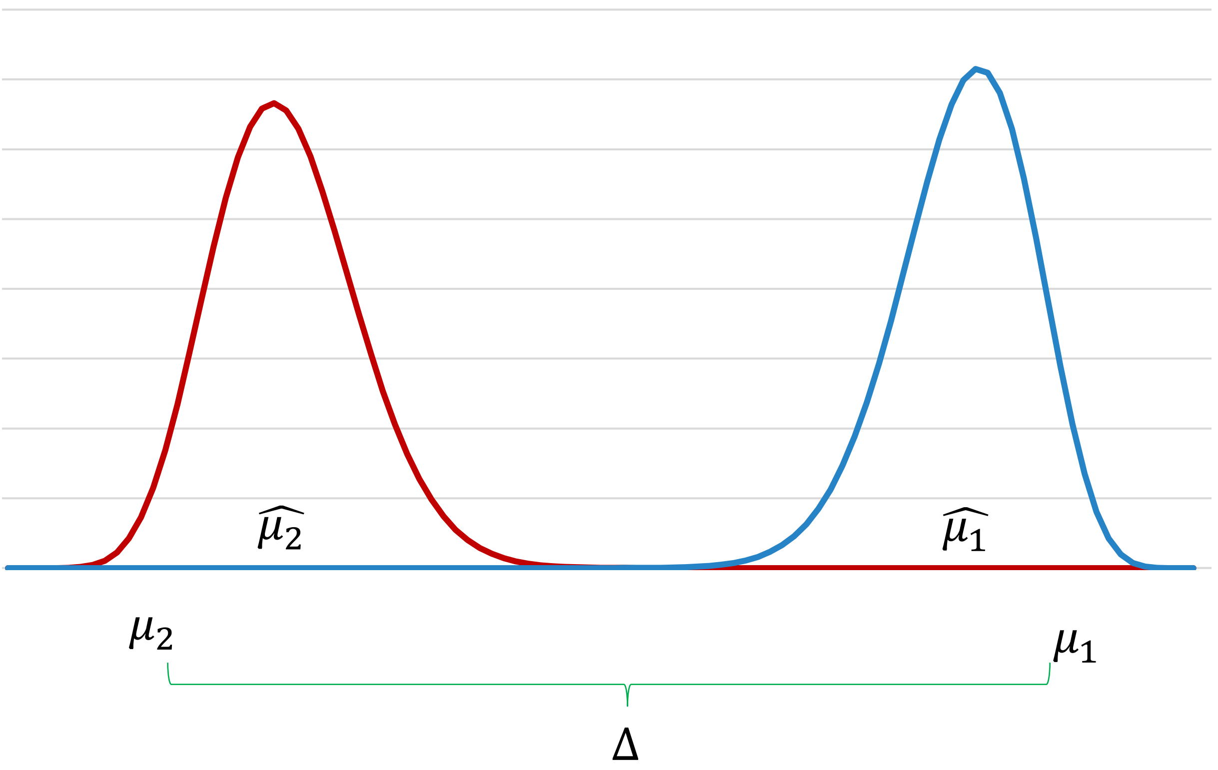
- Beta Posteriors are well separated

  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.

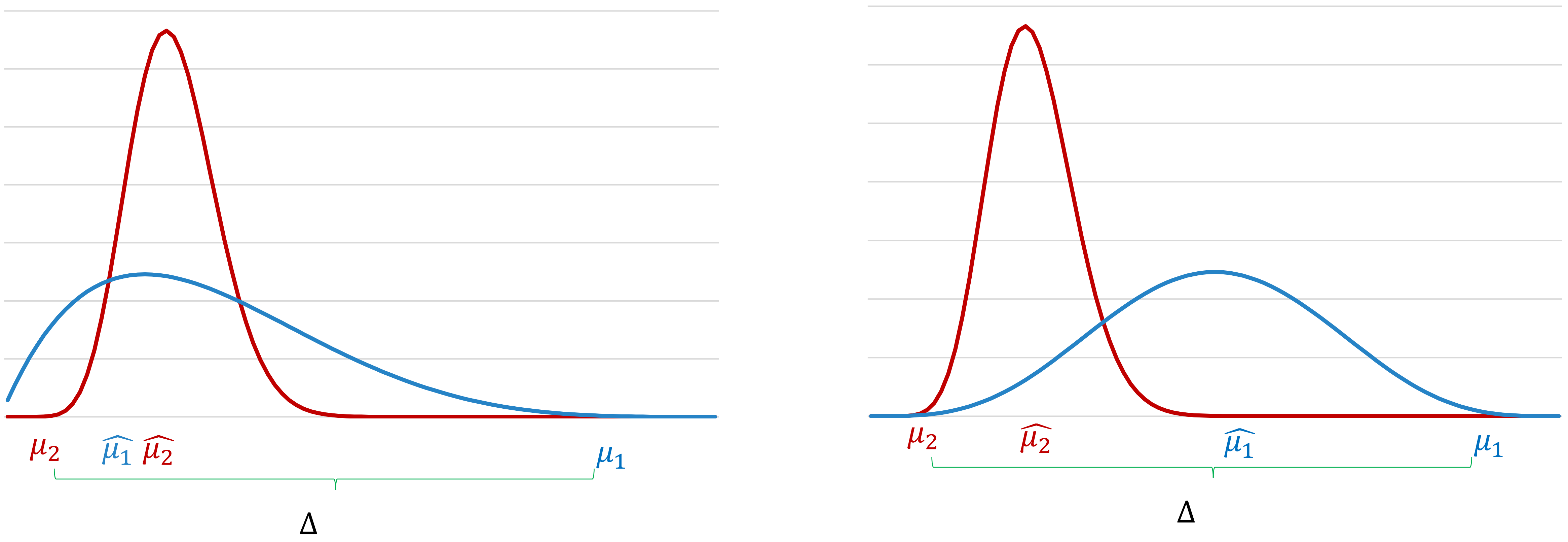


Easy situation

- 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 $O(\frac{\log T}{\Delta^2})$

- 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

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

- p is the probability of anti-concentration of posterior sample for the best arm
  - E.g., $p_a := \Pr(\theta_{i^*} \geq \mu_{i^*} - \frac{\Delta_i}{4})$
- Best arm gets played roughly every $\frac{1}{p}$ plays **of arm $\boldsymbol{i}$**
  - p can be lower bounded by Δ_i 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 i without playing arm i^* sufficiently

Next: Useful Generalizations of the basic MAB problem

Recall: The basic Stochastic multi-armed bandit problem

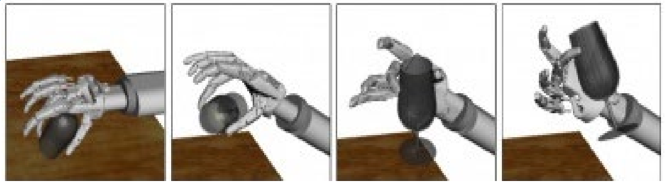
- Online decisions
  - At every time step $t = 1, \ldots, T$, pull one arm out of N arms
- Stochastic feedback
  - For each arm i, reward is generated **i.i.d. across time** from a **fixed but unknown distribution** support [0,1], mean μ_i
- Bandit feedback
  - Only the reward of the pulled arm can be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

$$E\left[\sum_{t=1}^{T}(\mu^* - \mu_{i_t})\right]$$

- Personalization
  - Linear contextual bandits

- Customer Choice behavior
  - Dynamic assortment selection

- Revenue management and resource allocation
  - Budget/supply constraints, nonlinear utilities



- Reinforcement learning
  - State-dependent response

- Inventory management, Dynamic Pricing
  - Learning continuous state MDPs

#1: Handling context in MAB

- Large number of products and customer types
- Utilize similarity?
- Content based recommendation (Supervised learning)
  - Customers and products described by their features
  - Similar features means similar preferences
  - Parametric models mapping customer and product features to customer preferences
    - E.g. linear regression
- Contextual bandits
  - Exploration-exploitation to learn the parametric models

Linear Contextual Bandits

- N arms, possibly very large N
- A d-dimensional context (feature vector) $x_{i,t}$ for every arm i, time t
- Linear parametric model
  - Unknown parameter **vector** $\boldsymbol{\mu}$
  - Expected reward for arm i at time t is $x_{i,t} \cdot \mu$
- Algorithm picks $x_t \in \{x_{1,t}, \ldots, x_{N,t}\}$, observes $r_t = x_t \cdot \mu + \eta_t$
- Optimal arm depends on context: $\mathrm{x}_\mathrm{t}^* = \arg\max_i x_{i,t} \cdot \mu$
- Goal: Minimize regret
  - Regret(T) = $\sum_t (x_t^* \cdot \mu - x_t \cdot \mu)$

UCB for linear contextual bandits

Linear regression

- Least square solution $\widehat{\mu_t}$ of set of $t-1$ equations

$$x_s \cdot \mu = r_s, \quad s = 1, \dots, t-1$$

- $\widehat{\mu_t} \simeq B_t^{-1}(\sum_{s=1}^{t-1} x_s r_s)$ where $B_t = I + \sum_{s=1}^{t-1} x_s x_s'$
- B_t^{-1} covariance matrix of this estimator

High confidence interval for θ

- With high probability $||\mu - \widehat{\mu_t}||_{B_t} \le C\sqrt{d\log(\mathrm{Td})}$

[Rusmevichientong and Tsitsiklis 2010] [Abbasi-Yadkori et al 2011]

UCB algorithm

- At time t
  - Observe the contexts $x_{i,t}$ for different arms $i = 1, \dots N$
  - Compute confidence interval for the unknown parameter
  - Choose the best arm according to the most optimistic parameter in C_t

$$C_t = \{z \ : \ ||z - \hat{\mu}||_{B_t} \leq C\sqrt{d \log(Td)}\}$$

Algorithm : LinUCB algorithm

```
foreach t = 1, ..., T do
  Observe set A_t ⊆ [N], and context x_{i,t} for all i ∈ A_t.
  Play arm I_t = arg max_{i∈A_t} max_{z∈C_t} z^T x_{i,t} with C_t as defined
  Observe r_t. Compute C_{t+1}
end
```

Regret bounds

- LinUCB [Auer 2002] With probability $1 - \delta$, regret

$$Regret(T) \leq \tilde{O}\left(d\sqrt{T}\ \right)$$

- Note : no dependence on number of arms

- Lower bound $\Omega\left(d\sqrt{T}\right)$

Proof outline

Thompson Sampling for linear contextual bandits

Linear regression

- Least square error solution $\widehat{\mu_t}$ of set of $t-1$ equations

$$x_s \cdot \mu = r_s, \quad s = 1, \dots, t-1$$

- $\widehat{\mu_t} \simeq B_t^{-1}(\sum_{s=1}^{t-1} x_s r_s)$ where $B_t = I + \sum_{s=1}^{t-1} x_s x_s'$
- B_t^{-1} covariance matrix of this estimator

Gaussian posterior

- $N(0, I)$ starting prior on μ,
- Reward distribution given μ, $x_{i,t}$: $N(\mu^T x_{i,t}, 1)$,
- posterior on μ at time t is $N(\hat{\mu}_t, B_t^{-1})$

Thompson Sampling for linear contextual bandits

[A., Goyal 2013] Algorithm:

At Step t,

- Sample $\tilde{\mu}_t$ from $N(\hat{\mu}_t,\ v^2 B_t^{-1})$
- Observe context x_t
- Pull arm with feature x_t where

$$x_t = \max_i x_{i,t} \cdot \tilde{\mu}_t$$

Regret bounds

- LinUCB [Auer 2002] With probability $1-\delta$, regret

$$Regret(T) \le \tilde{O}\left(d\sqrt{T}\ \right)$$

- Thompson Sampling [A. and Goyal 2013] With probability $1-\delta$, regret

$$Regret(T) \le \tilde{O}\left(d^{3/2}\sqrt{T}\ \right)$$

  - *Any* likelihood, *unknown* prior, only assumes bounded or sub-Gaussian noise

- Note : no dependence on number of arms

- Lower bound $\Omega\left(d\sqrt{T}\right)$

Many other contextual formulations

More general functions modeling expected reward on playing arm with context x

- Generalized linear bandits $\mathrm{g}(\mu^T x)$

[Filippi et al. 2010]

- Convex bandits: $f(x)$ for f convex in x

[Agarwal et al. 2011][Bubeck et al. 2015, 2016, 2017]

- Lipschitz bandits : $f(x)$ for Lipschitz function f on a metric space

[Kleinberg 2004] [Kleinberg et al. 2008] [Slivkins 2011] [Bubeck et al. 2011]

#2: Assortment selection as multi-armed bandit

- Consider arms as products
- Limited display space, k products displayed at a time
- Probability that customer choses product i from assortment S: $p_i(S)$
- Challenge: Customer response on one product is influenced by other products in the assortment
  - Feedbacks from individual arms are no longer independent

Flexion KS-901 Kinetic Series Wireless Bluetooth Headphones Noise Cancelling Headphones with Microphone/Running...

by Flexion

$29.99 ~~$49.99~~ Prime

Get it by **Tuesday, Mar 29**

More Buying Choices
$13.50 new (14 offers)
$22.24 used (1 offer)

1,512

FREE Shipping on eligible orders

Product Features
... bluetooth 4.0 + EDL in *headphones* 60% than the wireless ...

Sports & Outdoors: See all 60,247 items

Bluetooth Headphones, Liger MH770 High Quality Wireless Stereo Bluetooth 4.1 Sport Headphone with Magnetic Tips...

by Liger

$29.95 ~~$75.00~~ Prime

Get it by **Tuesday, Mar 29**

More Buying Choices
$29.95 new (2 offers)

56

FREE Shipping on eligible orders

Product Features
... MAGNET *HEADPHONES* DESIGN Hang like a necklace around your neck, and a ...

Electronics: See all 2,011,479 items

Liger BLAZE Bluetooth 4.1 Sweatproof Earbuds Noise Cancelling Headphones with Mic - Black

by Liger

$44.95 ~~$99.95~~ Prime

Get it by **Tuesday, Mar 29**

More Buying Choices
$44.95 new (4 offers)
$38.50 used (1 offer)

165

FREE Shipping on eligible orders

Electronics: See all 2,011,479 items

Panasonic ErgoFit In-Ear Earbud Headphones RP-HJE120-K (Black) Dynamic Crystal Clear Sound, Ergonomic Comfort-Fit

by Panasonic

$7.24 ~~$9.99~~ Prime

Get it by **Tuesday, Mar 29**

More Buying Choices
$2.33 new (139 offers)

30,424

FREE Shipping on eligible orders

Product Features
Black ultra-soft ErgoFit in-ear earbud *headphones* conform instantly to your ears

Electronics: See all 2,011,479 items


eal Time with Bill Maher:
eason 14
VICE
Game of Thrones: Sn 6
Silicon Valley: Sn 3

Customer choice modeling

Multinomial logit choice model [Luce 1959, McFadden 1978]

- Probability of choosing product i (feature vector x_i) in assortment S

$$p_i(S) = \frac{e^{\theta_i}}{1 + \sum_{j \in S} e^{\theta_j}}$$

- Probability of no purchase

$$p_i(S) = \frac{1}{1 + \sum_{j \in S} e^{\theta_j}}$$

- Key property: Independence of irrelevant alternatives
- Fixed reward r_i for product i
- Given a $\theta = (\theta_1, \theta_2, \dots, \theta_N)$, the optimal assortment is efficiently computable [Rusmevichientong et al. 2010] [Davis et al. 2013]

The MNL bandit problem

[A., Avadhanula, Goyal, Zeevi, 2016]

N products, Unknown parameters $\theta_1, \theta_2, \ldots, \theta_N$

At every step t,

- recommend an *assortment* S_t of size at most K,
- *observe customer choice* i_t, revenue r_{i_t}
- update parameter estimates

Goal:

- optimize total expected revenue $\mathrm{E}[\sum_{t=1}^{T} r_{i_t}]$
- or minimize regret compared to the optimal assortment $\mathrm{S}^* = \underset{\mathrm{S}}{\operatorname{argmax}} \sum_{i=1}^{N} r_i \, p_i(S)$

Main challenges and techniques

- Censored feedback
  - Feedback for product i effected by other products in assortment
  - Combinatorial choice: N^{K} possible assortments

[A., Avadhanula, Goyal, Zeevi, 2016, 2017]

- Technique to get unbiased estimate of individual parameters:
  - offer an assortment until no-purchase
  - Number of times i is purchased is unbiased estimate of its parameter e^{θ_i}
- Then, use standard UCB or Thompson Sampling techniques

Regret bounds

UCB based algorithm [A., Avadhanula, Goyal, Zeevi, 2016]

- $\tilde{O}(\sqrt{NT})$ regret bounds (under an assumption on no-purchase probability)
  - Parameter independent, no dependence on K
  - Matching lower bound of $\Omega(\sqrt{NT})$ [Chen and Wang 2017]

Thompson Sampling [A., Avadhanula, Goyal, Zeevi, 2017]

- Similar regret bounds, significantly more attractive empirical results

More recent work

- Contextual settings in [Chen et al. 2018][Ou et al 2018][Oh and Iyengar 2019]
- Nested logit models [Chen, Wang & Zhu, 2018]
- With resource constraints [Cheung & Simchi-Levi 2017]

#3: Bandits with constraints and non-linear aggregate utility

Regular bandits

- Total number of pulls constrained by T
  - No other global constraint on decisions across time
- Maximize sum of rewards

More global constraints

- Resource constraints in pricing and network revenue management
- Multiple Budget constraints in advertising campaigns
  - Nonlinear risk constraints
- Covering constraints in network routing and scheduling, sensor networks, crowdsourcing
- In pay-per-click advertising multiple performance criteria to be satisfied simultaneously
  - revenue, user satisfaction, diversity, minimum impressions

More than sum of rewards

- Smooth delivery in advertising
  - Minimize variance over time
- Demographics of clicks
  - maximizing minimum number of each type
- Nonlinear functions converting number of clicks to user satisfaction, or revenue
- Crowd sourcing: Need diversity among workers
- Sensor measurements: cover variety of locations
  - maximizing minimum number of successful sensor measurements from each location

Generalizing MAB

- Classic MAB
  - Observe reward r_t on pulling an arm i_t
  - Maximize $\sum_t r_t$

- Bandits with knapsacks (BwK) [Badanidiyuru, Kleinberg, Slivkins 2013, Besbes and Zeevi 2009, 2012]

  Observe non-negative reward r_t and cost vector $\boldsymbol{c}_t$

$$\text{maximize} \sum_t r_t$$

$$\text{s.t.} \sum_t \boldsymbol{c}_{t,j} \leq B, \forall j$$

Bandits with convex knapsacks and concave rewards (BwCR)

[Agrawal, Devanur 2014]

- Pulling an arm i_t generates a d dimensional vector $\boldsymbol{v}_t$, unknown mean V_{i_t}
- Total number of pulls constrained by T

+ Arbitrary convex global constraints on average of observations across time

$$\frac{1}{T}\sum_t \boldsymbol{v_t} \in S, \quad S \text{ is arbitrary convex set}$$

- Maximize arbitrary concave function $f\left(\frac{1}{T}\sum_t \boldsymbol{v}_t\right)$

  Minimize distance $\mathrm{dis}\left(\frac{1}{T}\sum_t \boldsymbol{v}_t, S\right)$ from convex set S

UCB like optimistic algorithm for BwCR

What is an optimistic estimate of the mean observation vectors?

- Need to estimate for every arm i and every coordinate j
- Non-decreasing f : upper bound (UCB)
  - The function value at the estimate will be more than actual
- Downward closed S: lower bound (LCB)
  - If actual mean is in S, the estimate will be in S
- In general
  - Most optimistic estimate in the confidence interval?

Optimistic algorithm for BwCR

- Play the (distribution over) arm that appears to be the best according to the most optimistic estimates in the confidence interval
  - Two levels of optimizations

- Actual mean lies in confidence intervals

$$H_t = \{\tilde{V} : \tilde{V}_{\mathrm{ij}} \in [\mathrm{LCB}_{t,ij}, \mathrm{UCB}_{t,ij}]\}$$

- Play best distribution over arms according to most optimistic estimate

$$\boldsymbol{p}_t = \arg\max_{\boldsymbol{p}} \max_{\tilde{\boldsymbol{V}} \in H_t} f\left(\sum_i p_i \tilde{V}_i\right)$$
$$\text{s.t.} \min_{\tilde{\boldsymbol{U}} \in H_t} dis\left(\sum_i p_i \tilde{U}_i, S\right) \le 0$$

Regret bounds

- [A. and Devanur 2014] UCB like optimistic algorithm that
  - achieves near-optimal average regret

$$\text{Regret in objective} \leq \tilde{O}\left(L\sqrt{N/T}\right), \qquad \text{Regret in constraints} \leq \tilde{O}\left(\sqrt{N/T}\right)$$

  - achieves problem specific optimal bounds on regret for Bandits with knapsacks

$$\text{Regret} \leq \tilde{O}\left(\text{OPT}\sqrt{N/B} + \sqrt{N\ \text{OPT}}\right)$$

  - is polynomial time implementable
- Recent Extensions to contextual bandits