Data Mining

CS57300
Purdue University

March 27, 2018

Recap last class

- So far we have seen how to:
  - Test a hypothesis in batches (A/B testing)
  - Test multiple hypotheses (Paul the Octopus-style)

The New York Times Daily Dilemma


"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

NEW YORK, TUESDAY, APRIL 16, 1912.—TWENTY-FOUR PAGES.

ONE CENT

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG; 866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH; ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

Women and Children Put Over in Lifeboats and Are Supposed to be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Vincent Astor Calls at White Star Office for News of His Father and Leaves Weeping.

FRANKLIN HOPEFUL ALL DAY

Manager of the Line Insisted Titanic Was Unsinkable Even After She Had Gone Down.

HEAD OF THE LINE ABOARD

J. Bruce Ismay Making First Trip on Gigantic Ship That Was to Surpass All Others.

Biggest Liner Plunges to the Bottom at 2:20 A. M.

RESCUERS THERE TOO LATE

Except to Pick Up the Few Hundreds Who Took to the Lifeboats.

WOMEN AND CHILDREN FIRST

Cunarder Carpathia Rushing to New York with the Survivors.

SEA SEARCH FOR OTHERS

The California Stands By on Chance of Picking Up Other Boats or Rafts.

OLYMPIC SENDS THE NEWS

Only Ship to Flash Wireless Messages to Shore After the Disaster.

LATER REPORT SAVES 866.

- Select 50% users to see headline A
  - Titanic Sinks
- Select 50% users to see headline B
  - Ship Sinks Killing Thousands
- Do people click more on headline A or B?
- If A much better than B we could do better…
- We refer to decision A or B as choosing an **arm**

Truth is…

- Sometimes we don’t only want to find whether hypothesis A is better than hypothesis B

- We really want to use the best-looking hypothesis at any point in time

- Deciding if H_0 : A = B or H_1 :A = B is irrelevant

Real-world Problem

- Web in perpetual state of feature testing

- Goal:
  Acquire just enough information about suboptimal arms to ensure they are suboptimal

$$X_k^{(i)} = \begin{cases} 1 & \text{, if } k\text{-th user seeing headline } i \text{ clicks} \\ 0 & \text{, otherwise} \end{cases}$$

(arm A) Titanic Sinks

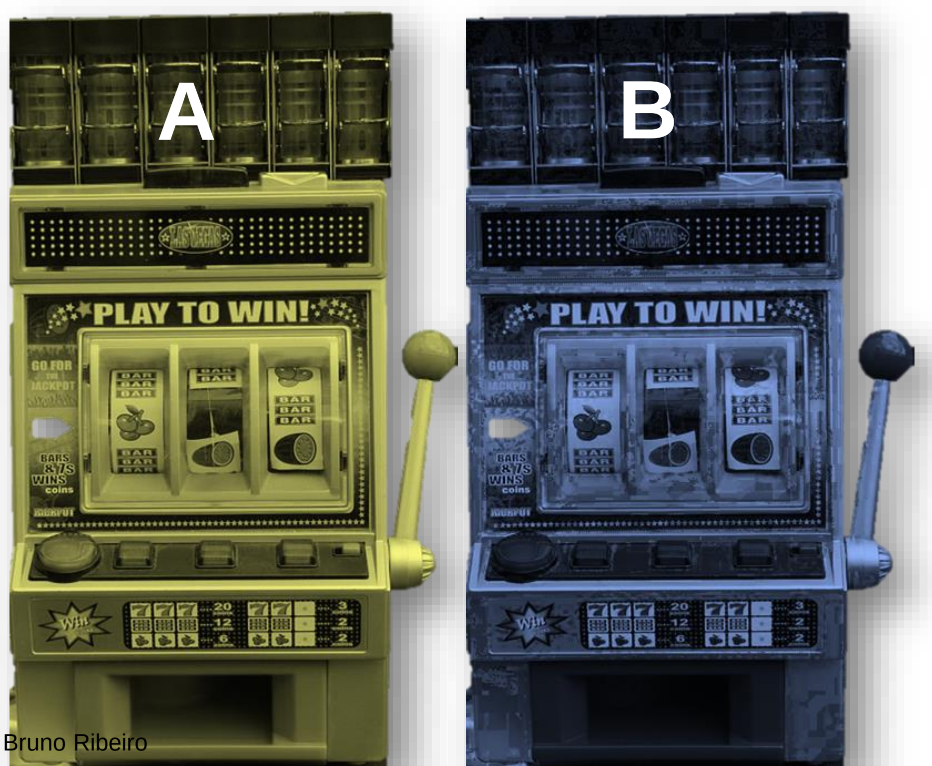
reward

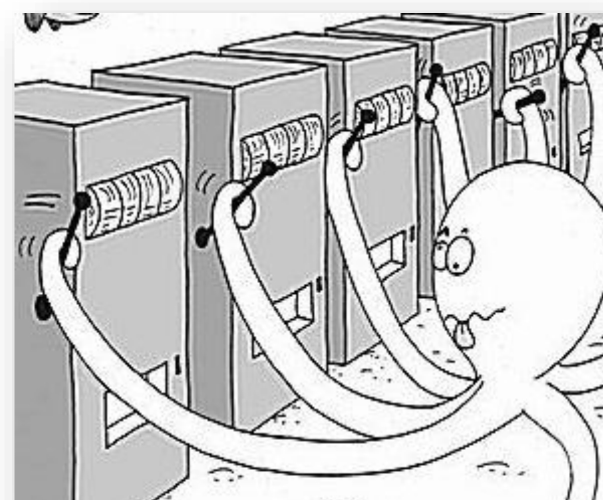
$$X_k^{(1)} = \begin{cases} 1 & \text{, with probability } p_1 \\ 0 & \text{, otherwise} \end{cases}$$

(arm B) Ship Sinks Killing Thousands

$$X_k^{(2)} = \begin{cases} 1 & \text{, with probability } p_2 \\ 0 & \text{, otherwise} \end{cases}$$

Multi-armed Bandits

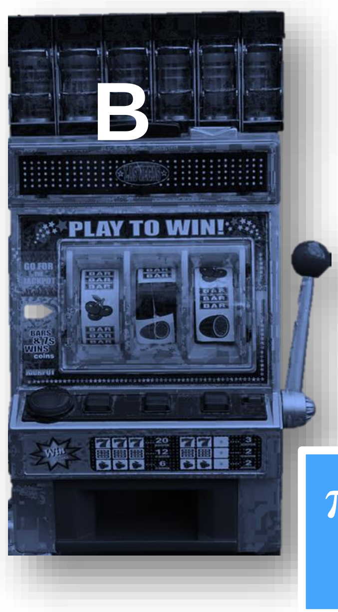




Multi-armed Bandit Dynamics

$$X_k^{(1)} = \begin{cases} 1 & \text{, with probability } p_1 \\ 0 & \text{, otherwise} \end{cases}$$

k-th time arm A is played



$$X_k^{(2)} = \begin{cases} 1 & \text{, with probability } p_2 \\ 0 & \text{, otherwise} \end{cases}$$

π is a vector of the arms we play
π_t is the t-th played arm

- Play t times
- Each time choose arm $i \in \{1,2\}$
- **Goal:** Maximize total expected reward

$$R_T = \sum_{t=1}^{T} X_{n_{\pi_t}(t)}^{(\pi_t)},$$

$$\text{where } n_i(t) = \sum_{h=1}^{t} \mathbf{1}\{\pi_h = i\}$$

Problem Characteristics

- Exploration-exploitation trade-off
  - Play arm with highest (empirical) average reward so far?
  - Play arms just to get a better estimate of expected reward?

- Classical model that dates back multiple decades [Thompson ‘33, Wald ‘47, Arrow et al. ‘49, Robbins ‘50, …, Gittins & Jones ‘72, ...]

Formal Bandit Definition

- $K \geq 2$ arms
- Pulling n_i times arm i produces rewards $X_1^{(i)}, \ldots, X_{n_i}^{(i)}$ with (unknown) joint distribution $f(x_1, \ldots, x_{n_i} | \theta_i)$, $\theta_i \in \Theta$
- At time $t \geq n_i(t)$ we know $X_1^{(i)}, \ldots, X_{n_i(t)}^{(i)}$, where $n_i(t)$ is the number of pulls of arm i at time t.
- Many formulations assume $X_1^{(i)}, \ldots, X_{n_i(t)}^{(i)}$ form a Markov chain

Markov chain: $P[X_k^{(i)} | X_{k-1}^{(i)}, X_{k-2}^{(i)}, \ldots] = P[X_k^{(i)} | X_{k-1}^{(i)}]$

- Most applications assume i.i.d. draws $P[X_k^{(i)} | X_{k-1}^{(i)}, \ldots] = P[X_k^{(i)}]$

Assumptions (can be easily violated in practice)

(A1) only one arm is operated each time

(A2) rewards in unused arms remain frozen

(A3) arms are independent

(A4) frozen arms contribute no reward

I.i.d. Stochastic Bandit Definition

- Assumption: Conditional Independence ($P[X_k^{(i)} | X_{k-1}^{(i)}, \ldots, \theta_i] = P[X_k^{(i)} | \theta_i]$)
- $K \geq 2$ arms
- Pulling n_i times arm i produces rewards $X_1^{(i)}, \ldots, X_{n_i(t)}^{(i)}$ i.i.d. with distribution $f(x|\theta_i)$, $\theta_i \in \Theta$
- At time $t \geq n_i(t)$ we know $X_1^{(i)}, \ldots, X_{n_i(t)}^{(i)}$

Relies on a model $f(x|\theta_i)$, $\theta_i \in \Theta$.

Depending on model can be more general than i.i.d. and Markov assumption.

Goal

Regret

$$R_t = \max_{j^\star=1,\dots,K} \sum_{h=1}^{t} X^{(j^\star)}_{n_{j^\star}(h)} - \sum_{h=1}^{t} X^{(\pi_h)}_{n_{\pi_h}(h)},$$

where π is the sequence of arm choices (or way to choose arms [policy]), $n_{\pi_h}(h)$ is the number of times we pull arm π_h at time h.

We can seek to minimize average regret

$$\min_{\pi} E[R_t]$$

Future actions depend
on past actions. Policy takes into
consideration past values

or minimized regret with high probability

$$P[R_t \geq \epsilon] \leq \delta$$

Regret Growth with i.i.d. Rewards

- Standard deviation of empirical $\sum_{k=1}^{n_i} X_k^{(i)}$ grows like $\sqrt{t}$
- Thus, at best $E[R_t] \propto \sqrt{t}$
- Rather, we minimize over π w.r.t. best policy (Pseudo-regret)

$$\bar{R}_t = \max_{i^\star=1,\ldots,K} E\left[\sum_{h=1}^{n} X_{n_{i^\star}(h)}^{(i^\star)} - \sum_{h=1}^{n} X_{n_{\pi_h}(h)}^{(\pi_h)}\right]$$

Optimal policy Chosen policy

Reward Definitions

- Mean reward $\mu_i = E[X_1^{(i)}]$
- Highest reward $\mu^\star = \max_{i^\star=1,\ldots,K} \mu_i$
- Reward gap: $\Delta_i = \mu^\star - \mu_i$

Lower Bound on Expected Pseudo-Regret

- Recall

$$\bar{R}_t = \max_{i^\star=1,\ldots,K} E\left[\sum_{h=1}^{t} X^{(i^\star)}_{n_{i^\star}(h)} - \sum_{h=1}^{t} X^{(\pi_h)}_{n_{\pi_h}(h)}\right]$$

$$= t \max_{i^\star=1,\ldots,K} \mu_{i^\star} - E\left[\sum_{h=1}^{t} X^{(\pi_h)}_{n_{\pi_h}(h)}\right]$$

$$= t \max_{i^\star=1,\ldots,K} - \sum_{k=1}^{K} E\left[n_k(t)\Delta_k\right]$$

- Asymptotically (Theorem 2, Lai & Robbins, 1985)

Valid for large values of n

$$E[n_i(t)] \gtrsim \frac{\log t}{D_{\mathrm{KL}}(f(x|\theta_i), f(x|\theta_{i^\star}))},$$

where D_{KL} is the KL divergence metric.

*The KL-divergence of two distributions can be thought of as a measure of their statistical distinguishability

Playing Strategies

Pure Exploration

- Algorithm for K arms
    - Play i with probability 1/K
- Pure exploration does not work well
    - Worst case: $E[R_t] \propto t$

Play-the-winner (Pure Exploitation)

- Algorithm
  - Let arm i be the arm with the maximum average reward at step t
  - Play i
- Play-the-winner does not work well
  - Worst case: $E[R_t] \propto t$

ε-greedy

- Assume rewards in [0,1]
- ε-greedy: at time t
  - with probability $1-\varepsilon_t$ play the best arm so far
  - with probability ε_t play random arm

- Theoretical guarantee (Auer, Cesa-Bianchi, Fischer 2002)
  - $\Delta = \min_{i:\Delta_i>0} \Delta_i$ and let $\epsilon_t = \min\left(\frac{12}{\Delta^2 t}, 1\right)$
  - If $t \geq \frac{12}{\Delta^2}$, the probability of choosing a suboptimal arm i is bounded by $\frac{C}{\Delta^2 t}$ for some constant $C > 0$
  - Then we have a logarithmic regret as $E[n_i(t)] \leq \frac{C}{\Delta^2} \log t$ and $R_t \leq \sum_{i:\Delta_i>0} \frac{C\Delta_i}{\Delta^2} \log t$

  - In practice we use larger values for Δ than the minimum reward gap (too conservative)

Problems of ε-greedy

- For K > 2 arms we play suboptimal arms with same probability
- Very sensitive to high variance rewards
- Real-world performance worst than next algorithm (UCB1)

Detour (Confidence Intervals)

Confidence Intervals

Confidence Interval: An interval of values computed from the observations, that covers the true value with X% probability.

Interpretation example: We are 95% confident that the true average will be in the interval

Computing Confidence Intervals

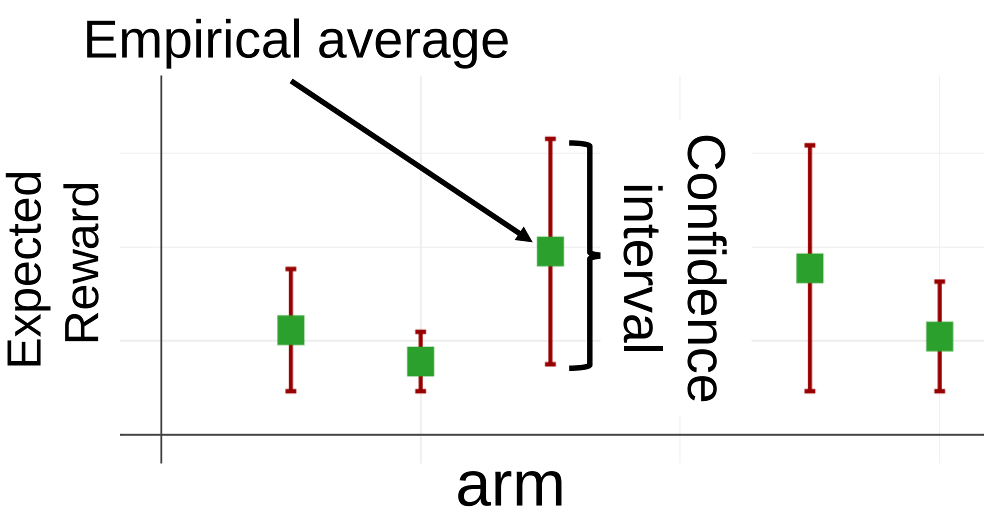
A simple way to compute confidence intervals is to assume observations come from a Normal distribution.

In a Normal distribution, we know that 95% of the observations are within about 2 standard errors* of the true average.

*empirical standard deviation of the empirical average (average of samples)

<u>Origin of the 2</u>:

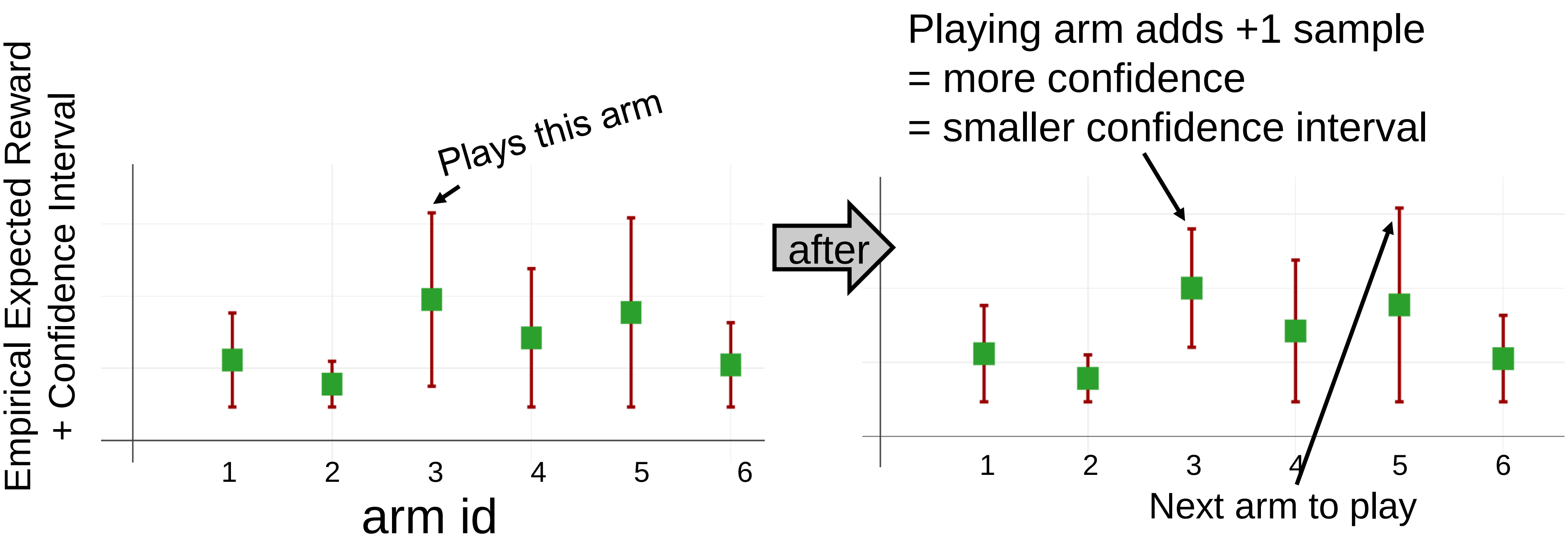
In a Normal distribution (actually, t-distribution) 95% of the data will be between +/- 1.96 standard deviations.



Back to Bandits

Optimism in Face of Uncertainty (UCB)

- Using a probabilistic argument we can provide an upper bound of the expected reward for arm $i=1,\ldots,K$ with a given level of confidence
- Strategy: play arm with largest upper bound
- Algorithm known as Upper Confidence Bound 1 (UCB1)



Using the Chernoff-Hoeffding Theorem

Let $X_1, \ldots, X_{n_i}$ be i.i.d. rewards from arm i with distribution bounded in $[0,1]$, then for any $\epsilon \in (0,1)$

$$P\left[\sum_{k=1}^{n_i} X_k \leq n_i E[X_1] - \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{n_i}\right)$$

Why?

$$P\left[\frac{1}{n_i(t)}\sum_{k=1}^{n_i(t)} X_k + \sqrt{\frac{2\log t}{n_i(t)}} \leq E[X_1]\right] \leq t^{-4}$$

- UCB1 algorithm
  - t total plays
  - Let $\epsilon = \sqrt{2 n_i(t) \log t}$
  - Gives algorithm:
    - Play arm i with largest $\displaystyle \frac{1}{n_i(t)}\sum_{k=1}^{n_i(t)} X_k + \sqrt{\frac{2\log t}{n_i(t)}}$

UCB 1 Regret Bound

- Each sub-optimal arm i is pulled on average at most

$$E[n_i(t)] \leq \frac{8 \log t}{\Delta_i^2} + \frac{\pi^2}{3}$$

times.

- Note that the MAB lower bound is O(log t)

Improving Bound ➡ Smaller Regret

- Use Empirical Bernstein's inequality

- Play arm i at time t if i has the maximim

$$\frac{1}{n_i(t)} \sum_{h=1}^{n_i(t)} X_h^{(i)} + \sqrt{\frac{2 \log t \, \mathrm{var}(X_1^{(i)}, \ldots, X_{n_i(t)}^{(i)})}{n_i(t)}} + \frac{8 \log t}{3 n_i(t)}$$

UCB for High-variance Distributions

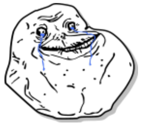
- Bubeck et al. gives more details on these “robust” versions of UCB.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, “Bandits with heavy tail,” Arxiv preprint arXiv:1209.1727, 2012.

Optimal Solution?

- Optimal solutions via stochastic dynamic programming
  - Gittins index

- Suffer from Incomplete Learning (Brezzi and Lai 2000, Kumar and Varaiya 1986)
  - Playing the wrong arm forever with non-zero probability



Forever Wrong

  - One more reason to be warry of average rewards

- Only applicable to infinite horizon & complex to compute

- Poor performance on real-world problems

Extra

Bayesian Bandits

- Thompson sampling
  - Strategy: select arm i according to posterior probability

$$P[\mu_i = \mu^\star | X_1^{(i)}, \ldots, X_{n_i}^{(i)}]$$

  - Can be used in complex problems (dependent priors, complex actions, dependent rewards)
  - Great real-world performance
  - Great regret bounds

Bernoulli Bandits

- Let $\mu_i \in (0, 1)$
- Reward of arm $i = 1, \ldots, K$ at step k is $X_k^{(i)} \sim \text{Bernoulli}(\mu_i)$

Thompson Sampling (1933)

- Strategy:
  - Uniform prior $\mu_i \sim U(0,1)$
  - Play arm i at time t as to maximize posterior $P[\mu_i = \mu^\star | X_1^{(i)}, \ldots, X_{n_i(t)}^{(i)}]$

This is the posterior probability that arm i has largest average

Bernoulli rewards + Beta priors

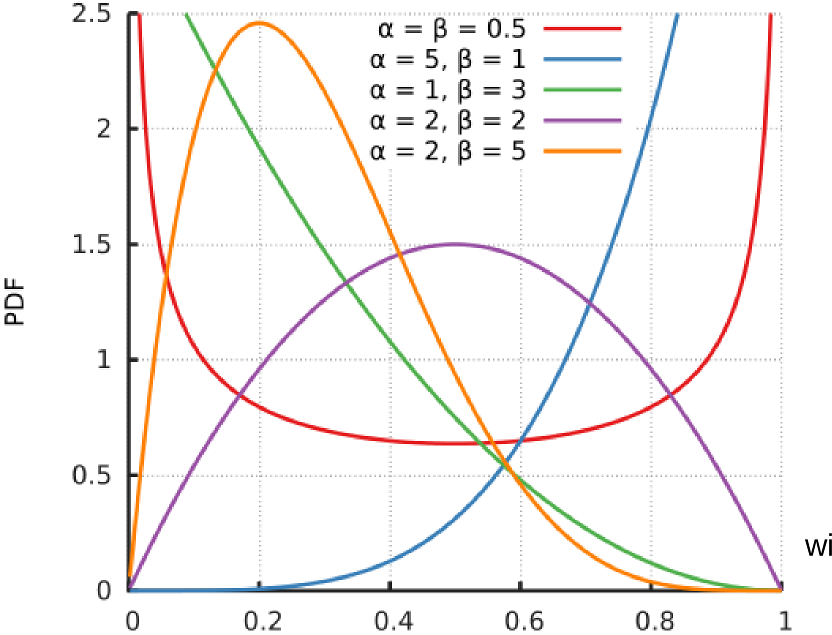
$$Y_t | I_t \sim \text{Bernoulli}(\mu_{I_t})$$

Prior: Beta distribution

$$P[\mu_i | \alpha, \beta] = \frac{\mu_i^{\alpha-1}(1-\mu_i)^{\beta-1}}{\int_0^1 p^{\alpha-1}(1-p)^{\beta-1} dp}$$

$$\text{Posterior } \mu_i \sim \text{Beta}(\alpha + \sum_{k=1}^t Y_k \mathbf{1}\{I_k = i\}, \beta + \sum_{k=1}^t (1 - Y_k)\mathbf{1}\{I_k = i\})$$

Beta distribution PDF



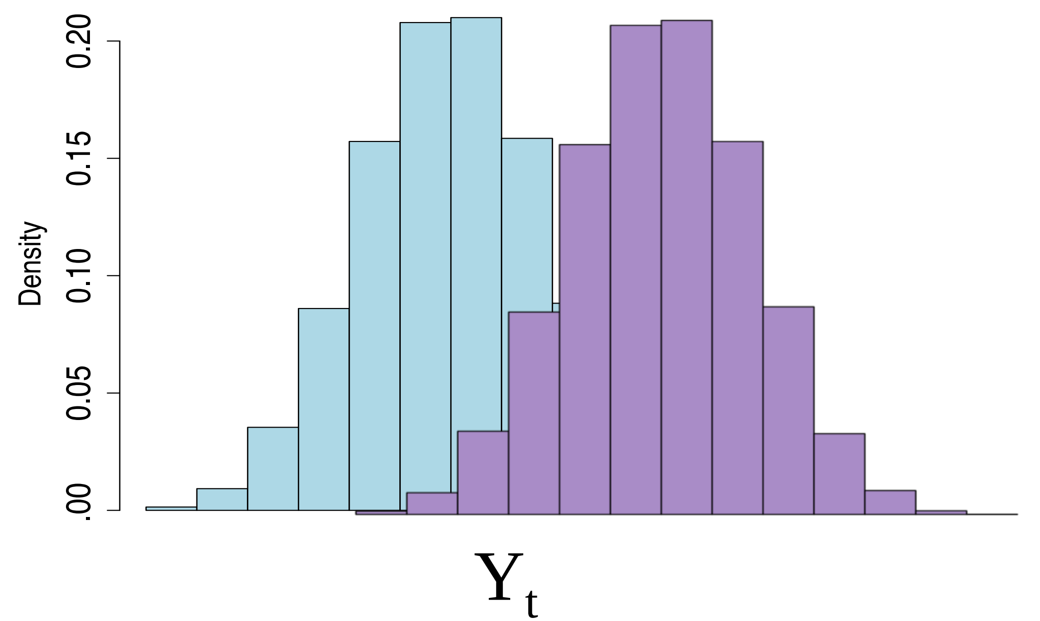
wikipedia

Thompson Algorithm (for Bernoulli rewards)

Prior arm i: $\mu_i \sim \text{Beta}(\alpha, \beta)$

$S_i = 0$; $F_i = 0$ // no. successes and failures of arm i

1. $\forall i$, draw $\hat{\mu}_i \sim \text{Beta}(S_i + \alpha, F_i + \beta)$
2. Choose arm $I_t = \arg\max_i \hat{\mu}_i$ and get reward Y_t
3. $S_{I_t} = S_{I_t} + Y_t$
4. $F_{I_t} = F_{I_t} + (1 - Y_t)$



TS: Bernoulli Reward Regret

- Theorem (Agrawal and Goyal, 2012)

For all $\mu_1, \ldots, \mu_K$ there is a constant C such that $\forall \epsilon > 0$,

$$\bar{R}_t \leq (1+\epsilon) \sum_{i:\Delta_i > 0} \frac{\Delta_i \log t}{\mathrm{D}_{\mathrm{KL}}(\mu_i, \mu^\star)} + \frac{Ck}{\epsilon^2}$$

Proof idea

- Posterior gets concentrated as more samples are obtained

Some Shortcomings of MAB

- Facebook & Linkedin use two-sample hypothesis tests instead of MAB. Why?

- More generally, which MAB assumptions often does not hold in real-life applications?

Assumptions most violated in practice

(A2) rewards in arms not used remain froze

  Recommendation influence users

(A3) arms are independent

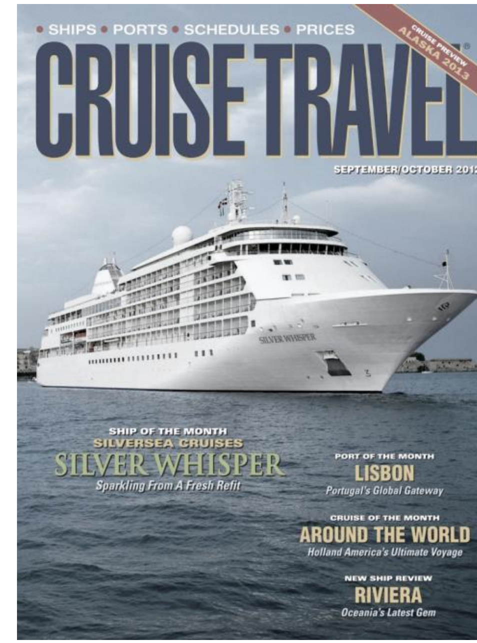
  Users are influenced by what they see and by each other

Advanced Topic

(Contextual Bandits)

Contextual Bandits

- Bandits with side information
- We know reader subscribes to magazine
- Headline A may be more successful in this subpopulation
  - Titanic Sinks
- Headline B better for general population
  - Ship Sinks Killing Thousands



"All the News That's Fit to Print."

The New York Times.

THE WEATHER.

TITANIC SINKS FOUR HOURS AFTER HITTING ICEBERG;
866 RESCUED BY CARPATHIA, PROBABLY 1250 PERISH;
ISMAY SAFE, MRS. ASTOR MAYBE, NOTED NAMES MISSING

Col. Astor and Bride, Isidor Straus and Wife, and Maj. Butt Aboard.

"RULE OF SEA" FOLLOWED

Women and Children Put Over in Lifeboats and Are Supposed to be Safe on Carpathia.

PICKED UP AFTER 8 HOURS

Vincent Astor Calls at White Star Office for News of His Father and Leaves Weeping.

FRANKLIN HOPEFUL ALL DAY

Manager of the Line Insisted Titanic Was Unsinkable Even After She Had Gone Down.

HEAD OF THE LINE ABOARD

Biggest Liner Plunges to the Bottom at 2:20 A. M.

RESCUERS THERE TOO LATE

Except to Pick Up the Few Hundreds Who Took to the Lifeboats.

WOMEN AND CHILDREN FIRST

Cunarder Carpathia Rushing to New York with the Survivors.

SEA SEARCH FOR OTHERS

OLYMPIC SENDS THE NEWS

LATER REPORT SAVES 866.

Contextual Bandits: Problem Formulation

- Consider a hash (random function) with deterministic projection $h:\{0,1\}^n \rightarrow \mathbb{R}^m$
- At each play:

1. Observe features $X_t \in \mathcal{X}$
2. Choose arm $i_t \in \{1, \ldots, K\}$
3. In theory we will get reward $Y_t = f(h(x_t, z_{i_t})|\theta) + \epsilon_t$

Learned from observations
(e.g. recursive least squares)

User features

Headline
features (context)

some useful assumptions about f

- $f(x|\theta) = \theta^{\mathsf{T}} x$ (linear bandit)
- $f(x|\theta) = g(\theta^{\mathsf{T}} x)$ (generalized linear bandit)

Contextual Bandits (linear model)

- First we build model

$$
\begin{array}{ccccc}
\vec{Y}_t & & \mathbf{X}_t & & \\
\begin{bmatrix} y_1 \\ \vdots \\ y_t \end{bmatrix} & = & \begin{bmatrix} h(x_1^{\mathtt{T}}, z_1^{\mathtt{T}}) \\ \vdots \\ h(x_t^{\mathtt{T}}, z_t^{\mathtt{T}}) \end{bmatrix} & \theta + & \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_t \end{bmatrix}
\end{array}
$$

We can estimate $\hat{\theta}_t$ using a regularized least-squares estimate of θ at time t

$$\hat{\theta}_t = (\lambda I + \mathbf{X}_t^{\mathtt{T}} \mathbf{X}_t)^{-1} \mathbf{X}_t^{\mathtt{T}} \vec{Y}_t \,,$$

$\lambda > 0$

Thompson Sampling for Linear Contextual Bandits 1/2

Assume noise is Gaussian

$$\epsilon_t \sim N(0, \sigma^2)$$

and that θ has prior

$$\theta \sim N(0, \kappa^2 I)$$

The posterior distribution of θ is given by

$$p(\theta | \mathbf{X}_t, \vec{Y}_t) = N(\hat{\theta}_t, \Sigma_t)$$

where

$$\Sigma_t = \lambda I + \mathbf{X}_t^{\mathsf{T}} \mathbf{X}_t,$$

where $\lambda = \frac{\sigma^2}{\kappa^2}$.

Thompson Sampling for Linear Contextual Bandits 2/2

Thompson Sampling heuristic:

$$\tilde{\theta}_t \sim N(\hat{\theta}_t, \Sigma_t)$$

and obtain best arm

$$i^\star = \underset{i \in \{1,\ldots,K\}}{\arg\max}\; h(x_{t+1}, z_i)\tilde{\theta}_t$$

The above draws each context ∝ posterior probability of being optimal

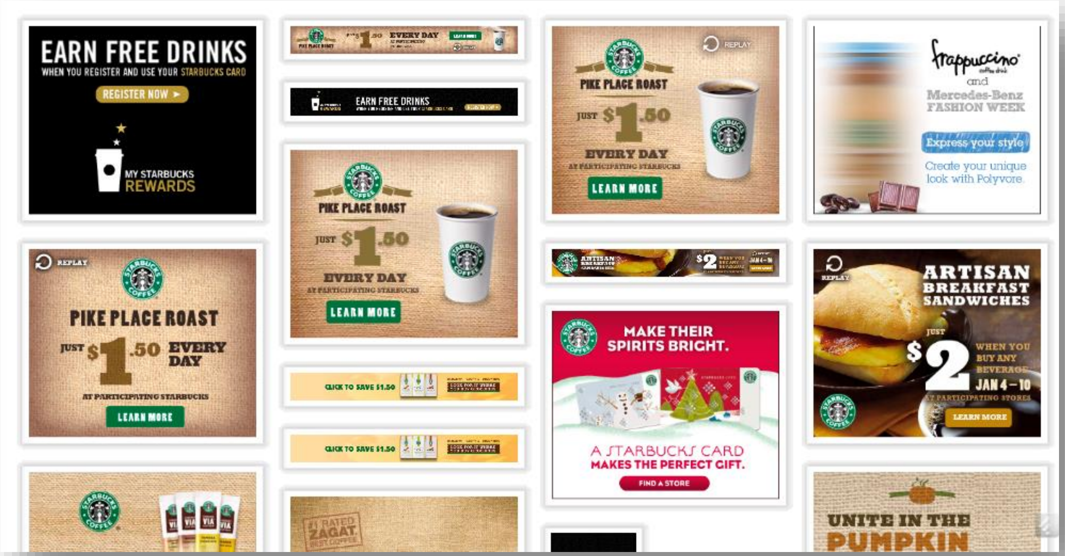
- From [Russo, Van Roy 2014] pseudo-regret is

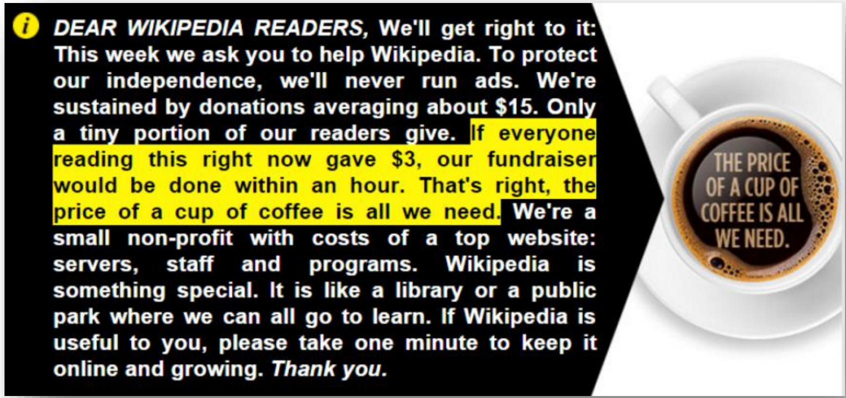
$$\bar{R}_t = \tilde{O}(d\sqrt{t})$$

PS: $\tilde{O}$ ignores logarithmic factors

Example

Effectiveness of ad may depend on consumer





Response Prediction for Display Advertising

- Example:
  - Chapelle et al. (2012)

- The features used:
  - Ω_{t+1} set of sparse binary entries
  - Concatenate categorical features of user with features of all bandits (ad, headline)
  - Use hash function h() maps from categorical space to lower dimensional $\mathbb{R}^m$ space

Algorithm for Display Advertising

Goal: maximize the number of clicks or conversions

Model: Logistic regression

$$P[Y_t = 1 | x_t, z_{I_t}, \theta] = \frac{1}{1 + \exp(-\theta^{\mathtt{T}} h(x_t, z_{I_t}))}$$

Response prediction based on training set $\Omega'_t = \{(x_k, z_{I_k}, y_k)\}_{k=1,\ldots,t}$

$$\hat{\theta} = \arg\min_{\alpha \in \mathbb{R}^m} \frac{\lambda}{2} \|\alpha\|^2 + \sum_{k=1}^{t} \log(1 + \exp(-y_k \alpha^{\mathtt{T}} h(x_k, z_{I_k})))$$

Display Ads (cont)

If prior $\theta \sim N(0, \frac{1}{\beta} I)$, the posterior $P[\theta|D]$ has no closed form expression but we can use the Laplace approximation of the integral

$$P[\theta|D] = N(\hat{\theta}, \mathrm{diag}(q_i)^{-1})$$

where

$$q_i = \sum_{j=1}^{t} w_{j,i}^2 p_j (1 - p_j) \qquad \text{with } p_j = (1 + \exp(-\hat{\theta}^{\mathsf{T}} h(x_j, z_{I_j})))^{-1}$$

and $w_j = h(x_j, z_{I_j})$

Using Thompson Sampling Algorithm for Ad Display

1. A new user arrives at time $t+1$

2. Form the set $\Omega_{t+1} = \{(x_t, z_i) : i \in \{1, \ldots, K\}\}$ of context corresponding to the different items that can be recommended to user

3. Sample vector from the current (approximate) posterior

$$\tilde{\theta}_t \sim N(\hat{\theta}, \mathrm{diag}(q_i)^{-1})$$

4. Choose the context (x_t, z_i) that maximizes probability of positive response according to

$$i = \arg\max_{i=1,\ldots,K} \frac{1}{1+\exp(-\tilde{\theta}_t^{\mathsf{T}} h(x_t, z_i))}$$

5. Recommend item and get response Y_{t+1}