Multi-Armed Bandits for Minimizing Regret

David Kriegman

San Diego Women in Data Science
November 10, 2016

Computer Science & Engineering

UCSD

Machine Learning Dropbox

Kriegman Group

Computer Science & Engineering
University of California, San Diego

- Cameras looking at people
- Vision for biology and ecology
- Reflectance and lighting for vision and graphics

RECOGNIZING URBAN TRIBES:

Microgroups of people who share common interests in metropolitan areas. The members of these relatively small groups tend to have similar worldviews, dress styles and behavioral patterns

- Maffesoli 1985

Bikers

Surfers

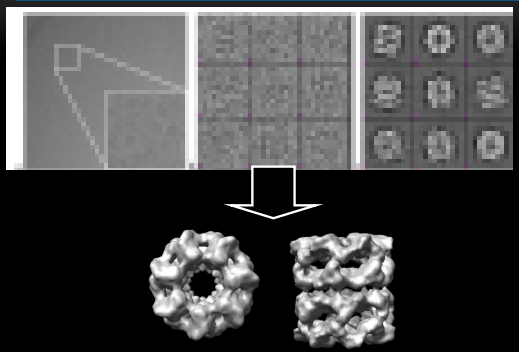
Some Tribes

Anarcho-punk, biker, cyber-goth, clubber, cosplayer, emo, goth, heavy metal, hiphop, hippie, hipster, juggalo, lolita, punk, skater, surfer

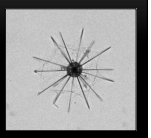
VISION FOR BIOLOGY AND ECOLOGY

FROM SMALL TO BIG

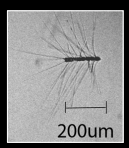
SMALLEST: RECONSTRUCTING PROTEIN MACROMOLECULES FROM CRYO-ELECTRON MICROSCOPY



SMALL: THE KECK 3D PLANKTON MICROSCOPE



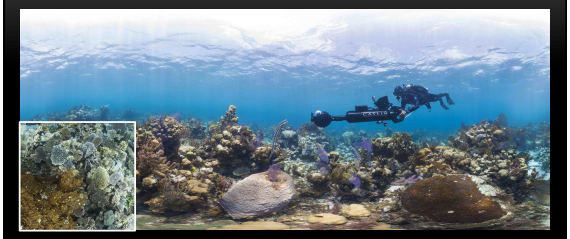
Acantharian



Diatom

- *Underwater, in situ microscope with 2 micron resolution, 1000 micron field of view*
- *Tracking, recognizing, understanding the behavior of hundreds micro plankton.*

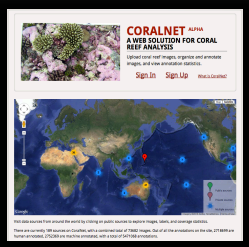
BIGGER: CORAL REEF ECOLOGY



- *Great Barrier Reef is 344.000 km²*
- *How can we monitor such vast areas?*

CORALNET (CORALNET.UCSD.EDU)

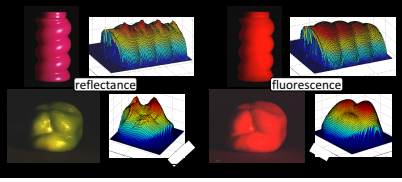
- Online random point annotation tool.
- Freely available.
- Stats
  - 300k images
  - 402 surveys
  - 443 users
  - 6.3 Million *manual* point annotations.
- Funded by NSF and NOAA



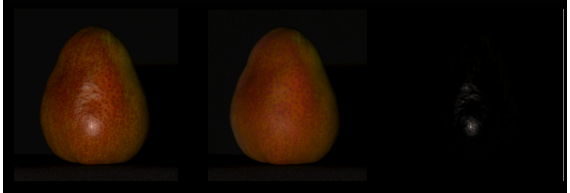
REFLECTANCE AND LIGHTING FOR VISION AND GRAPHICS

RESEARCH IN REFLECTANCE AND LIGHTING

- Photometric stereo
- Shape from fluorescence
- BRDF's
- Shape reconstruction in turbid media
- Estimating lighting in natural images
- Photorealistic augmented reality using learned appearance models



SPECULAR SEPARATION AND EDITING

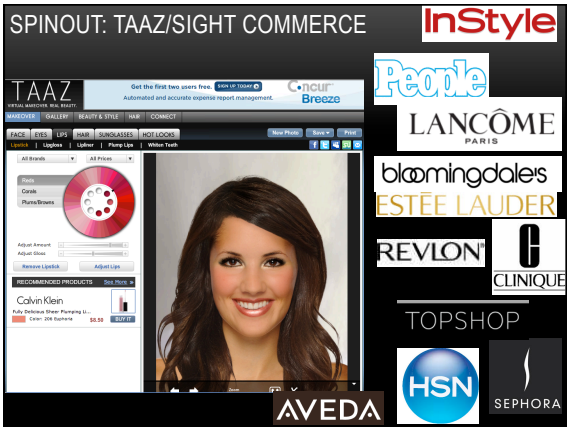


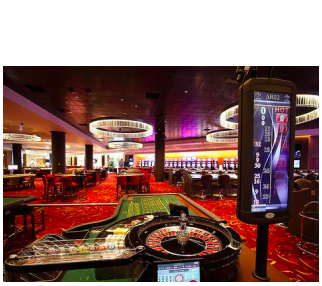
VIRTUAL MAKEUP: FOUNDATION AND LIP GLOSS

Before

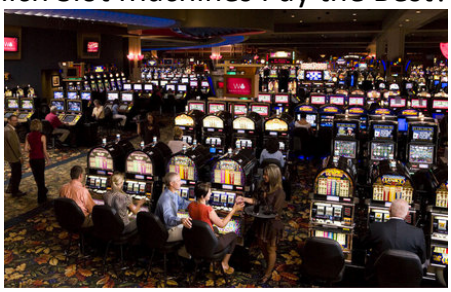
After Foundation and Lip Glosn

SPINOUT: TAAZ/SIGHT COMMERCE





Which Slot Machines Pay the Best?

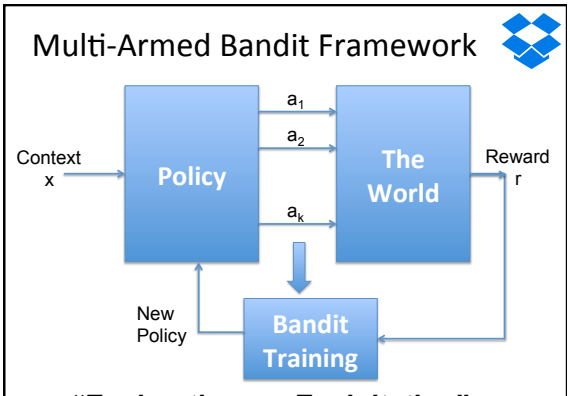


http://www.gamblersbookcase.com/How-to-Find-the-Best-Slot-Machine.htm

The location within the casino is also important for many players. Some players claim that aisle machines are the best, while others believe that the loosest slots are hidden in back corners so that they don't get much play.

"The **multi-armed bandit problem** is the problem a gambler faces at a row of slot machines when deciding which machines to play, how many times to play each machine and in which order to play them. When played, each machine provides a random reward from a distribution specific to that machine. The objective of the gambler is to maximize the sum of rewards earned through a sequence of lever pulls."

Multi-Armed Bandit Framework



"Exploration vs. Exploitation"

Context-free bandits

- A={a_1,…,a_k} : Set of k actions
- There exists some unknown distribution D of rewards over the k actions
- T trials, t = 1,2,…,T
- At trial t, the bandit policy chooses arm a_t and receives a payoff $r_{t,at}$
- Note, we don't learn anything about payoff for unchosen arms
- After each round, use <a_t, $r_{t,at}$> to update policy

Let's start real simple – 1 bandit

- For T trials, we get a sequence of rewards R={r_1,…,r_T}
- We can compute the mean reward μ = E[R]
- For binary payoffs (win or lose, ad clicked, payment succeeds), Bernoulli trials
- Confidence Interval: $\mu \pm k\sqrt{\frac{1}{T}\mu(1-\mu)}$
- For 95% confidence, k=1.96

Multi-arm Bandit Algorithms

- A/B testing
- ε-Greedy
- Upper Confidence Bound
- Thompson Sampling
- Many, many other algorithms and variants.

A/B Testing

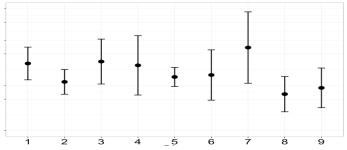
- For the first N trials, randomly select amongst the k actions and record reward.
- Identify the action **a*** with the highest average reward.
- Always use **a*** after the N trials.
- N is chosen in advanced based on being 95% confident **a*** is actually the better than the alternatives.

Epsilon Greedy

- Choose epsilon 0 ≤ ε ≤ 1 (often 0.01 or 0.1)
- At each trial, identify the action with the best mean reward and use 1-ε of the time.
- For the other ε fraction of trials, pick the action randomly.
- A the number trials approaches infinite, guaranteed to find optimal action.

Upper Confidence Bound (UCB)

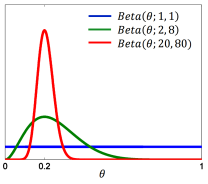
- For each action a_i, compute :
  - the mean reward reward r_i
  - the upper confidence bound UCB_i – It's a function of mean reward received for the action and the number of times the action was applied.
- Pick the action with the highest UCB.
- Different ways to compute confidence interval



Thompson Sampling

- Bayesian approach: For each action, maintain P(expected reward | samples of that action)
- Beta distribution for binary bandits

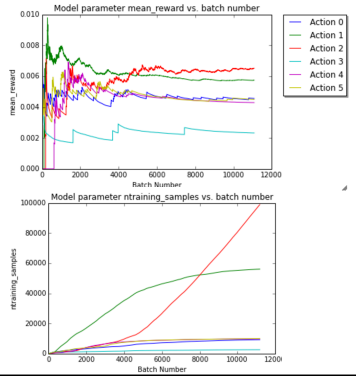
$$Beta(\theta:\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \Pr(\theta)$$



Algorithm 1 Thompson Sampling for Bernoulli bandits

For each arm $i = 1, \ldots, N$ set $S_i = 0, F_i = 0$.
foreach $t = 1, 2, \ldots,$ **do**
  For each arm $i = 1, \ldots, N$, sample $\theta_i(t)$ from the Beta$(S_i + 1, F_i + 1)$ distribution.
  Play arm $i(t) := \arg\max_i \theta_i(t)$ and observe reward r_t.
  If $r = 1$, then $S_{i(t)} = S_{i(t)} + 1$, else $F_{i(t)} = F_{i(t)} + 1$.
end

6 Armed Bandit – Thompson Sampling

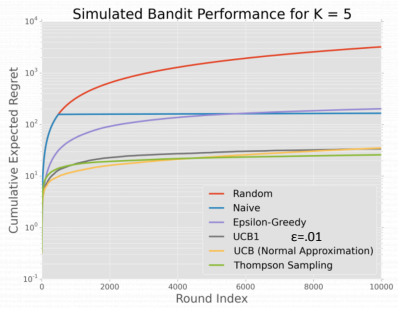


Regret

re·gret
/rəˈgret/
verb
1. feel sad, repentant, or disappointed over (something that has happened or been done, especially a loss or missed opportunity).
"she immediately regretted her words"
synonyms: be sorry about, feel contrite about, feel remorse about/for, be remorseful about, rue, repent (of), feel repentant about, be regretful at/about
"they came to regret their decision"

- Best action: a^*
- Reward for best action r_{t,a^*}
- Reward for best action over T trials $\sum_{t=1}^{T} r_{t,a^*}$
- Lots of variation from Trial to trial, so really care about expected (average) value
- T-trial Regret: $R(T) = E\left[\sum_{t=1}^{T} r_{t,a^*}\right] - E\left[\sum_{t=1}^{T} r_{t,a_t}\right] r$

How do bandit algorithms perform



From Blog: Sergey Feldman, Bandits for Recommendation Systems
http://engineering.richrelevance.com/bandits-recommendation-systems/

Real World Considerations

1. Carefully identify reward
2. Rewards may not be immediately observed and there could be significant delays (days).
3. Many actions may be taken before reward for even first action is observed.
4. Associating rewards to specific actions.
5. Batches – policy updating in batches, rather than one at a time.

Some applications

- Web site or mobile app optimization and experiments (Google content experiments, Microsoft Multiworld Testing Service)
- Email campaign optimization
- Content or advertisement selection and placement (Yahoo)
- Product recommendations (Rich Relevance)
- Clinical trials

Bandits vs. A/B Testing

- Advantage of Multi-armed Bandits
  - Generally handle larger number of variants
  - Converges to optimal policy
  - Lower regret
  - Set and forget
  - Time varying rewards
  - Often faster to reach significance
- Advantage of A/B testing
  - All exploration up front. Making product/strategic decisions.

Let's add context, x

Let $\mathcal{X}$ be an arbitrary input space, and $\mathcal{A} = \{1, \ldots, k\}$ be a set of actions. An instance of the *contextual bandit problem* is specified by a distribution D over tuples $(x, \vec{r})$ where $x \in \mathcal{X}$ is an input and $\vec{r} \in [0, 1]^k$ is a vector of rewards [6]. Events occur on a round-by-round basis where on each round t:

1. The world draws $(x, \vec{r}) \sim D$ and announces x.
2. The algorithm chooses an action $a \in \mathcal{A}$, possibly as a function of x and historical information.
3. The world announces the reward r_a of action a, but not $r_{a'}$ for $a' \neq a$.

From Strehl, Langford, Li, Kakade, "Learning from Logged Implicit Exploration Data"

- Simple solution: Treat as set of "context-free" bandit problems, conditioned on **x**.
- That is for each possible feature value, we can do things like estimate the expected payoff $Q_t(a|x)$, the UCB, etc.
- OK, if number of possible features is relatively small.

Contextual Bandits with Linear Payoff Functions

Using the notation of Section 2.1, we assume the expected payoff of an arm a is linear in its d-dimensional feature $\mathbf{x}_{t,a}$ with some unknown coefficient vector $\boldsymbol{\theta}_a^*$; namely, for all t,

$$\mathbf{E}[r_{t,a}|\mathbf{x}_{t,a}] \quad = \quad \mathbf{x}_{t,a}^\top \boldsymbol{\theta}_a^*. \qquad (2)$$

Called Disjoint model because parameters theta_a are not shared between different arms.

- Contextual Bandits with Linear Payoff Functions, W. Chu, L. Li, L. Reyzin, R. Schapire, AISTATS 2011

Estimate Parameters with Ridge Regression

Let $\mathbf{D}_a$ be a design matrix of dimension $m \times d$ at trial t, whose rows correspond to m training inputs (*e.g.*, m contexts that are observed previously for article a), and $\mathbf{b}_a \in \mathbb{R}^m$ be the corresponding response vector (*e.g.*, the corresponding m click/no-click user feedback). Applying ridge regression to the training data $(\mathbf{D}_a, \mathbf{c}_a)$ gives an estimate of the coefficients:

$$\hat{\boldsymbol{\theta}}_a = (\mathbf{D}_a^\top \mathbf{D}_a + \mathbf{I}_d)^{-1} \mathbf{D}_a^\top \mathbf{c}_a, \qquad (3)$$

where $\mathbf{I}_d$ is the $d \times d$ identity matrix.

Upper Confidence Bound on Payoff

When components in $\mathbf{c}_a$ are independent conditioned on corresponding rows in $\mathbf{D}_a$, it can be shown [27] that, with probability at least $1 - \delta$,

$$\left| \mathbf{x}_{t,a}^\top \hat{\boldsymbol{\theta}}_a - \mathbf{E}[r_{t,a}|\mathbf{x}_{t,a}] \right| \leq \alpha \sqrt{\mathbf{x}_{t,a}^\top (\mathbf{D}_a^\top \mathbf{D}_a + \mathbf{I}_d)^{-1} \mathbf{x}_{t,a}} \qquad (4)$$

for any $\delta > 0$ and $\mathbf{x}_{t,a} \in \mathbb{R}^d$, where $\alpha = 1 + \sqrt{\ln(2/\delta)/2}$ is a constant.

LinUCB Policy: Choose action as

$$a_t \stackrel{\text{def}}{=} \arg\max_{a \in \mathcal{A}_t} \left(\mathbf{x}_{t,a}^\top \hat{\boldsymbol{\theta}}_a + \alpha \sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}} \right)$$

where $\mathbf{A}_a \stackrel{\text{def}}{=} \mathbf{D}_a^\top \mathbf{D}_a + \mathbf{I}_d$.

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $\alpha \in \mathbb{R}_+$
for $t = 1, 2, 3, \ldots, T$ **do**
  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$
  for all $a \in \mathcal{A}_t$ **do**
    if a is new **then**
      $\mathbf{A}_a \leftarrow \mathbf{I}_d$ (d-dimensional identity matrix)
      $\mathbf{b}_a \leftarrow \mathbf{0}_{d\times 1}$ (d-dimensional zero vector)
    end if
    $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1}\mathbf{b}_a$
    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha\sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1}\mathbf{x}_{t,a}}$
  end for
  Choose arm $a_t = \arg\max_{a\in\mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_t
  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t}\mathbf{x}_{t,a_t}^\top$
  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t\mathbf{x}_{t,a_t}$
end for

Regret Bounds

Context Free Bandits

- Epsilon Greedy. Per trial regret converges to epsilon. Idea decay epsilon.
- UCB. Total regret grows as O(log T). Optimal

Contextual Bandits

- Epoch greedy: Total regret grows as $O(T^{2/3})$
- EXP4, LineRel: O(sqrt(T))
- LinUCB: Total regret grows as O(sqrt(KdT))
  - K: # actions
  - d: dimension of feature space x
  - T : number of samples

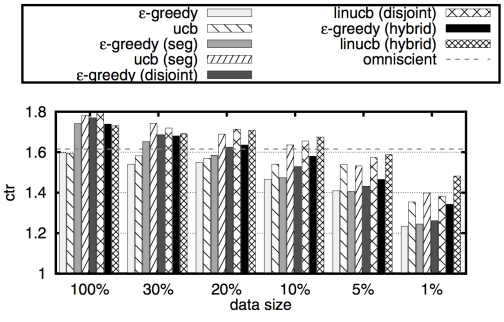
Application: Today Module on Yahoo! Front Page

- 4.7M events used for parameter setting (one day)
- 36M events from 7 days used for evaluation
- Action: Specific article from article pool to place in F1/Story. Number of actions is size of pool
- Payoff: 1 if user clicks on article F1, 0 otherwise

Raw Features

- User: 1193 features
  - Gender (2 classes)
  - Age (discretized to 10 segments)
  - Geography (200 metro locations worldwide)
  - Behavioral cateogires (1000 binary features based on past consumption history on Yahoo! Properties).
- Article: 83 features
  - URL categories –inferred from URL
  - Editorial categories

Click-Through Rate Results



More to explore

- Practical implementations and applications
- Evaluating on logged data
- Time varying rewards
- Contextual Bandits – beyond linear (random forests, deep nets)
- Very large (continuous) action spaces
- Adversarial bandits
- Constrained bandits (actions have costs)
- Reinforcement learning

Some References

- Lihong Li, Wei Chu, John Langford, Robert Schapire, **A Contextual-Bandit Approach to Personalized News Article Recommendation Systems,** WWW 2010.
- S. Bubeck, N. Cesa-Bianchi, **Regret Analysis of Stochastic and Nonstochastic multiarmed bandits**
- Microsoft's Multiworld Testing Service, https://www.microsoft.com/en-us/research/project/multi-world-testing-mwt/
- Sergey Feldman, Rich Relevance Blog http://engineering.richrelevance.com/bandits-recommendation-systems