

□

.

# Bandit Learning with Biased Human Feedback

Wei Tang, Chien-Ju Ho  
Washington University in St. Louis

# Multi-armed Bandit learning

# Slot Machines



...



# Multi-armed Bandit learning

- $T$  rounds, in each round, choose a slot machine/arm to pull
- IID Rewards: each arm reward is IID drawn from unknown distribution
- Bandit feedback: observe only the reward of your choice
- Goal:
  - Maximize the cumulative reward
  - Minimize regret  $R(T) = OPT - ALG$ 
    - No-regret learning  $R(T) = o(T)$

## Exploration vs Exploitation

# Bandit learning with humans in the loop

- In the literature
  - Arms can be strategically selected by the myopia users
    - “External” incentives: *monetary payments*. FKKK EC’14,
    - “Intrinsic” incentives: *information asymmetry*. YAVW EC’15 EC’16 , KG  
Econometrica’11, KG AER’14
  - Arms can strategically reporting their rewards
    - Treat each arm as a strategic agent. BMS COLT’19
- In this work, we consider **biased signal of unobservable reward**

# User-generated content System

YouTube search results for "arizona":

- ARIZONA - Oceans Away (3:16)
- ARIZONA - CROSS MY AUDIO (3:38)
- AMAZING Phoenix Arizona Donut I've Ever Had! (8:45)
- Carne Asada and Camping (18:06)

Quora question: What is your PhD thesis in one sentence?

Richard Peng, Assistant Professor at Georgia Institute of Technology (2015-present)  
Answered Jul 23, 2014 · Upvoted by Jessica Su, CS PhD student at Stanford and Karthik Abinav, PhD student in Computer Science from UMD

Viewing graphs as matrices lets you play with them as if they're positive real numbers, and leads to some really fast (parallel) algorithms for classical problems.

For reference: Algorithm Design Using Spectral Graph Theory

12.2k Views · 119 Upvotes

Abhinav Maurya  
Hasn't this been a theme in computer science research for a while now? Any addition...

Scott E. Fahlman, Professor Emeritus, Carnegie Mellon, LTI and CSD  
Answered Thu · Upvoted by Jessica Su, CS PhD student at Stanford and John L. Miller, B.S., M.S., PhD with 25 years industry experience

A2A: My PhD thesis (and much of my subsequent career in AI) has been focused on a single question: How is it that I can tell you one thing — “Clyde is an elephant” — and suddenly, without any apparent effort, you now know so much about Clyde?

It's harder than it sounds...

18.6k Views · 494 Upvotes · Answer requested by Fatma Faruq

Upvote | 494 · Downvote

Scott E. Fahlman  
Several commenters have asked about this. A scanned but more-or-less readable versl...

1,504,905 views

42K 1K

12.2k Views · 119 Upvotes

# User-generated content System

- When each new user arrives
  - Show the user some (set of) content
  - Obtain **feedback** (upvotes, likes, shares, etc) from the user
- Goal:
  - Maximize the **total user's happiness**
- A standard bandit learning problem
  - Arm: the content chosen to show to users

Feedback = happiness?

Answer · Marketing Strategy X

**What are some examples of great marketing?**

 Shenal, IT Student, Teenager  
Answered May 5 · Upvoted by Saloni Bhargava, MBA Marketing (2018)

Martial arts school Tattoo artist wanted Mondo Pasta  
BBC World Service 3M Security Glass Livegreen  
Toronto ... [\(more\)](#)

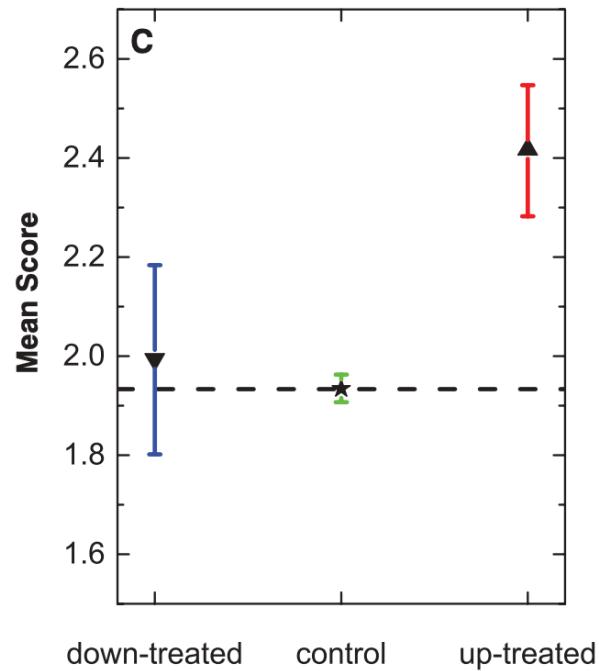


# Users' feedback might be biased

- Social Influence Bias: In a Reddit-like platform, randomly insert an upvote to some posts right after they are posted.



Herding effect

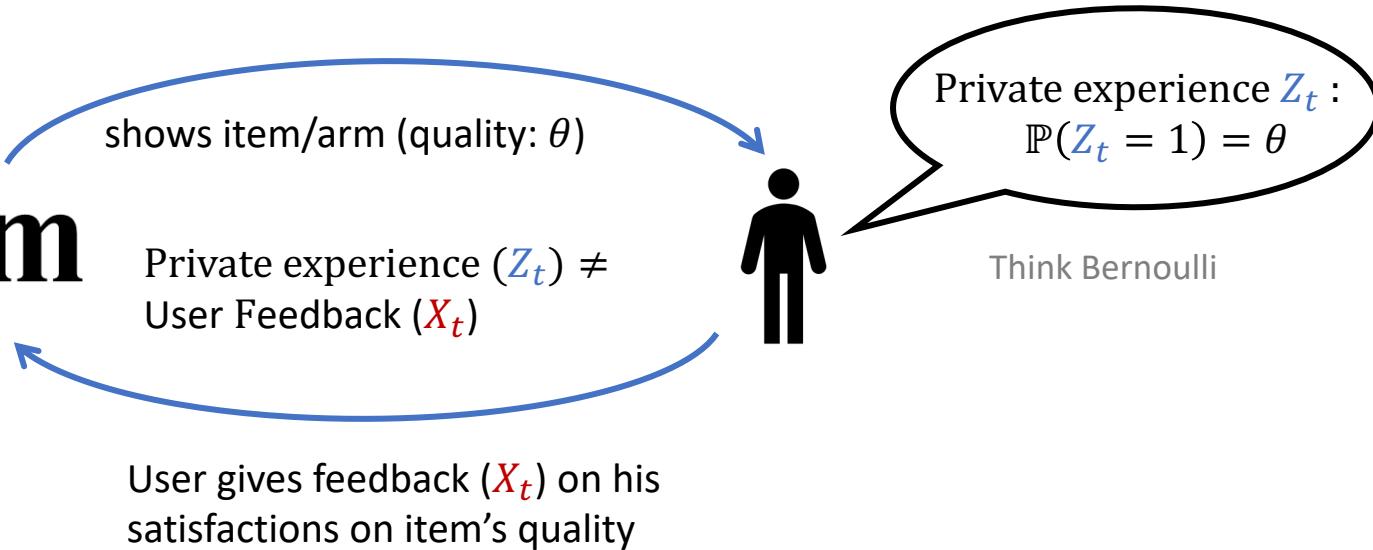


Social Influence Bias: A Randomized Experiment. Muchnik et al. Science 2013.

Can we still be able to design no-regret learning algorithms  
when true reward is not observable, while only biased  
feedback is available?

# Feedback model

# System



The probability for user to provide positive feedback:

$$\mathbb{P}(X_t = 1) = \text{Feedback}(\theta, \rho, n)$$

$\rho$ : positive feedback ratio

$n$ : total feedback received

# Summary of our results

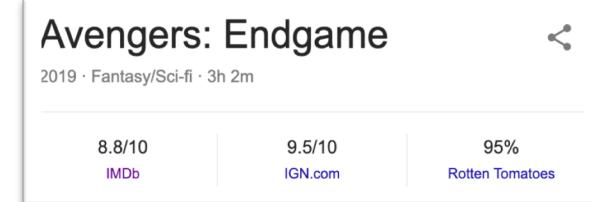
- Biased by the empirical average (Avg-Herding model):
  - User' feedback are biased by the average feedback ( $\rho$ ).
  - **Positive results:** Achieve no-regret learning.
- Biased by the whole history (Beta-Herding model):
  - User's feedback are biased by average feedback ( $\rho$ ) and total # of feedback ( $n$ ).
  - Consider a stylized model that users are performing Bayesian updating.
  - **Negative results:** no bandit algorithm could achieve no-regret learning.

# Biased by the empirical average (Avg-Herding model)

- Feedback function  $\mathbb{P}(X_t = 1 | \rho_t) = F(\theta, \rho_t)$

$\theta$ : item quality (ratio of users liking the movie)

$\rho_t$ : empirical feedback so far



- How does average feedback change over time for a single arm?

$$\begin{aligned}\rho_{t+1} &= \frac{t\rho_t + X_t}{t+1} \\ &= \rho_t - \frac{1}{t+1} (\rho_t - F(\theta, \rho_t) + F(\theta, \rho_t) - X_t)\end{aligned}$$



Re-naming the variables,  $\frac{\partial G}{\partial \rho} = \rho - F(\theta, \rho)$

$$\rho_{t+1} = \rho_t - \eta_{t+1} (\nabla_\rho G(\theta, \rho_t) + \xi_{t+1})$$

Users are collectively performing online gradient descent.

# Biased by the empirical average (Avg-Herding model)

- Utilize the connection to online gradient descent
  - The average feedback **asymptotically converges to some value**

LEMMA 4.2. Let  $\mathcal{S}_\theta := \{\rho : \rho - F(\theta, \rho) = 0\}$ . We have  $\mathbb{P}(\lim_{t \rightarrow \infty} \rho_t \in \mathcal{S}_\theta) = 1$ .

- Derive the **convergence rate**

THEOREM 4.4. Given  $L_F^\rho < 1$ , i.e.,  $G$  is strongly convex.  $\forall \epsilon > 0$ , we have,

$$\mathbb{P}(|\rho_t - \rho^*| \geq \epsilon) \leq \exp\left(-\frac{(\epsilon - \epsilon_t)^2}{2 \sum_{i=1}^t L_i}\right),$$

- Mapping from the converged feedback to the quality is **unique**
- Key interpretations:
  - The average feedback might not be accurate in representing item's quality
  - We can infer true item quality from average ratings (when # feedback is large)
  - Designing bandit algorithms with no-regret learning is possible

# Biased by the empirical average (Avg-Herding model)

- Algorithm:
  - Maintain a quality estimator for each arm (unique mapping)
  - Compute the confidence interval of each arm (convergence rate)
  - Select the arm with highest upper confidence
    - Apply UCB

[Regret Bound]: The expected regret is bounded by:

$$\mathbb{E}[R(T)] = \mathcal{O}\left(\frac{(\ln T)^{\bar{\lambda}'}}{\Delta_{\min}^{2\bar{\lambda}'-1}}\right)$$

$\bar{\lambda}'$  smaller,  
more biased,  
more regret

where  $\bar{\lambda}'$  : hardness of the problem;  $\Delta_{\min} = \min \Delta_k$ .

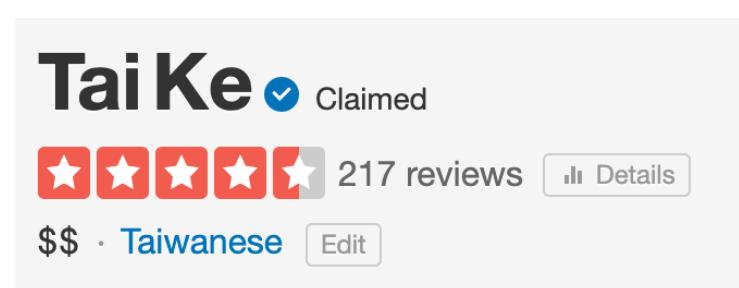
# Biased by the whole history (Beta-Herding model)

- Given history information  $(n, \rho)$ , users update their beliefs about the arm quality in a **Bayesian** manner:
  - $m \geq 0$ : the weight that users put on private experience.

$$\mathbb{P}(X_t = 1 | \rho_t) = \text{Feedback}(\theta, \rho_t, n_t) = \frac{m\theta + n\rho}{m + n}$$

when  $m = 0$ ,  $F(\theta, \rho, n) = \rho$ : totally biased;

when  $m \rightarrow \infty$ ,  $F(\theta, \rho, n) = \theta$ : unbiased



# Biased by the whole history (Beta-Herding model)

- How does average feedback change over time for a single arm?
  - $\lim_{t \rightarrow \infty} \rho_t$  converges to a random variable with **non-zero variance**.
$$\lim_{t \rightarrow \infty} \rho_t \sim \text{Beta}(m\theta, m(1 - \theta))$$
when  $m \rightarrow \infty$ , the Beta distribution will shrink to a Dirac delta function with the point mass in  $\theta$ .
  - Implication: impossible to infer true item quality from the average feedback
- Impossibility result
  - Using information theoretic arguments, there **exists no bandit algorithms** that achieve sublinear regrets in this setting.

Proof Sketch: Step 1. No single feedback path allows to learn  $\theta$ .

**Cumulative Fisher information on  $\theta$  given infinite feedback is bounded.**

Step 2. Any unbiased estimator has non-zero variance.

Step 3. Impossibility to infer arm's true quality. → Linear regret

# Biased by the whole history (Beta-Herding model)

- A natural approach to get over this impossibility results is to break the assumption by taking **interventions**:
  - designs the information structure to induce certain types of “feedback”.
- A toy example: consider binary choice in information design
  - either **showing no history information** (users provide unbiased feedback)
  - or **showing all history information** to users (users’ feedback follow beta-herding feedback model)
- Future work: learn to design information structure to **nudge** human decisions.

# Conclusions and Future work

- We consider bandit learning with different natural user biased behavior which lead to different learning results.
- Future work
  - User behavior: social learning or other behavior models
  - Information structure design

Questions?