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 $R(T) = o(T)$

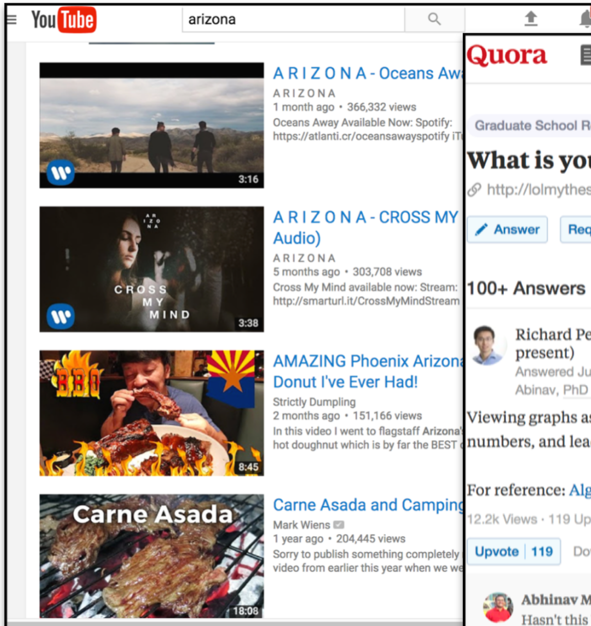
Exploration vs **Exploitation**

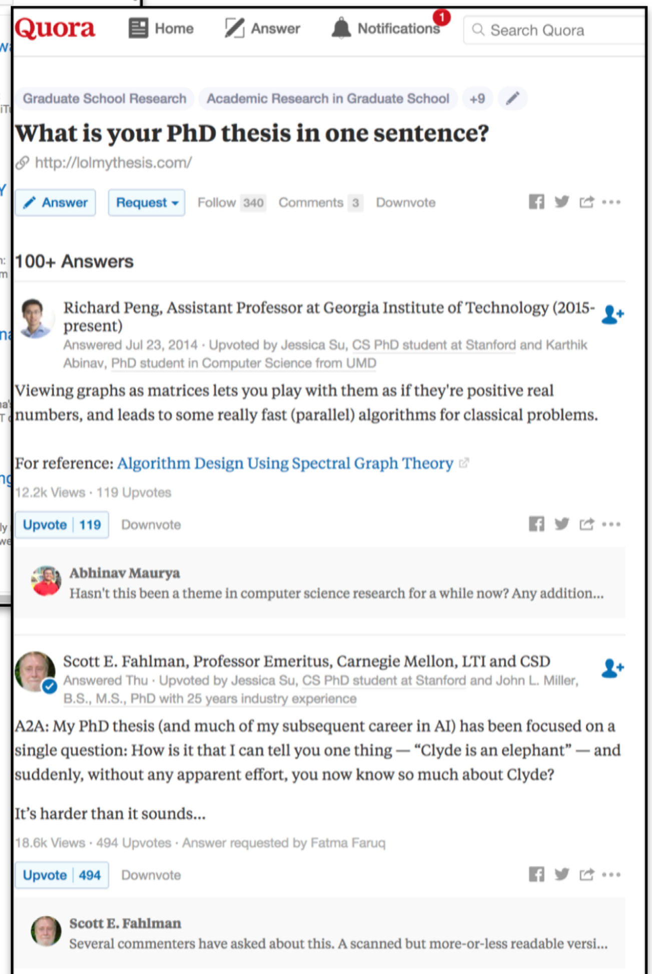
Bandit learning with humans in the loop

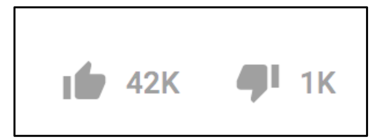
- In the literature
  - Arms can be strategically selected by the myopia users
    - "External" incentives: *monetary payments.* FKKK EC'14,
    - "Intrinsic" incentives: *information asymmetry.* YAVW EC'15 EC'16 , KG Econometrica'11, KG AER'14
  - Arms can strategically reporting their rewards
    - Treat each arm as a strategic agent. BMS COLT'19

- In this work, we consider biased signal of unobservable reward

User-generated content System







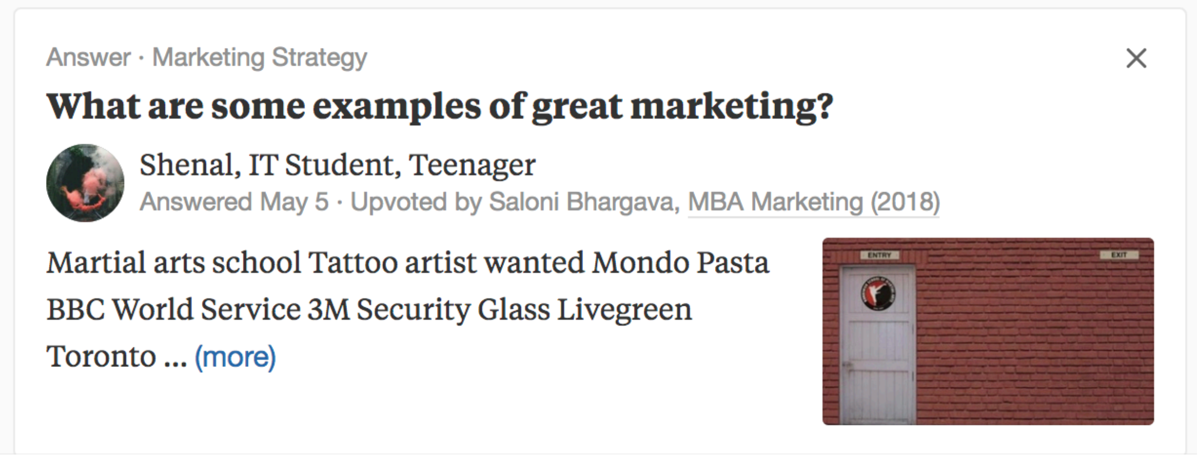
User-generated content System

- 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

Feedback = happiness?

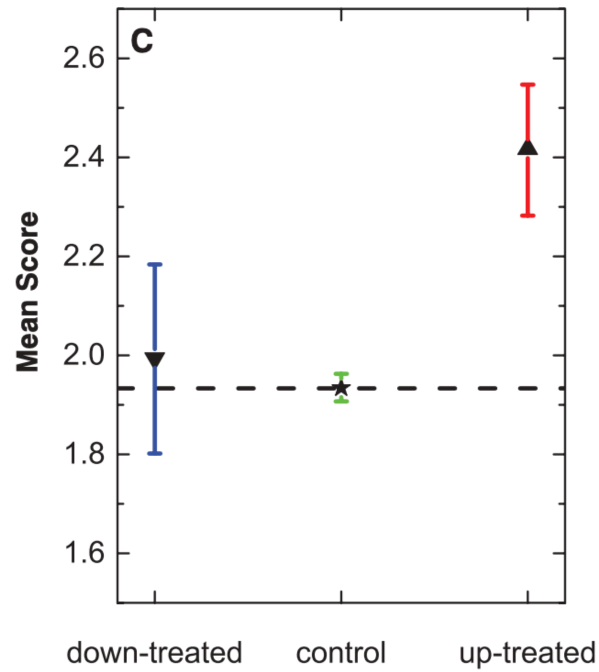
- A standard bandit learning problem
  - Arm: the content chosen to show to users



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.

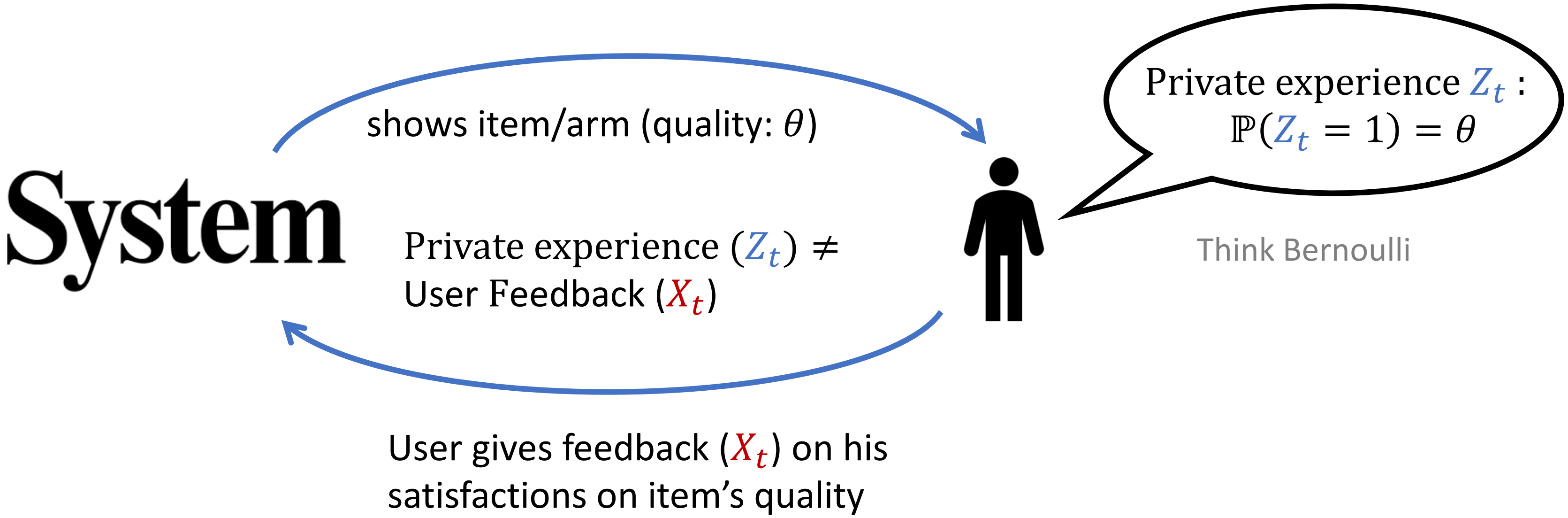
Herding effect



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



The probability for user to provide positive feedback:

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

ρ: 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 (ρ).
  - Positive results: Achieve no-regret learning.

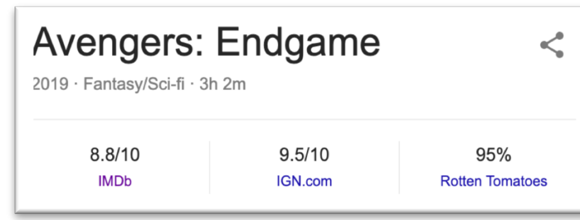
- Biased by the whole history (Beta-Herding model):
  - User's feedback are biased by average feedback (ρ) and total # of feedback (n).
  - Consider a stylized model that users are performing Bayesian updating.
  - Negative results: no bandit algorithm could achieve no-regret learning.

Biased by the empirical average (Avg-Herding model)

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

  θ: item quality (ratio of users liking the movie)

  ρ_t: empirical feedback so far



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

$$\rho_{t+1} = \frac{t\rho_t + X_t}{t+1}$$

$$= \rho_t - \frac{1}{t+1}\left(\rho_t - F(\theta, \rho_t) + F(\theta, \rho_t) - X_t\right)$$

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

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

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* $\mathbb{P}(\lim_{t \to \infty} \rho_t \in \mathcal{S}_\theta) = 1$.

  - 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;* $\Delta_{min} = \min \Delta_k$.

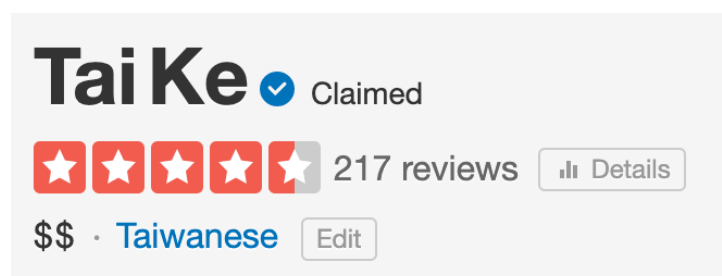
Biased by the whole history (Beta-Herding model)

- Given history information (n, ρ), 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 $m \to \infty, F(\theta, \rho, n) = \theta$: unbiased



Biased by the whole history (Beta-Herding model)

- How does average feedback change over time for a single arm?
  - $\lim_{t\to\infty} \rho_t$ converges to a random variable with non-zero variance.

$$\lim_{t\to\infty} \rho_t \sim \mathrm{Beta}\big(m\theta, m(1-\theta)\big)$$

when $m \to \infty$, the Beta distribution will shrink to a Dirac delta function with the point mass in θ.

  - 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 θ.

Cumulative Fisher information on θ given infinite feedback is bounded.

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

Step 3. Impossibility to infer arm's true quality. $\longrightarrow$ 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?