Bandit Learning with Human Biased Feedback

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

Multi-armed Bandit learning

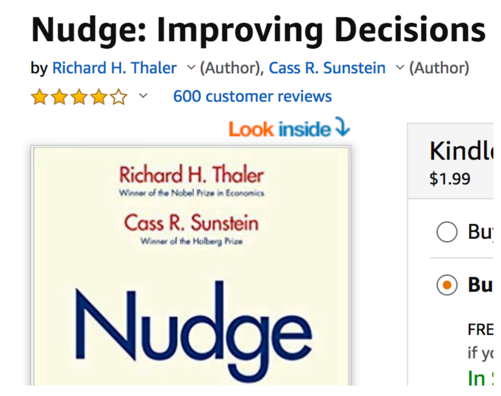
Arm 1	Arm 2	Arm 3	Arm 4	
50%	70%	35%	45%	Reward distributions are unknown

Which arm to pick next

- Sequential game, T rounds, K arms, binary reward;
- At time t, select arm I_t, observe reward $Z_t \in \{0,1\}$
- Minimize the cumulative regret:

$$\mathbb{E}[R(T)] = T\theta^* - \mathbb{E}\left[\sum_{t=1}^{T} Z_t\right]$$

Biased Human Feedback



Recommend item/arm (quality: θ)

amazon

Private experience (Z_t) $\neq$ Feedback (X_t)

$\mathrm{Prob}(Z_t = 1) = \theta$

User gives feedback (X_t) on his satisfactions on item's quality

Can Amazon learn item's quality while only having access to the biased feedback $X_1, \dots, X_t$?

Feedback function: $\mathbb{P}(X_t = 1 | \rho_t) = \mathrm{Feedback}(\theta, \rho, n)$

- ρ: positive votes ratio
- n: total votes received

Avg-Herding Model

- User feedback is biased by the average feedback of the arm. Particularly, the feedback function has the form:

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

- Given current history information of item (n_t, ρ_t), the update rule of ρ_{t+1} is given as follows:

$$\rho_{t+1} = \frac{t\rho_t + X_t}{t+1} = \rho_t - \underbrace{\frac{1}{t+1}}_{\text{Learning rate}}\left(\rho_t - F(\theta,\rho_t) + \underbrace{F(\theta,\rho_t) - X_t}_{\text{Martingale noise}}\right)$$

Key Observation: $\rho_{t+1} = \rho_t - \eta_{t+1}(\rho_t - F(\theta, \rho_t) + \xi_{t+1})$

$\nabla_\rho G(\theta, \rho_t) = \rho - F(\theta, \rho)$

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

Theoretical Result

- **Result 1** ρ_t almost surely converges to a deterministic value in the set of $\mathcal{S}_\theta = \{\rho : \rho - F(\theta, \rho) = 0\}$, $\mathbb{P}\left(\lim_{t\to\infty} \rho_t \in \mathcal{S}_\theta\right) = 1$

Below focus on the case when *G is strongly convex*

- **Result 2** [Convergence rate]: In the order of $\mathcal{O}(1/t^{\bar{\lambda}'})$
  $\mathbb{P}(|\rho_t - \rho^*| \geq \delta) \leq \exp\left(\frac{(\delta - \delta_t)}{\mathcal{O}(t^{\bar{\lambda}'})}\right)$
- **Result 3** [Smoothness of F]: *unique mapping between item quality and converged* ρ_t: $F(\hat{\theta}_t, \rho_t) = \rho_t$ *unique solution of* $\hat{\theta}_t$

Algorithm Avg-UCB:

- Maintain a quality estimator for each arm **[Result 2]**
- Compute the confidence interval of each arm **[Result 3]**
- Select the arm with highest upper confidence
  - Apply UCB

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

more biased, $\bar{\lambda}'$ increasing, more regret.

where $\Delta_{min} = \min \Delta_k$, $\bar{\lambda}' = \max\{1, 1/(2\bar{\lambda})\}$, $\bar{\lambda} = \inf \nabla_\rho^2 G = \nabla_\rho(\rho - F(\theta, \rho))$

Beta-Herding Model

- Given history information (n, ρ), users update their beliefs about the arm quality in a Bayesian manner:

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

$m \geq 0$: the weight that users put on private experience.

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

Theoretical Result

- **Result 1** $\lim_{t\to\infty} \rho_t$ converges almost surely to a random variable which has non-zero variance: $\lim_{t\to\infty} \rho_t \sim \mathrm{Beta}(m\theta, m(1-\theta))$

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

[Impossibility Result]: There exists no bandit algorithm that can achieve sublinear regrets!

- Taking **interventions to re-design the information structure.**
- What's the minimal intervention we can do to get over this impossibility result?
- Two-level policy: consider binary choice in information design
  - either showing no history information [in First T^α, Apply UCB]
  - or showing all history information to users [Present best arm in next $T - T^\alpha$ rounds.]

As longs as $\alpha = \Omega(1/\ln(T))$: $\mathbb{E}[R(T)] = \mathcal{O}\left(\sqrt{\alpha T^{1-\alpha}\ln(T)}\right)$

Conclusions and Future Work

Investigate two natural class of models:

- Avg-Herding model: Positive results
- Beta-Herding model: Negative results

A small change on information structure leads to dramatical difference in learnability.