Multi-Armed Bandit Algorithms for Strategic Agents

Touqir Sajed

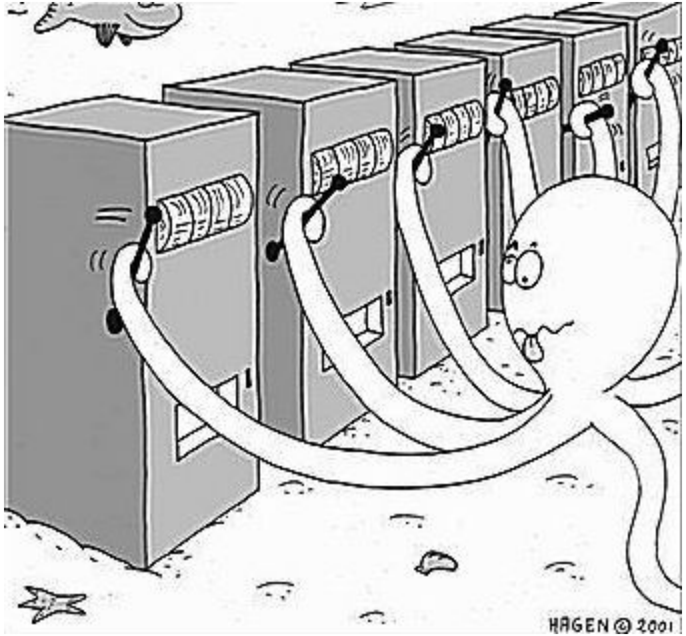
N-Armed Stochastic Bandit Problem

- There are N arms
- The learner pulls an arm at rounds : 1, ... ,**T**
- Pulling an arm $\mathbf{i_t}$ at round **t** generates a reward:

$$r_t \sim \mathcal{D}_{i_t}(\mu_{i_t}) | r_t \in [0, 1]$$

- Goal of the learner: Maximize $\sum_{t=1}^{T} r_t$
- Bound Pseudo Regret :

$$\sum_{t=1}^{T} \max_{i \in \{1,\ldots,K\}} \mu_i - \sum_{t=1}^{T} \mu_{i_t}$$



Incentivizing exploration in the presence of strategic agents

- At each round a **new** agent comes
- Agents are selfish i.e maximize own utility
- Principle recommends arms/items.
- Principle needs information about arms.
- Let the agent explore arms by providing incentives.

BIC Bandit Exploration (Mansour et al 2015)

- Bayesian Incentive Compatible Bandit Exploration.
- The reward means are sampled from known prior distribution.
- Principle sends a recommendation σ_t
- The agent maximizes $\mathbb{E}[\mu_i|\sigma_t]$

BIC Bandit Exploration (Mansour et al 2015)

Definition 2.1. Let $\mathcal{E}_{t-1}$ be be the event that the agents have followed the algorithms recommendations up to (and not including) round t. Then, a recommendation algorithm is Bayesian incentive compatible (BIC) if

$$\mathbb{E}[\mu_i|\sigma_t, I_t = i, \mathcal{E}_{t-1}] \geq \max_{j \in \{1,\ldots,N\}} \mathbb{E}[\mu_j|\sigma_t, I_t = i, \mathcal{E}_{t-1}] \qquad \forall t \in \{1,\ldots,T\},\ \forall i \in \{1,\ldots,N\}$$

- Ex-post regret:

$$R_\mu(T) = T(\max_i \mu_i) - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t} \mid \mu\right]$$

BIC Bandit Exploration (Mansour et al 2015)

- Proposed algorithms that are “sort of optimal”
- Ex-post regret of $\min(f(N), O(\sqrt{t \log(CT)}))$
- The algorithms are BIC
- Caveat : Needs information about priors

BIC Bandit Exploration (Mansour et al 2015)

- Lots of future directions possible!
- Regret bound holds for constant N.
- No problem specific lower bound
- Constrain the amount of information in σ_t

BIC Bandit Exploration (Mansour et al 2015)

- How useful is the setting in practice?
- Priors are usually not known in real world applications
- How about estimating them?
    - Only possible if the algorithm is run multiple times on the same arms
    - Usually, the algorithm runs once on the same arms.

Incentivizing Exploration (Frazier et al 2014)

- At round t, each arm i has state $S_{i,t}$
- Each arm has a markov chain from where the next state is sampled
- The reward sequence is a martingale :

$$\mathbb{E}[\ \mathbb{E}[r_{i,t+1}|S_{i,t+1}]\ |\ S_{i,t}\ ] = \mathbb{E}[r_{i,t}|S_{i,t}]$$

- Define the set of states of all arms as : $\mathbf{S}_t = \bigcup_{i\in\{1,\dots,N\}}\{S_{i,t}\}$

Incentivizing Exploration (Frazier et al 2014)

- At round t, a new agent comes
- Agent selects arm i^* myopically: $i^* = \arg\max_{i \in \{1,\ldots,N\}} \mathbb{E}[r_{i,t}|\mathbf{S}_t]$
- If incentivized, agent selects i^* : $i^* = \arg\max_{i \in \{1,\ldots,N\}} \left(\mathbb{E}[r_{i,t}|\mathbf{S}_t] + c_{i,t}\right)$
- Principle decides to recommend arm j.
- Principle sets incentive c_t : $c_t := c_{j,t} = \left(\max_{i \in \{1,\ldots,N\}} \mathbb{E}[r_{i,t}|\mathbf{S}_t]\right) - \mathbb{E}[r_{j,t}|\mathbf{S}_t]$

Incentivizing Exploration (Frazier et al 2014)

- Given a discount factor gamma :

$$R^{(\gamma)} = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t\right]$$

$$C^{(\gamma)} = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} c_t\right]$$

- They considered maximizing $R^{(\gamma)}$ with constraint $C^{(\gamma)}$ <= b
- Maximize a relaxed lagrangian:

$$R_\lambda^{(\gamma)} = R^{(\gamma)} - \lambda C^{(\gamma)}$$

Incentivizing Exploration (Frazier et al 2014)

- A randomized strategy ***TE***.
- With probability p , let the agent behave myopically
- With probability 1-p, incentivize agent based on algorithm A.

Given a parameter λ*, define* $p = \frac{\lambda}{\lambda+1}$*, and* $\eta = \frac{(1-p)\gamma}{1-p\gamma}$*. Then,*

$$R_{\lambda}^{(\gamma)}(\boldsymbol{TE}_{p,\mathcal{A}}) = \frac{1-\eta}{1-\gamma} \cdot R^{(\eta)}(\mathcal{A}).$$

Incentivizing Exploration by Heterogeneous Users (Chen et al 2018)

- At round t, an agent with type $\boldsymbol{\theta}_t$ comes.
- $\boldsymbol{\theta}_t$ is sampled iid from a known distribution.
- Agent pulls an arm i_t and observes a vector y_t.
- Vector y_t : $\mathbf{y}_t = \mu_{i_t} + \zeta_t$ s.t $\zeta_t \sim \mathbf{subG}(\sigma \cdot I_d)$
- Define for arm i with $\hat{\mu}_{t,i}$ the empirical mean over all past y_t s.t i_t=i
- Principal provides incentive $c_{t,i}$
- Agent selects arm i that maximizes : $(c_{t,i} + \boldsymbol{\theta}_t \cdot \hat{\boldsymbol{\mu}}_{t,i})$

Incentivizing Exploration by Heterogeneous Users (Chen et al 2018)

- Goal is to minimize expected regret $E[R_T]$ and expected payments $E[C_T]$:

$$\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^{T} \max_i(\boldsymbol{\mu}_i \cdot \boldsymbol{\theta}_t) - \boldsymbol{\theta}_t \cdot \boldsymbol{\mu}_{i_t}\right]$$
$$\mathbb{E}[C_T] = \mathbb{E}\left[\sum_{t=1}^{T} c_{t,i|I_t=i}\right]$$

- Their algorithm incurs $E[R_T]$ at most : $O\left(N \cdot \mathrm{e}^{2/p} + LN \log^3(T)\right)$
- And $E[C_T]$ at most : $O\left(N^2 \cdot \mathrm{e}^{2/p}\right)$
- Suboptimal Bounds.

Incentivizing Exploration by Heterogeneous Users (Chen et al 2018)

- Let $m_{t,i}$ be the number of times arm i has been pulled till round t.
- An arm i is "eligible" at phase s if:
  - If it has been pulled at most s times uptil round t. **AND**
  - If : $\mathbb{P}_\theta[\boldsymbol{\theta} \cdot \hat{\boldsymbol{\mu}}_{t,i} > \boldsymbol{\theta} \cdot \hat{\boldsymbol{\mu}}_{t,i'} \ \forall\ i' \neq i] < 1/\log(s)$

Incentivizing Exploration by Heterogeneous Users (Chen et al 2018)

Algorithm 1 Algorithm: Incentivizing Exploration

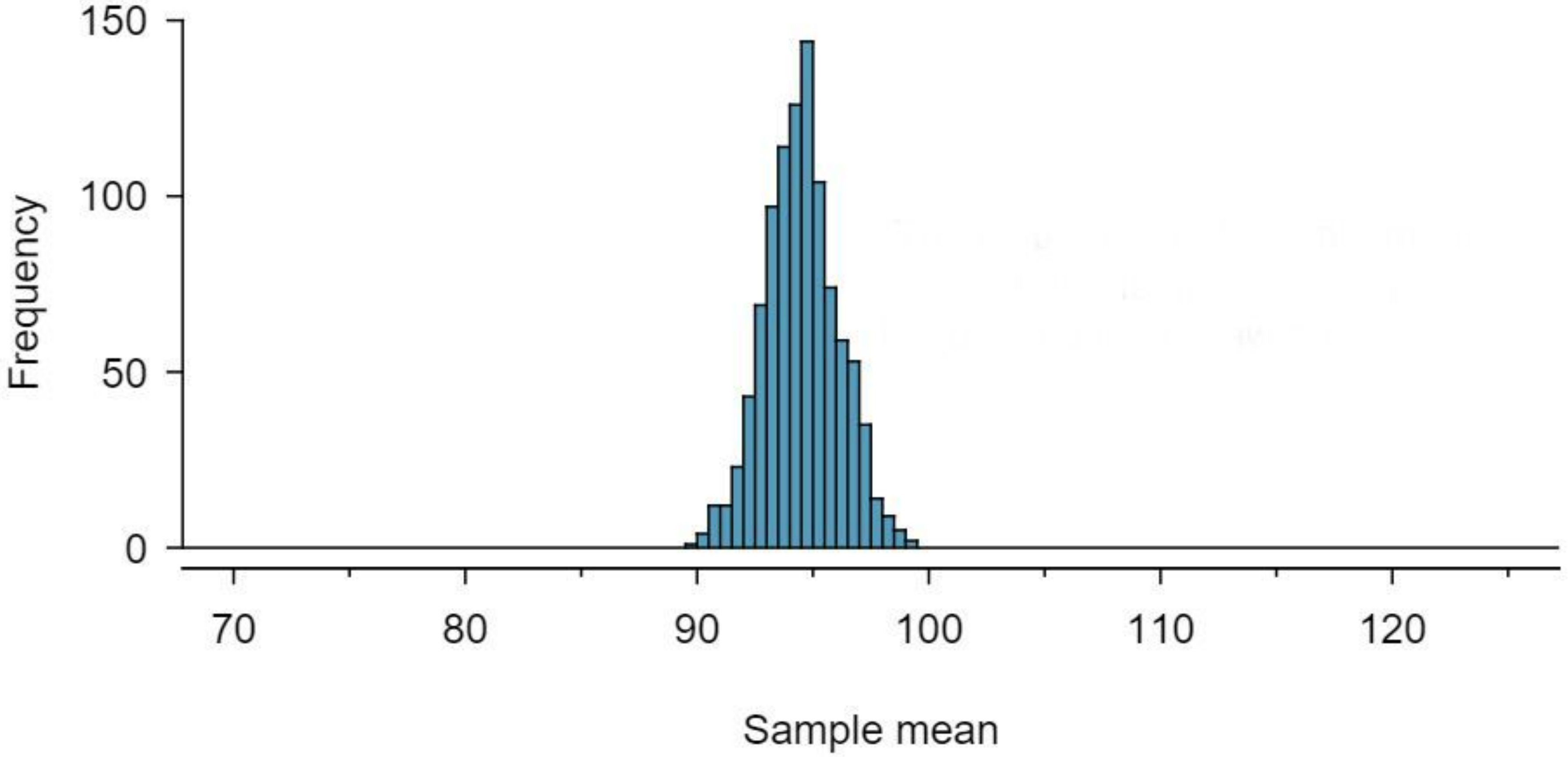
Set the current phase number $s = 1$. {Each arm is pulled once initially "for free."}
for time steps $t = 1, 2, 3, \ldots$ **do**
  if $m_{t,i} \geq s + 1$ for all arms i **then**
    Increment the phase $s = s + 1$.
  if there is a payment-eligible arm i **then**
    Let i be an arbitrary payment-eligible arm.
    Offer payment $c_{t,i} = \max_{\boldsymbol{\theta},i'} \boldsymbol{\theta} \cdot (\hat{\boldsymbol{\mu}}_{t,i'} - \hat{\boldsymbol{\mu}}_{t,i})$ for pulling arm i (and payment 0 for all other arms).
  else
    Let agent t play myopically, i.e., offer payments 0 for all arms.

Incentivizing Exploration by Heterogeneous Users (Chen et al 2018)

- Their algorithm is suboptimal.
- Offers payment based on worst case θ_t.
    - Why not make better use of theta's distribution?
- It randomly chooses an eligible arm for recommendation.
    - Rather, why not choose the **most eligible** arm?

Our contributions

- We address the two issues using a thompson sampling like strategy
- The theta distribution may not be given.
- Additionally we assume that with probability **p**:
    - A freeloader agent comes and takes action that only maximizes incentives.



150
100
50
0
Frequency
70
80
90
100
110
120
Sample mean

Thompson Sampler

- Suppose theta distribution is known and multinomial.
- Maintains posterior distributions over the means
- Uses beta-multinomial model
- At each round, it samples the means using thompson sampler
- Samples a theta
- Principle selects the arm that maximizes the dot product with theta and the sampled means for incentivizing
- How to select the incentive?

Thompson Sampler

- How to select the incentive?
- With probability 1-p use c_t
- With probability p, use ≅ 0.

Performance

- How well does it perform in contrast to Chen's algorithm?
- If p = 0, we are in the same setting as Chen's.
- Ran the algorithms on randomly generated data.
- Under p=0, preliminary results show better performance.

When $\boldsymbol{\theta}$ distribution is unknown

- Need to approximate $\boldsymbol{\theta}$ distribution.
- Compute empirical probability mass function (EMF) - point estimate.
- Construct a ball **B** around the point estimate
  - Such that with high probability the true PMF lies within the ball
  - Involves Concentration of measure analysis.
- How to choose a distribution $\mathcal{D}$ from the ball?
- Based on $\arg\max_{\mathcal{D} \in B} \mathbb{E}_{\boldsymbol{\theta}} \left[\max_{i,j} (\bar{\boldsymbol{\mu}}_i \cdot \boldsymbol{\theta} - \bar{\boldsymbol{\mu}}_j \cdot \boldsymbol{\theta})\right]$

Future Work

- Theoretically analyze regret and cumulative payments.
- Carry out empirical experiments on real world data (like Mechanical Turk)
- Is there a better strategy to sample from the ball **B**?

Thank you!
Questions?

References

1. Chen, B., Frazier, P., & Kempe, D. (2018, July). Incentivizing Exploration by Heterogeneous Users. In Conference On Learning Theory (pp. 798-818).
2. Frazier, P., Kempe, D., Kleinberg, J., & Kleinberg, R. (2014, June). Incentivizing exploration. In Proceedings of the fifteenth ACM conference on Economics and computation (pp. 5-22). ACM.
3. Mansour, Y., Slivkins, A., & Syrgkanis, V. (2015, June). Bayesian incentive-compatible bandit exploration. In Proceedings of the Sixteenth ACM Conference on Economics and Computation (pp. 565-582). ACM.