Exploration, Exploitation and Incentives

Yishay Mansour

Outline

- Incentives:
  - Actions are only recommendations
  - Agents decide whether to follow them
  - Need to induce exploration!
- Two deterministic actions
  - Optimal policy
- Two stochastic actions
  - Generic framework

Report Cards

- Report-card systems
  - Health-care, education, …

  Public disclosure of information
    - Patients health, students scores, …
- Pro:
  - Incentives to improve quality
  - Information to users
- Cons:
  - Incentives to “game” the system
    - avoid problematic cases

User Based Recommendations

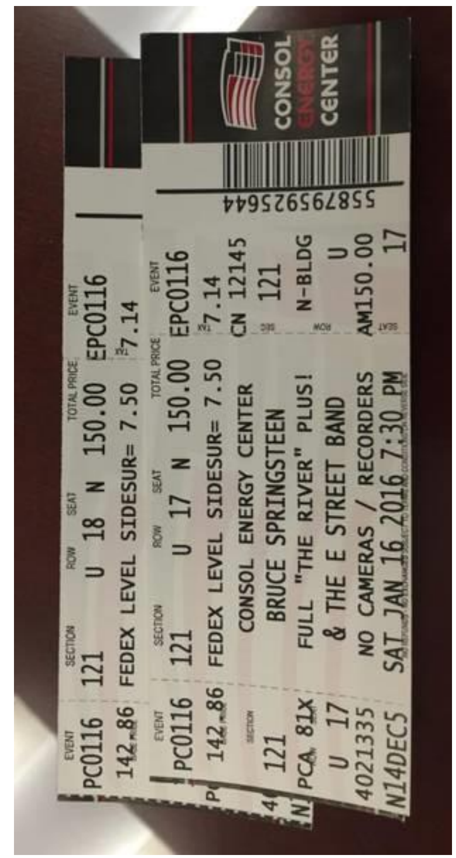
- Recommendation web sites
- Example: TripAdvisor
- User based reviews
- Popularity Index
  - Proprietary algo.
  - Self-reinforcement
- Can be used to induce exploration

Waze: User based navigation

- Real time navigation recommendations
- Based on user inputs
  - Cellular/GPS
- Recommendation dilemma:
  - Need to try alternate routes to estimate time
  - Actually, done in practice

Resell tickets

- Secondary market for show tickets
  - StubHub
- Matches sellers and buyers
- New feature: price recommendation
  - Implicit coordination between sellers



Multi-Arm Bandit

- Simple decision model
- Multiple independent actions
- Uncertainty regarding the rewards
- Repeated interaction
- Tradeoff between exploration and exploitation

MAB

Classical setting

- uncertainty regarding rewards
- action execution:
  - arbitrary

Today setting

- uncertainty regarding rewards
- action execution:
  - control by agents
  - Bayesian Incentive Compatible (BIC)

Our Motivation

- Agents need to select between few alternatives:
  - Hotels, Traffic routes, Doctors, ticket price
  - Known prior on the success
- Multiple agents arriving:
  - Each makes one decision, and get
  - Individual agents are strategic
    - Maximizing their reward
- Planner:
  - Would like to learn and implement the better alternative
    - Government, regulator, society, etc.
    - Maximize user satisfaction

Agents are both producers and consumers

Main Research Question

- Planner policy limitations:
  - No monetary incentives
  - Controlling revelation of information
- Can the planner induce exploration?
  - Guarantee that the best alternative is selected
- What is the expected regret
  - Compared to a non-strategic setting.
  - Bound the cost of exploration

Model

Environment

- K actions: a_1 … a_k
- Prior over μ_i
  - Realized only once, initially
- Given μ_i action i has reward R_i (r.v.) s.t. $E[R_i]=\mu_i$
  - Deterministic/stochastic
  - Range [0,1]
- Notation: $E[\mu_{i-1}] > E[\mu_i]$

Agents

- T agents
  - Arrive sequentially
    - Known arrival order
- Select once a single action
  - Get the reward of the selected action
- Risk neutral
- Agent optimal strategy:
  - Given all the observed information
  - Select the action that maximizes expected payoff

Model

Planner

- Controls the information
- Agents are Incentive Compatible
- No side payments
- **Planner goal:**
  - Social welfare maximization
  - Minimize regret
    - REGRET = T*max μ_i - E[Rew]
  - Arbitrary
    - Max-min, etc.

Planner actions:

- Gives agent $\boldsymbol{t}$ message $\boldsymbol{m_t}$
  - information about past.
  - W.l.o.g. recommendation $\boldsymbol{a_t}$
- Observes the outcome
  - Realization $\boldsymbol{r_{a_t}}$
- Cumulative Reward

$$\text{Rew} = \sum_t r_{a_t}$$

- Agents know Planner policy

Controlling Information

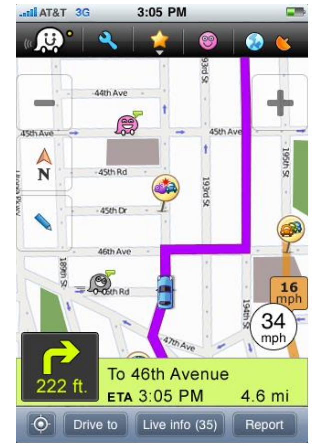
Report Cards

Public Recom.

Waze

Individual

Recom.

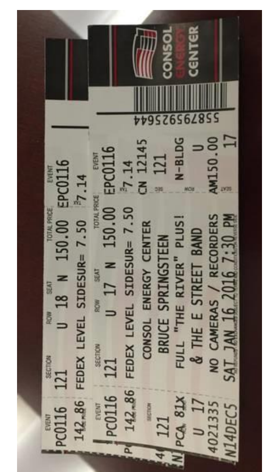


TripAdvisor

Time based

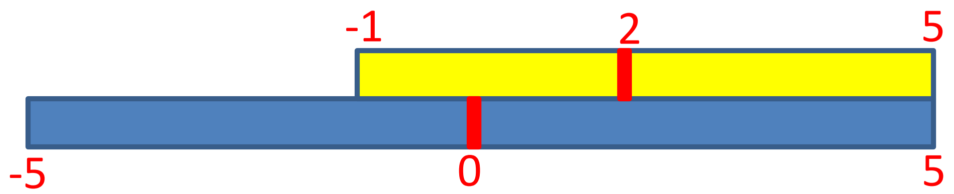
Ticket resell

Group recom.



Simple recommendations: No information

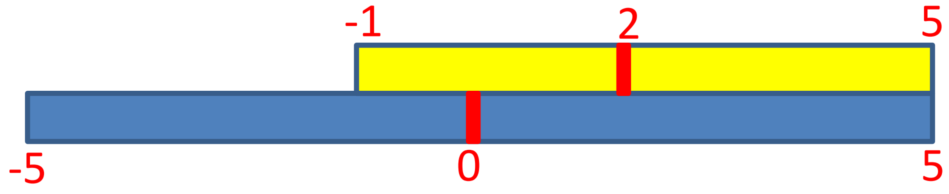
- Example:
  - $R_1 \sim \boldsymbol{U}[-1, 5]$
  - $R_2 \sim \boldsymbol{U}[-5, 5]$
  - T large (optimal to test the both alternatives).



- All agents prefer the better a priori alternative
  - Action 1
- No exploration!
- High regret: 2.6*T-2*T=0.6*T

Simple recommendations: Full Transparency

- Agent 1: chooses the first action.
- Agent 2: Observes r_1
  - If $r_1 > 0$: Selects action 1
    - All following agents select action 1
  - If $r_1 \leq 0$: Selects action 2
    - All following agents select the better action
- outcome is suboptimal for large T:
  - Regret = 2.6*T – 2.252*T =0.348*T



Public Recommendations

- Better than Full information
  - Only recommendations are public
  - In the example: recommend action 2

    If $r_1 < +1$

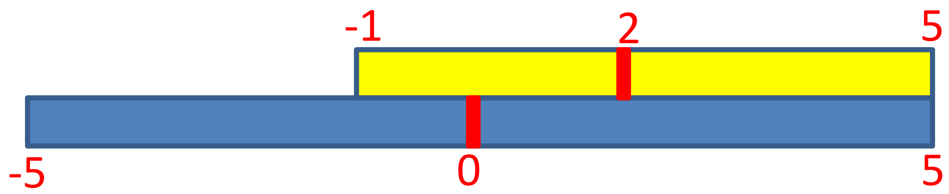
- Main Observation:

all exploration can move to second agent

  - Simple characterization
  - Significant limitation

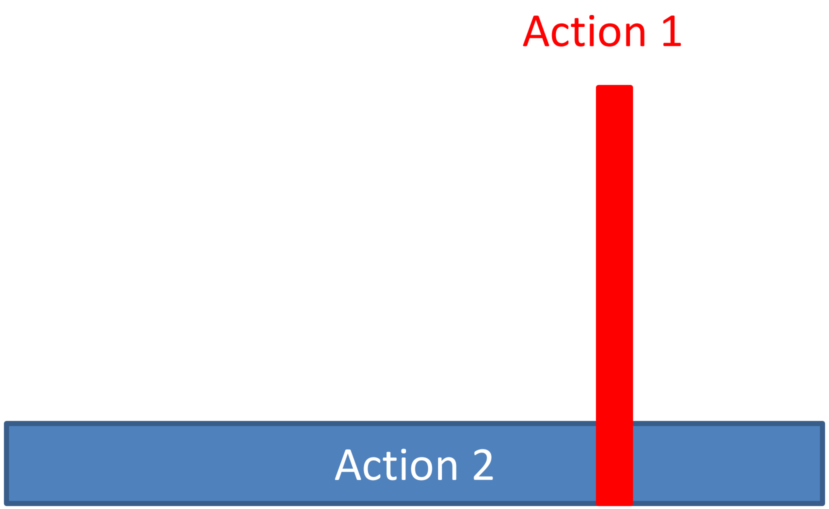
- Linear regret:

2.6*T – 2.42*T =0.18*T



Explorable Actions: Two deterministic actions

- Can we hope to explore any action?!
  - Main limitation is BIC
- Example:
  - Action 1 always payoff 0
  - Action 2 prior Unif[-2,+1]
    - $E[R_2] = -1/2 < 0$
- Agent *t* knows:
  - All prior agents preferred action 1
  - Planner has no info on action 2
  - Hence, will do action 1



Condition

$\Pr[\mu_1 < E[\mu_2]] > 0$

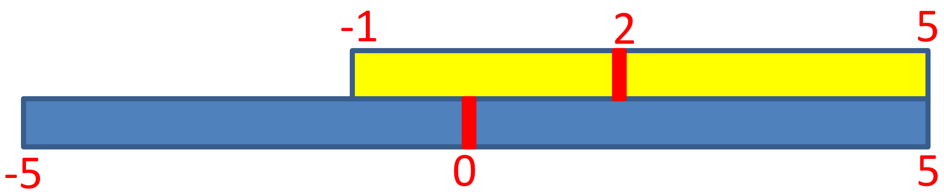
Explorable actions: Two stochastic actions

- Requirement
  - We need “Evidence” that action 2 might be better
    - For this we can use realizations of action 1

- Condition for a distribution P
  - There exists k_P such that there exists
  - $\Pr[\, E[\mu_2] > E[\mu_1 \mid \text{some } k_p \text{ outcomes}] \,] > 0$

Optimal Policy (first agent)

- Example:
  - R_1 ~ $\boldsymbol{U}$ [-1 , 5]
  - R_2 ~ $\boldsymbol{U}$ [-5 , 5]
  - T large (optimal to test the both alternatives).

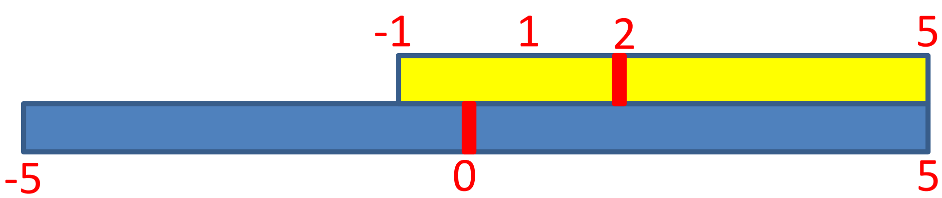


➢ Recommend action 1 to first agent

- The only recommendation agent 1 will follow

Optimal Policy (second agent)

- recommends 2nd alternative to agent two whenever $r_1 \leq 1$.
- This is ***IC*** because
  - $E[R_1 \mid \text{recommend}(2)\,] = 0$

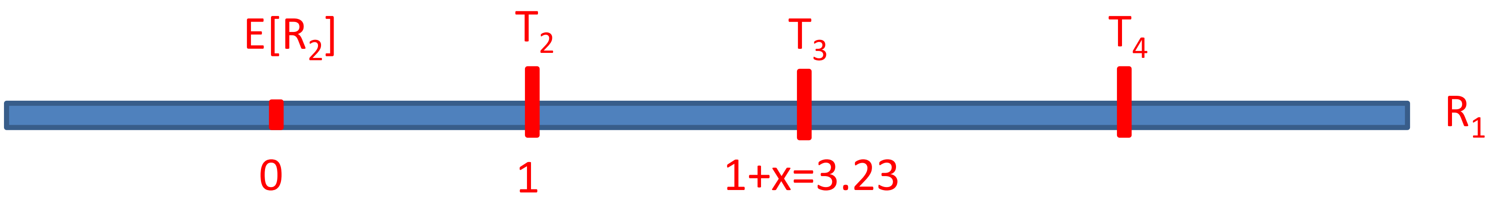


- Better than full transparency
  - more experimentation by the second agent.
    - full transparency is sub-optimal.

- But we can do even better.

Optimal Policy (3rd agent)

- recommends third agent to use 2nd action if one of two cases occurs
  i. Second agent tested 2nd action (***$R_1 \leq 1$***) and the planner learned that ***$R_2 > R_1$***
  ii. ***$1 < R_1 \leq 1+x$*** , so the third agent is the first to test 2nd action
  iii. Gain is constant. Loss due to exploration can be made arbitrarily small. We can always balance them.



Two deterministic actions

Optimal Algorithm

- Agent 1:
  - recommend action 1.
  - Observe reward r_1
- Agent t >1:
  - Both actions sampled: recommend the better action
  - Otherwise: If $r_1 < \theta_t$ then recommend action 2 otherwise action 1

Properties of optimal policy

- Recommendation sufficient
  - revelation principle
- IC constraints tight
- Generally: explore low values before high
  - threshold
- Intuition: tradeoff between potential reasons for being recommended action 2

Recommendation Policy

Recommendation Policy:

- For agent t,
  - Gives recommendation rec_t
- Recommendation is IC
  - $E[R_j - R_i \mid rec_t = a_j] \geq 0$
- Note that it requires IC:
  - Implies: recommend to agent 1 action a_1
- **Claim**: Optimal policy is a Recommendation Policy

Proof (Revelation Principle):

- $M(j,t)$ – set of messages that cause agent t to select action a_j.
- $H(j,t)$ – the corresponding histories
- $E[R_j - R_i | m] \geq 0$ for $m \in M(j,t)$
- Consider the recommendation a_j after $h \in H(j,t)$
- Still IC
- Identical outcomes

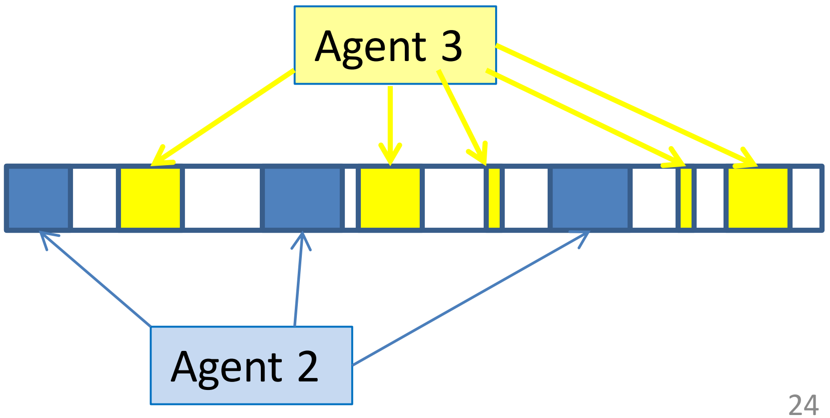
Partition Policy

Partition Policy:

- Recommendation policy
- Agent 1: recommending action a_1 and observing r_1
- Disjoint subsets $I_t, t \geq 2$
- If $r_1 \in I_t$
  - Agent t first to explore a_2
  - Any agent $t' > t$ uses the better of the two actions
    - Payoff $\max\{r_1, r_2\}$
- If $r_1 \in I_{T+1}$ no agent explores a_2

Optimal policy is a partition:

- Recommending the better action
  - when both are known
  - Optimizes sum of payoffs
  - Strengthen the IC



Only worse action is "important"

Lemma:

Any policy that is

IC w.r.t. a_2 is

IC w.r.t. a_1

Proof:

- Let $K_t = \{(R_1, R_2)\}$ set of event that cause $rec_t = a_2$
- If empty then $E[R_1 - R_2] \geq 0$
- Otherwise: $E[R_2 - R_1 | K_t] \geq 0$
  - Since it is an IC policy
- Originally: $E[R_2 - R_1] < 0$
- Therefore

$E[R_2 - R_1 \mid \text{not } K_t] < 0$

Second agent explores low values

- **Claim**: The second agent explores for any value $r_1 < \mu_2$

- Proof:
  - Consider an agent t that explores for $r_1 < \mu_2$
    - Call this set of values B
  - Move the exploration of B to agent 2
  - Agent 2: Improve the IC constraint for a_2
    - By $E_B[\mu_2 - r_1] > 0$
  - Agent t: Improve the IC constraint for a_2
    - When $r_1 \in B$ the payoff is $E_B[\max\{r_2, r_1\}]$

IC constraints

- Basic IC constraint:

$$E[R_2 - R_1 | rec_t = 2] \geq 0$$

- Alternatively,

$$F(M) = E[R_2 - R_1 | M] \Pr[M]$$

$$F(rec_t = a_2) = E[R_2 - R_1 | rec_t = 2] \Pr[rec_t = 2] \geq 0$$

- Recommendation policy:

$$F(r_1 \in \cup_{\tau<t} I_\tau, R_2 > R_1) + F(\{r_1 \in I_t\}) \geq 0$$

IC constraints

- Recommendation policy
  - With sets I_t

  - $\underbrace{F(r_1 \in \cup_{\tau<t} I_\tau \wedge \{R_2 > R_1\})}_{\text{Positive (exploitation)}} + \underbrace{F(\{r_1 \in I_t\})}_{\text{Negative (exploration)}} \geq 0$

Threshold policy

- Partition policy such that $I_t = (i_{t-1}, i_t]$
- $I_2 = (-\infty, i_2)$
- $I_{T+1} = (i_T, \infty)$

Agent 2	Agent 3	Agent 4	Agent 5	No exploration

- **Main Characterization**:

The optimal policy is a threshold policy

Optimal has Tight IC constraints

Lemma:

If agent t+1 explores ($Pr[I_{t+1}]>0$)

Then

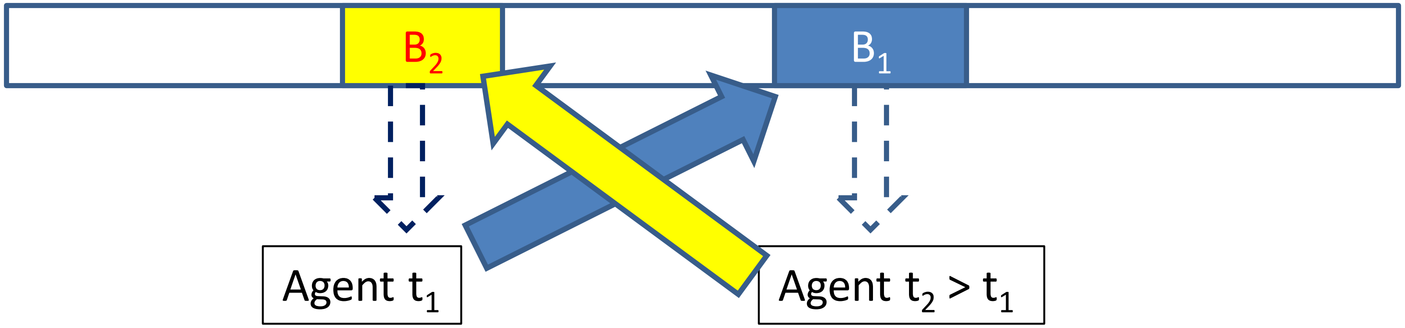
Agent t has a tight IC constraint.

Proof:

- Move exploration from agent t+1 to agent t
- Improves sum of payoffs
  - Replaces r_1+R_2 by $R_2 + \max\{r_1,r_2\}$
- Keeps the IC for agent t (since it was not tight)
- Keeps the IC for agent t+1 (remove exploration)

Threshold policy

➢ What is NOT a threshold policy:

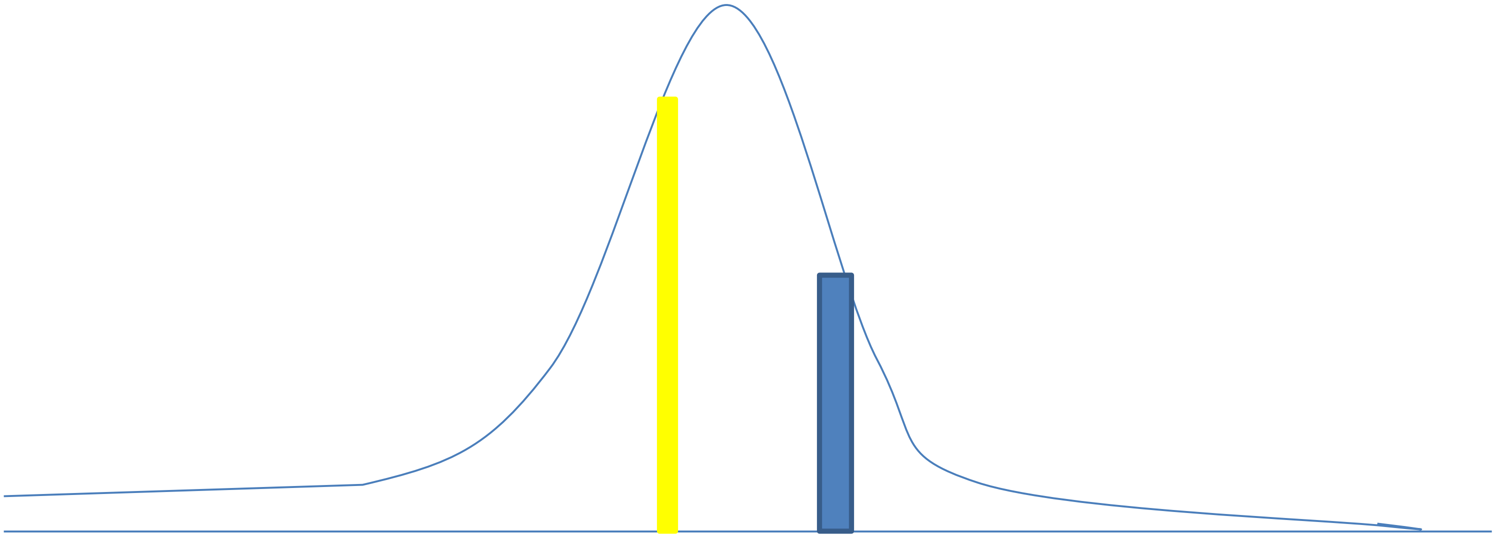


➢ Proper Swap: $F(\{r_1 \in B_1\}) = F(\{r_1 \in B_2\})$

$$F[r_1 \in B_*] = E[\mu_2 - R_1 | r_1 \in B_*] \Pr[r_1 \in B_*]$$

Proper Swap Operation

$$F(\{r_1 \in B_1\}) = F(\{r_1 \in B_2\})$$



Since $B_2 < B_1$ it Implies $Pr[B_2] > Pr[B_1]$

Proper Swap – IC Analysis

- Agent t_1 unchanged
  - Added B_2 subtracted B_1
  - Proper swap implies equal effect.
- Agents other than t_1 and t_2
  - Before t_1 and after t_2: unchanged
  - Between t_1 and t_2: increase willingness
    - Gain $(Pr[B_2] - Pr[B_1])\ max\{r_1,r_2\}$

Proper Swap – IC Analysis

- Agent t_2 (assuming real agent, not T+1)

$$F(r_1 \in B_1, R_2 > R_1) + F(\{r_1 \in B_2\})$$

before

$$F(r_1 \in B_2, R_2 > R_1) + F(\{r_1 \in B_1\})$$

after

$$F(r_1 \in B_2, R_2 > R_1) - F(r_1 \in B_1, R_2 > R_1)$$

diff

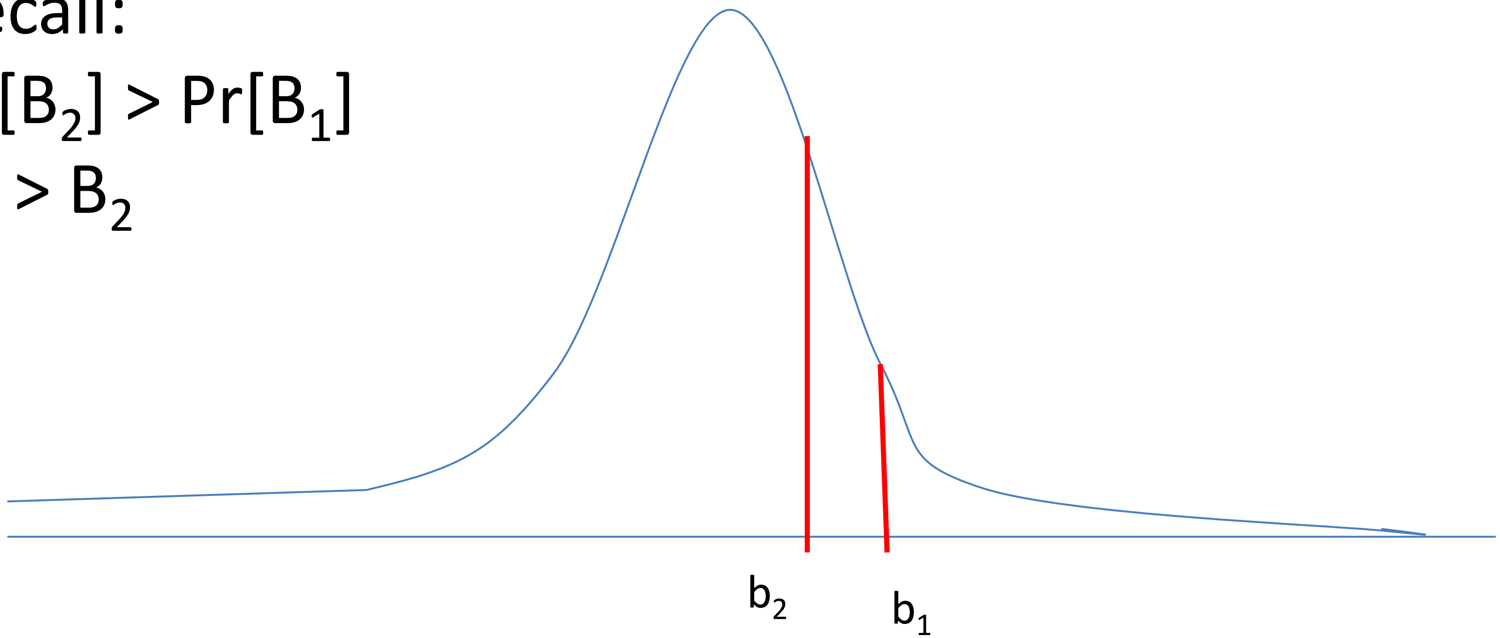
Proper Swap – IC Analysis

$$E(E[R_2 - R_1|R_2 > R_1]|r_1 \in B_2)\,\Pr[r_1 \in B_2]$$
$$> E(E[R_2 - R_1|R_2 > R_1]|r_1 \in B_1)\,\Pr[r_1 \in B_1]$$

Recall:

$\Pr[B_2] > \Pr[B_1]$

$B_1 > B_2$



Proper Swap – Payoff Analysis

- **Before Swap:**

Before	$\mathbf{B_2}$	$\mathbf{B_1}$
t_1	r_1	r_2
t_2	r_2	$\mathrm{Max}\{r_1, r_2\}$

- **After Swap:**

After	$\mathbf{B_2}$	$\mathbf{B_1}$
t_1	r_2	r_1
t_2	$\mathrm{Max}\{r_1, r_2\}$	r_2

$$\mathrm{GAIN} = (\Pr[B_2] - \Pr[B_1])\,(\mathrm{Max}\{r_1, r_2\} - r_1) > 0$$

Optimal Policy

- Threshold policy
- Define thresholds with infinite num. agents:
  - $\Theta_{t,\infty}$
- Compute for each t:
  - $(T-t)E[\max\{R_2-\theta_t, 0\}] = \theta_t - \mu_2$
- Let τ be the minimal index that
  - $\Theta_{t,\infty} > \theta_t$
- Threshold:
  - $\Theta_{t,T} = \min\{\Theta_{t,\infty}, \theta_t\}$

How good is optimal?!

- The loss due to IC
  - Constant (independent of T)
- Bounding the number of exploring agents:
  - $\frac{\mu_1 - \mu_2}{\alpha}$
  - $\alpha = F(\{R_1 < R_2\} \wedge \{R_1 < \mu_2\})$
  - $\alpha = E[R_2 - R_1 | R_1 < R_2, R_1 < \mu_2] \Pr[R_1 < R_2, R_1 < \mu_2]$

Two stochastic actions

- Need to sample multiple times
- How do we incentivize exploration?
- Simple scheme:
  - Same algorithm as deterministic
  - Each step extended to $1/\epsilon^2$ recommendations
- Performance
  - Maintain the BIC
  - High regret: $T^{\frac{2}{3}}$

Basic Technique: Hidden exploration

- Embed exploration in a lot of exploitation
- Exploitation
  - $a^*(h) = \arg\max E[\mu_a|h]$
- Exploration:
  - $a^0(h)$
  - Arbitrary function
- Recommendation:
  - rec

<u>Hidden exploration:</u>

- Input: prior P, history h, parameter $\epsilon > 0$,
- With probability ϵ:
  - $rec \leftarrow a^0(h)$ explore
- Else
  - $rec \leftarrow a^*(h)$ exploit

Hidden Exploration: BIC

- BIC property:

For any actions $a \neq a'$:

$$\Pr[rec = a] > 0 \Rightarrow E[\mu_a - \mu_{a'} | rec = a] \geq 0$$

- Posterior Gap: $G = E[\mu_2 - \mu_1 | h]$

- Lemma: For $\epsilon \leq \frac{1}{3} E[G \cdot I\{G > 0\}]$

algorithm HiddenExploration is BIC

Hidden Exploration: BIC

- Recall:
  - If ALG is BIC for $rec = a_2$ it is also for $rec = a_1$
- <u>Proof of the lemma:</u>
- $M_2 = \{rec = a_2\}, M_{explore}, M_{exploit}$
- $\Pr[M_2] > 0$
  - Otherwise trivial
- $F(M) = E[G|M]\Pr[M]$
- Need to show: $F(M_2) \geq 0$
  - $F(M_2) = F\left(M_{explore} \wedge M_2\right) + F(M_{exploit} \wedge M_2)$

- $F\left(M_{exploit} \wedge M_2\right) = E[G|G > 0]\Pr[G > 0](1-\epsilon)$
  $= (1-\epsilon)\, F(\{G > 0\})$

- $F\left(M_{explore} \wedge\ M_2\right) \geq F\left(M_{explore} \wedge\ M_2 \wedge G < 0\right)$
  $\geq F\left(M_{explore} \wedge\ G < 0\right)$
  $= E[G|G < 0]\Pr[G < 0]\,\epsilon$
  $= \epsilon\, F(\{G < 0\})$

- $F(M_2) \geq (1-\epsilon)\, F(\{G > 0\}) + \epsilon\, F(\{G < 0\})$

- $F(\{G > 0\}) + F(\{G < 0\}) = E[\mu_2 - \mu_1]$
- Sufficient:
  $F(M_2) \geq \epsilon\, E[\mu_2 - \mu_1] + (1 - 2\epsilon)F(\{G > 0\}) \geq 0$

- Holds for:

$$\epsilon \leq \frac{F(\{G>0\})}{2F(\{G>0\})+E[\mu_1-\mu_2]}$$

$$\epsilon \leq \frac{1}{3}F(\{G > 0\}) \leq \frac{F(\{G>0\})}{2F(\{G>0\})+E[\mu_1-\mu_2]}$$

Last inequality follows from simple algebra and because the rewards are in [0,1]

Two stochastic actions – black box

- Black-box reduction
- Goal: "compile" an arbitrary algorithm ALG
  - Arbitrary goal
- Input:

  Arbitrary algorithm ALG
  - Selects an action
  - Observes reward

- Method:
  - Run it using HiddenExploration
- Corollary:
  - BIC
  - vanishing regret

Repeated Hidden Exploration

- Parameters:
  - P, $\epsilon > 0, N_0$
- For $t \in [1, N_0]$
  - $a_t = 1$
- For $t > N_0$:
  - With prob ϵ:

    $a_t \leftarrow ALG$

    $ALG \leftarrow r_t$
  - *Else* $a_t \leftarrow a^*(h_t)$

- Claim: If for $t > N_0$:

$$\epsilon \leq \frac{1}{3} F(\{G_t > 0\})$$

the algorithm is BIC

Repeated Hidden Exploration

- Claim: If $\epsilon \leq \frac{1}{3} F(\{G_{N_0+1} > 0\})$ then for $t > N_0$: $\epsilon \leq \frac{1}{3} F(\{G_t > 0\})$
- Proof: We will show monotonicity
- $E[G_t | G_t > 0] = E[G_{t+1} | G_t > 0]$
- $$\begin{aligned} F(\{G_t > 0\}) &= E[G_t \cdot I\{G_t > 0\}] \\ &= E[G_{t+1} \cdot I\{G_t > 0\}] \\ &\leq E[G_{t+1} \cdot I\{G_{t+1} > 0\}] \\ &= F(\{G_{t+1} > 0\}) \end{aligned}$$

Repeated Hidden Exploration

- Regret Analysis
  - If ALG has Bayesian Regret $R(T) = \sqrt{T}$
  - Then RepeatedHiddenExploration has regret

$$R'(T) \leq N_0 + \frac{1}{\epsilon} E[R(N)] \approx \sqrt{T/\epsilon}$$

  - $N \approx \epsilon T$ number of exploration steps

Summary

- Adding incentives
- Two actions
  - Deterministic: optimal
  - Stochastic: Low regret
- Multiple actions
  - Deterministic: optimal policy?
  - Stochastic: same idea, low regret

Resources

- Optimal policy

  Deterministic actions
  - K=2 [Kremer, M, Perry, EC 2013 and JPE 2014]
  - $K \geq 3$ [Cohen, M EC 2019]
    - Limited domain

- Asymptotic Regret
  - Stochastic actions:
  - [M, Slivkins, Syrgkanis, EC 2015]
  - Multiple Agents:
  - [M, Slivkins, Syrgkanis, Wu, EC 2016]

- **Multiple Principals**
  - **[M, Slivkins, Wu, ITCS 2018]**

Bayesian Persuasion

- Kamenica & Gentzkow: AER 2011
- Two players:
  - principal and agent
- Agent selects action
  - Action effects both
- Principal selects information revelation
- How can the principal influence agent action?

- Example:
- Prosecutor and Judge
- Defendant:
  - guilty of innocent.
  - unobservable
- Trial:
  - Convicted or acquitted
- Prosecutor
  - max convictions
- Judge
  - minimizes errors

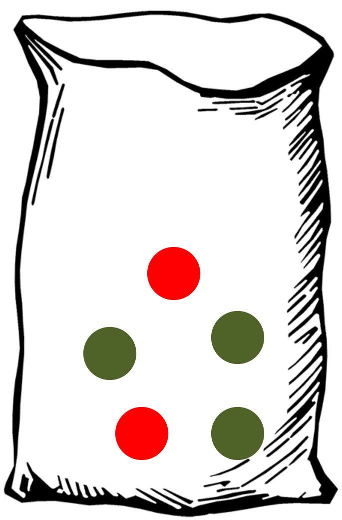
Bayesian Persuasion

- A priori 70% innocent
  - No information
    - judge equites
- Prosecutor
  - Controls which tests are done, and how
    - Information revelation
  - Selects a test s.t.
  - Pr[i | innocent]=4/7
  - Pr[i | innocent]=3/7
  - Pr[g | guilty] = 1

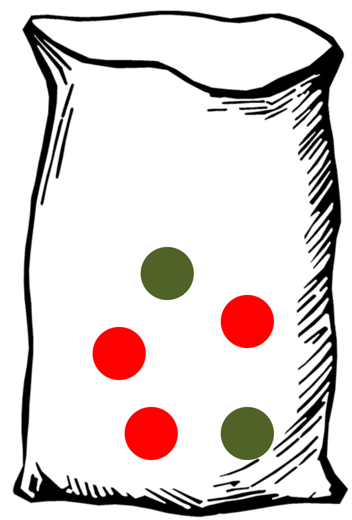
- Judge, given:
  - signal i: acquits
    - 40% defendants
    - All innocent
  - Signal g: convicts
    - 60% of defendants
    - Equally divided

- Although 30% guilty, 60% convicted !!!

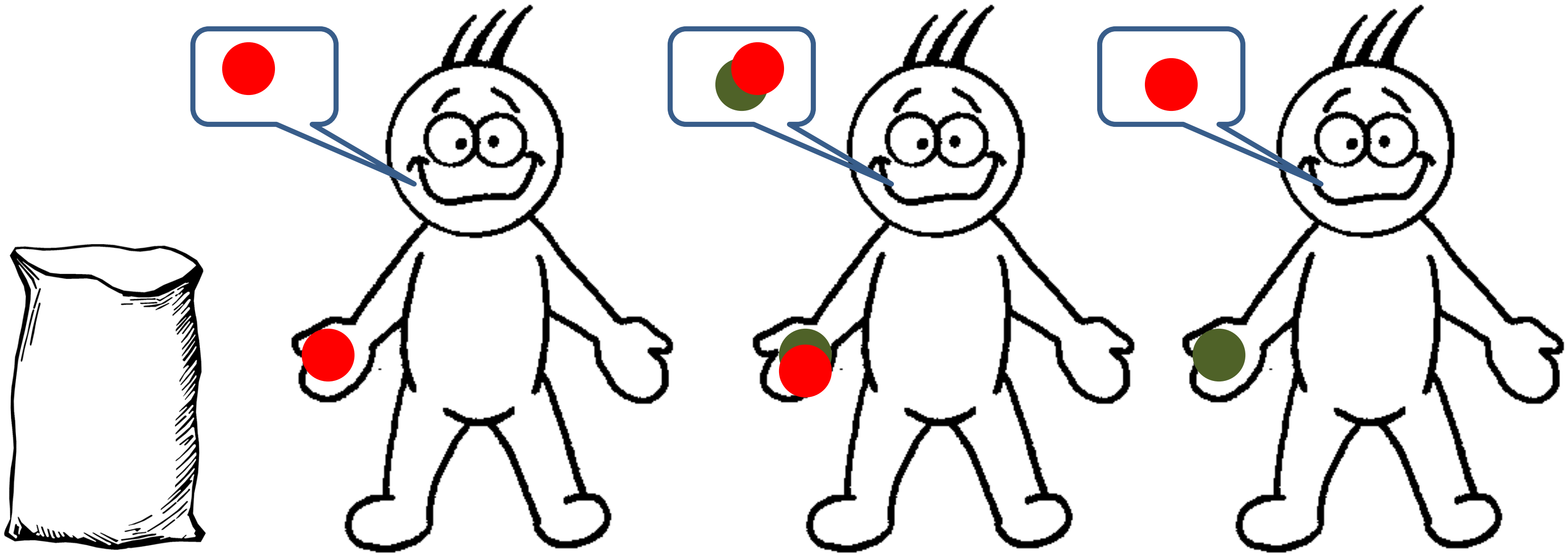
Information Cascading :



OR



Information Cascading



Agents ignore their input,
and information does not aggregate

Our Setting: Private recommendations