Incentivizing and Coordinating Exploration

Part II: Bayesian Models with Transfers

Bobby Kleinberg
Cornell University

ALT 2020 Tutorial
San Diego, 8 Feb 2020
Scope
- Mechanisms with monetary transfers
- Bayesian models of exploration
- Agents maximize expected utility minus cost
Scope
- Mechanisms with monetary transfers
- Bayesian models of exploration
- Agents maximize expected utility minus cost

Application: incentivizing “crowdsourced exploration”
E.g. online product recommendations with fully observable history.
Scope

- Mechanisms with monetary transfers
- Bayesian models of exploration
- Agents maximize expected utility minus cost

**Key abstraction:** Undiscounted terminal decision process (UTDP)

- Generalizes multi-armed bandits & Weitzman’s “box problem”
- A simple “index-based” policy is optimal.
- Proof introduces a key quantity: *deferred value*. [Weber, 1992]
  - Aids in adapting analysis to strategic settings.
Application 1: Multi-Armed Bandit

- One planner
- $n$ choices ("arms")

- Arm $i$ has random payoff sequence drawn i.i.d. from $F_i$
- Pull an arm: receive next element of payoff sequence.
- Maximize geometric discounted reward, $\sum_{t=0}^{\infty} (1 - \delta)^t r_t$. 
Application 2: Job Search

- One applicant
- $n$ firms
- Firm $i$ has interview cost $c_i$, match value $v_i \sim F_i$
- Special case of the “box problem”. [Weitzman, 1979]
Strategic issues

Firms compete to hire $\rightarrow$ inefficient investment in interviews.
Firms compete to hire $\rightarrow$ inefficient investment in interviews.

Competition $\rightarrow$ sunk cost.
Firms compete to hire $\rightarrow$ inefficient investment in interviews.
Competition $\rightarrow$ sunk cost.
Anticipating sunk cost $\rightarrow$ too few interviews.
Strategic issues

Firms compete to hire $\rightarrow$ inefficient investment in interviews.

Competition $\rightarrow$ sunk cost.

Anticipating sunk cost $\rightarrow$ too few interviews.
Strategic issues

Firms compete to hire → inefficient investment in interviews.
Competition → sunk cost.
Anticipating sunk cost → too few interviews.

Social learning → inefficient investment in exploration.
Each individual is myopic, prefers exploiting to exploring.
Strategic issues

“Arms” are strategic.

*Time steps are strategic.*
Undiscounted terminal decision process

Given *n* Markov chains, each with...
- state set $S_i$, terminal states $T_i \subset S_i$
- transition probabilities
- reward function $R_i : S_i \rightarrow \mathbb{R}$

Design policy $\pi$ that, in any state-tuple $(s_1, \ldots, s_n)$,
- chooses one Markov chain, $i$, to undergo state transition,
- receives reward $R(s_i)$

Stop the first time a MC enters a terminal state.

Maximize expected total reward.
Interview Markov Chain

-1

0  1  5  10  25

Interview
Evaluate
Hire
Interview UTDP
Interview UTDP
Interview UTDP
Interview UTDP
Interview UTDP
Interview UTDP
Multi-Stage Interview Markov Chain

Diagram:

-1

-5

0 5 10 25

Interview

Fly-Out

Evaluate

Hire
Multi-Armed Bandit as UTDP

Markov chain interpretation

State of an arm represents Bayesian posterior, given observations.

\[
\text{Beta}(1, 1)
\]

\[
\frac{1}{2}
\]
Multi-Armed Bandit as UTDP

Markov chain interpretation
State of an arm represents Bayesian posterior, given observations.

\[
\begin{align*}
\text{Beta}(2, 1) & \quad \text{Beta}(1, 2) \\
\frac{1}{3} & \quad \frac{2}{3}
\end{align*}
\]
Multi-Armed Bandit as UTDP

Markov chain interpretation
State of an arm represents Bayesian posterior, given observations.
Multi-Armed Bandit as UTDP

Markov chain interpretation
State of an arm represents Bayesian posterior, given observations.

$\beta(1, 1)$
$\beta(2, 1)$
$\beta(1, 2)$

$\delta$
Part 2:

Solving Undiscounted Terminal Decision Processes

Dumitriu, Tetali, & Winkler, *On Playing Golf with Two Balls*
Kleinberg, Waggoner, & Weyl, *Descending Price Optimally Coordinates Search*
Consider one Markov chain (arm) in isolation.

**Stopping game** $\Gamma(M, s, \sigma)$
- Markov chain $M$ starts in state $s$.
- In a non-terminal state $s'$, you may **continue** or **stop**.
- **Continue**: Receive payoff $R(s')$. Move to next state.
- **Stop**: game ends.
- In a terminal state, game ends and you pay penalty $\sigma$.

**Gittins index**
The *Gittins index* of (non-terminal) state $s$ is the maximum $\sigma$ such that the game $\Gamma(M, s, \sigma)$ has an optimal policy with positive probability of reaching a terminal state.
Consider one Markov chain (arm) in isolation.

The **Gittins index** of (non-terminal) state $s$ is the maximum $\sigma$ such that the game $\Gamma(M, s, \sigma)$ has an optimal policy with positive probability of reaching a terminal state.
Gittins Index and Deferred Value

Consider one Markov chain (arm) in isolation.

\[ \sigma(s_1) = 0 \]

\[ \sigma(s_2) = 5 \]

Gittins index

The **Gittins index** of (non-terminal) state \( s \) is the maximum \( \sigma \) such that the game \( \Gamma(M, s, \sigma) \) has an optimal policy with positive probability of reaching a terminal state.
Consider one Markov chain (arm) in isolation.

The Gittins index of (non-terminal) state $s$ is the maximum $\sigma$ such that the game $\Gamma(M, s, \sigma)$ has an optimal policy with positive probability of reaching a terminal state.
Consider one Markov chain (arm) in isolation.

\[ \sigma(s_0) = 2 \]
\[ \sigma(s_1) = 0 \]
\[ \sigma(s_2) = 5 \]

Deferred value

The \textit{deferred value} of Markov chain \( M \) is the random variable

\[ \kappa = \min_{1 \leq t < T} \{ \sigma(s_t) \} \]

where \( T \) is the time when the Markov chain enters a terminal state.
An optimal stopping rule for $\Gamma(\mathcal{M}, s_0, \sigma)$ must
- always stops in a state $s$ with $\sigma(s) < \sigma(s_0)$
- never stop in a state $s$ with $\sigma(s) > \sigma(s_0)$.
Amortization Lemma

Non-exposed stopping rules

A stopping rule for Markov chain $\mathcal{M}$ is *non-exposed* if it never stops in a state with $\sigma(s_\tau) > \min\{\sigma(s_t) | t < \tau\}$.

For a stopping rule $\tau$, define $\mathbb{A}(\tau)$ (abbreviated $\mathbb{A}$) by

$$
\mathbb{A}(\tau) = \begin{cases} 
1 & \text{if } s_\tau \in \mathcal{T} \\
0 & \text{otherwise.}
\end{cases}
$$

Assume Markov chain $\mathcal{M}$ satisfies

1. **Almost sure termination (AST):** With probability 1, the chain eventually enters a terminal state.

2. **No free lunch (NFL):** In any state $s$ with $R(s) > 0$, the probability of transitioning to a terminal state is positive.
Amortization Lemma

If Markov chain $\mathcal{M}$ satisfies AST and NFL, then every stopping rule $\tau$ satisfies $\mathbb{E} \left[ \sum_{0 < t < \tau} R(s_t) \right] \leq \mathbb{E}[\mathcal{A}_\kappa]$, with equality if the stopping rule is non-exposed.

Proof Sketch.

1. Time step $t$ is non-exposed if $\sigma(s_t) = \min\{\sigma(s_1), \ldots, \sigma(s_t)\}$.
2. Break time into “episodes”: subintervals consisting of one non-exposed step followed by zero or more exposed steps.
3. Prove the inequality by summing over episodes.
Amortization Lemma

If Markov chain $\mathcal{M}$ satisfies AST and NFL, then every stopping rule $\tau$ satisfies $\mathbb{E} \left[ \sum_{0 < t < \tau} R(s_t) \right] \leq \mathbb{E}[A_\kappa]$, with equality if the stopping rule is non-exposed.
A UTDP policy is optimal if and only if, in each state-tuple \((s_1, \ldots, s_n)\), it advances a Markov chain whose state \(s_i\) has maximum Gittins index, or if all indices are negative then it stops.

**Proof Sketch.** Gittins index policy induces a non-exposed stopping rule for each \(\mathcal{M}_i\) and always advances \(i^* = \arg\max_i \{\kappa_i\}\) into a terminal state unless \(\kappa_{i^*} < 0\). Hence

\[
\mathbb{E}[\text{Gittins}] = \mathbb{E}[\max_i (\kappa_i)^+] 
\]

whereas amortization lemma implies

\[
\mathbb{E}[\text{OPT}] \leq \mathbb{E}[\max_i (\kappa_i)^+].
\]
Part 3:

Social Learning

Frazier, Kempe, Kleinberg, & Kleinberg, *Incentivizing Exploration*. 
A Model Based on Multi-Armed Bandits

$k$ arms have independent random types that govern their (time-invariant) reward distribution when selected.

Users observe all past rewards before making their selection.
\( k \) arms have independent random types that govern their (time-invariant) reward distribution when selected.

Users observe all past rewards before making their selection.

Platform’s goal: maximize \( \sum_{t=0}^{\infty} (1 - \delta)^t r_t \)

User \( t \)’s goal: maximize \( r_t \)
Incentivized Exploration

Incentive payments

At time $t$, announce reward $c_{t,i} \geq 0$ for each arm $i$. User now chooses $i$ to maximize $\mathbb{E}[r_{i,t}] + c_{i,t}$.

Our platform and users have a common posterior at all times, so platform knows exactly which arm a user will pull, given a reward vector.

An equivalent description of our problem is thus:

- Platform can adopt any policy $\pi$.
- Cost of a policy pulling arm $i$ at time $t$ is $r_{t}^{\text{max}} - r_{i,t}$, where $r_{t}^{\text{max}}$ denotes myopically optimal reward.
Suppose, for platform’s policy $\pi$:
- reward $\geq (1 - a) \cdot \text{OPT}$.
- payment $\leq b \cdot \text{OPT}$.

We say $\pi$ achieves loss pair $(a, b)$.

**Definition**

$(a, b)$ is achievable if for every multi-armed bandit instance, $\exists$ policy achieving loss pair $(a, b)$. 

**Main Theorem**

Loss pair $(a, b)$ is achievable if and only if
$$\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}.$$
Suppose, for platform's policy $\pi$:
- reward $\geq (1 - a) \cdot \text{OPT}$.
- payment $\leq b \cdot \text{OPT}$.

We say $\pi$ achieves loss pair $(a, b)$.

**Definition**

$(a, b)$ is achievable if for every multi-armed bandit instance, $\exists$ policy achieving loss pair $(a, b)$.

**Main Theorem**

Loss pair $(a, b)$ is achievable if and only if $\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}$. 
Achievable region is convex, closed, upward monotone.

Main Theorem

Loss pair \((a, b)\) is achievable if and only if \(\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}\).
Incentive Cost

Achievable region is convex, closed, upward monotone.

Set-wise increasing in $\delta$.

Main Theorem

Loss pair $(a, b)$ is achievable if and only if $\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}$. 
The Achievable Region

Achievable region is convex, closed, upward monotone.
Set-wise increasing in $\delta$.
$(0.25,0.25)$ and $(0.1,0.5)$ achievable for all $\delta$.

Main Theorem

Loss pair $(a, b)$ is achievable if and only if $\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}$. 
Achievable region is convex, closed, upward monotone.

Set-wise increasing in $\delta$.

$(0.25,0.25)$ and $(0.1,0.5)$ achievable for all $\delta$.

You can always get $0.9 \cdot \text{OPT}$ while paying out only $0.5 \cdot \text{OPT}$.

**Main Theorem**

Loss pair $(a, b)$ is achievable if and only if $\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}$. 
A Hard Instance

Infinitely many “collapsing” arms $M$ with prob. $\frac{1}{M} \delta^2$, else 0.

*(Type fully revealed when pulled.)*
A Hard Instance

Infinitely many “collapsing” arms $M$ with prob. $\frac{1}{M} \delta^2$, else 0.

One arm whose payoff is always $\phi \cdot \delta$.

Extreme points of achievable region correspond to:

- **OPT**: pick a fresh collapsing arm until high payoff is found.
- **MYO**: always play the safe arm.
A Hard Instance

Infinitely many “collapsing” arms $M$ with prob. $\frac{1}{M} \delta^2$, else 0.

One arm whose payoff is always $\phi \cdot \delta$.

Extreme points of achievable region correspond to:

- **OPT**: reward $\approx 1$, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- **MYO**: always play the safe arm.
A Hard Instance

Infinitely many “collapsing” arms $M$ with prob. $\frac{1}{M} \delta^2$, else 0.

One arm whose payoff is always $\phi \cdot \delta$.

Extreme points of achievable region correspond to:

- **OPT**: reward $\approx 1$, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- **MYO**: reward $\phi$, cost 0. $(a, b) = (1 - \phi, 0)$
A Hard Instance

Infinitely many “collapsing” arms $M$ with prob. $\frac{1}{M} \delta^2$, else 0.

One arm whose payoff is always $\phi \cdot \delta$.

Extreme points of achievable region correspond to:

- **OPT**: reward $\approx 1$, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- **MYO**: reward $\phi$, cost 0. $(a, b) = (1 - \phi, 0)$
A Hard Instance

Infinitely many “collapsing” arms $M$ with prob. $\frac{1}{M} \delta^2$, else 0.

One arm whose payoff is always $\phi \cdot \delta$.

Extreme points of achievable region correspond to:

- **OPT:** reward $\approx 1$, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- **MYO:** reward $\phi$, cost 0. $(a, b) = (1 - \phi, 0)$
The line segment joining $(0, \phi - \delta)$ to $(1 - \phi, 0)$ is tangent to the curve $\sqrt{\frac{x}{1-\delta}} + \sqrt{\frac{y}{1-\delta}} = \sqrt{1-\delta}$ at

\[
x = \frac{1}{1-\delta} (1 - \phi)^2
\]
\[
y = \frac{1}{1-\delta} (\phi - \delta)^2
\]

- **OPT:** reward $\approx 1$, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- **MYO:** reward $\phi$, cost 0. $(a, b) = (1 - \phi, 0)$
Diamonds in the Rough

The line segment joining \((0, \phi - \delta)\) to \((1 - \phi, 0)\) is tangent to the curve \(\sqrt{x} + \sqrt{y} = \sqrt{1 - \delta}\) at

\[
x = \frac{1}{1-\delta}(1 - \phi)^2
\]

\[
y = \frac{1}{1-\delta}(\phi - \delta)^2
\]

- **OPT**: reward \(\approx 1\), cost \(\approx \phi - \delta\). \((a, b) = (0, \phi - \delta)\)
- **MYO**: reward \(\phi\), cost 0. \((a, b) = (1 - \phi, 0)\)
The inequality

\[ \sqrt{x} + \sqrt{y} \geq \sqrt{1 - \delta} \]

holds if and only if

\[ \forall p \in (0, 1) \quad \frac{x}{p} + \frac{y}{1 - p} \geq 1 - \delta \]

- **OPT:** reward \( \approx 1 \), cost \( \approx \phi - \delta \). \( (a, b) = (0, \phi - \delta) \)
- **MYO:** reward \( \phi \), cost 0. \( (a, b) = (1 - \phi, 0) \)
Proof of achievability is by contradiction. Suppose \((a, b)\) unachievable and \(\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}\). Then \(\exists p \in (0, 1)\) such that \(\forall\) achievable \(x, y\),

\[
\frac{x}{p} + \frac{y}{1-p} > 1 - \delta
\]

To reach a contradiction, must show that \(\forall p \in (0, 1)\) \(\exists \pi\) s.t.

\[
\mathbb{E} \left[ \frac{1}{p} \text{Payoff}(\pi) - \frac{1}{1-p} \text{Cost}(\pi) \right] \geq \frac{1}{p} - (1 - \delta).
\]
Lagrangean Relaxation

Proof of achievability is by contradiction. Suppose \((a, b)\) unachievable and \(\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}\).
Then \(\exists p \in (0, 1)\) such that \(\forall\) achievable \(x, y\),

\[
\frac{1-x}{p} - \frac{y}{1-p} < \frac{1}{p} - (1 - \delta)
\]

To reach a contradiction, must show that \(\forall p \in (0, 1)\) \(\exists \pi\) s.t.

\[
\mathbb{E} \left[ \frac{1}{p} \text{Payoff}(\pi) - \frac{1}{1-p} \text{Cost}(\pi) \right] \geq \frac{1}{p} - (1 - \delta).
\]
Time-Expanded Policy

We want a policy that makes \( \mathbb{E} \left[ \frac{1}{p} \text{Payoff}(\pi) - \frac{1}{1-p} \text{Cost}(\pi) \right] \) large.

The difficulty is \( \text{Cost}(\pi) \). Cost of pulling an arm depends on its state and on the state of the myopically optimal arm.

**Game plan.** Use randomization to bring about a cancellation that eliminates the dependence on the myopically optimal arm.

Play MYO with probability \( p \), \( \pi \) with probability \( 1 - p \).

If \( \pi \) is set to earn \( r \), MYO set to earn \( r + \Delta \), then

\[
\frac{1}{p} \text{Payoff} = \frac{r}{p} \text{ for sure, plus } \frac{\Delta}{p} \text{ with prob } p \\
\frac{1}{1-p} \text{Cost} = \frac{\Delta}{1-p} \text{ with prob } 1 - p.
\]
The time-expansion of policy $\pi$ with parameter $p$; $\text{TE}(\pi, p)$

Maintain a FIFO queue of states for each arm, tail is current state. At each time $t$, toss a coin with bias $p$.

**Heads:** Offer no incentive payments.
- User plays *myopically*. Push new state into tail of queue.

**Tails:** Apply $\pi$ to heads of queues to select arm.
- Push that arm’s new state into tail of queue, remove head.
- Pay user the difference vs. myopic.
The time-expansion of policy $\pi$ with parameter $p$; $\text{TE}(\pi, p)$

Maintain a FIFO queue of states for each arm, tail is current state. At each time $t$, toss a coin with bias $p$.

**Heads:** Offer no incentive payments.
   User plays *myopically*. Push new state into tail of queue.

**Tails:** Apply $\pi$ to heads of queues to select arm.
   Push that arm’s new state into tail of queue, remove head.
Pay user the difference vs. myopic.
The time-expansion of policy $\pi$ with parameter $p$; $\text{TE}(\pi, p)$

Maintain a FIFO queue of states for each arm, tail is current state.
At each time $t$, toss a coin with bias $p$.

**Heads:** Offer no incentive payments.
User plays myopically. Push new state into tail of queue.

**Tails:** Apply $\pi$ to heads of queues to select arm.
Push that arm’s new state into tail of queue, remove head.
Pay user the difference vs. myopic.
Time-Expanded Policy

The time-expansion of policy $\pi$ with parameter $p$; $\text{TE}(\pi, p)$

Maintain a FIFO queue of states for each arm, tail is current state. At each time $t$, toss a coin with bias $p$.

**Heads:** Offer no incentive payments.
- User plays *myopically*. Push new state into tail of queue.

**Tails:** Apply $\pi$ to heads of queues to select arm.
- Push that arm’s new state into tail of queue, remove head.
- Pay user the difference vs. myopic.
The time-expansion of policy $\pi$ with parameter $p$; $\text{TE}(\pi, p)$

Maintain a FIFO queue of states for each arm, tail is current state. At each time $t$, toss a coin with bias $p$.

**Heads:** Offer no incentive payments.
   - User plays *myopically*. Push new state into tail of queue.

**Tails:** Apply $\pi$ to heads of queues to select arm.
   - Push that arm’s new state into tail of queue, remove head.
   - Pay user the difference vs. myopic.
The time-expansion of policy $\pi$ with parameter $p$; $\text{TE}(\pi, p)$

Maintain a FIFO queue of states for each arm, tail is current state. At each time $t$, toss a coin with bias $p$.

**Heads:** Offer no incentive payments.
   User plays *myopically*. Push new state into tail of queue.

**Tails:** Apply $\pi$ to heads of queues to select arm.
   Push that arm’s new state into tail of queue, remove head.
   Pay user the difference vs. myopic.
### Time-Expanded Policy

<table>
<thead>
<tr>
<th>The time-expansion of policy $\pi$ with parameter $p$; $\text{TE}(\pi, p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintain a FIFO queue of states for each arm, tail is current state.</td>
</tr>
<tr>
<td>At each time $t$, toss a coin with bias $p$.</td>
</tr>
<tr>
<td><strong>Heads:</strong> Offer no incentive payments.</td>
</tr>
<tr>
<td>User plays <em>myopically</em>. Push new state into tail of queue.</td>
</tr>
<tr>
<td><strong>Tails:</strong> Apply $\pi$ to heads of queues to select arm.</td>
</tr>
<tr>
<td>Push that arm’s new state into tail of queue, remove head.</td>
</tr>
<tr>
<td>Pay user the difference vs. myopic.</td>
</tr>
</tbody>
</table>

**Lagrangian payoff analysis.** In a state where MYO would pick $i$ and $\pi$ would pick $j$, expected Lagrangian payoff is $\frac{r_{j,t}}{p}$. Due to cancellation of cost with MYO bonus.

If $s$ is at the head of $j$’s queue at time $t$, then $\mathbb{E}\left[\frac{r_{j,t}}{p} | s\right] = \frac{1}{p} R_j(s)$. 
Stuttering Arms

The “no-op” steps scale payoffs by $\frac{1}{p}$ and modify the Markov chain to have self-loops in every state with transition probability $(1 - \delta)p$.
Lemma

Let $\phi = 1 - (1 - \delta)p$. If $\tilde{\sigma}(s)$ denotes the Gittins index of state $s$ in the modified Markov chain, we have $\tilde{\sigma}(s) \geq \phi \cdot \sigma(s)$ for every $s$. 
Lemma

Let $\phi = 1 - (1 - \delta)p$. If $\tilde{\sigma}(s)$ denotes the Gittins index of state $s$ in the modified Markov chain, we have $\tilde{\sigma}(s) \geq \phi \cdot \sigma(s)$ for every $s$.

If true, this implies . . .

1. $\tilde{\kappa}_i \geq \phi \cdot \kappa_i$

2. Gittins index policy $\pi$ for modified Markov chains has expected payoff $\mathbb{E}[\max_i \tilde{\kappa}_i] \geq \phi \cdot \mathbb{E}[\max_i \kappa_i] = \phi$.

3. Policy $\text{TE}(\pi, p)$ achieves

$$\mathbb{E} \left[ \frac{1}{p} \text{Payoff} - \frac{1}{1-p} \text{Cost} \right] \geq \frac{\phi}{p} = \frac{1}{p} - (1 - \delta).$$

. . . which completes the proof of the main theorem.
Lemma

Let $\phi = 1 - (1 - \delta)p$. If $\tilde{\sigma}(s)$ denotes the Gittins index of state $s$ in the modified Markov chain, we have $\tilde{\sigma}(s) \geq \phi \cdot \sigma(s)$ for every $s$.

By definition of Gittins index, $\mathcal{M}$ has a stopping rule $\tau$ such that

$$\mathbb{E} \left[ \sum_{t < \tau} R(s_t) \right] \geq \sigma(s) \cdot \Pr(s_\tau \in \mathcal{T}) > 0.$$ 

Let $\tilde{\tau}$ be the equivalent stopping rule for $\tilde{\mathcal{M}}$, i.e. $\tilde{\tau}$ keeps going until it reaches a state in $\mathcal{T}$ or a state where $\tau$ stops.

**Observation 1:** $\tilde{\tau}$ stochastically dominates $\tau$, due to stuttering.

**Observation 2:** By Wald’s equation,

$$\mathbb{E}[\sum_{t < \tilde{\tau}} R(\tilde{s}_t)] = R(s) \cdot \mathbb{E}\tilde{\tau} \geq R(s) \cdot \mathbb{E}\tau = \mathbb{E}[\sum_{t < \tau} R(s_t)].$$
Lemma

Let $\phi = 1 - (1 - \delta)p$. If $\tilde{\sigma}(s)$ denotes the Gittins index of state $s$ in the modified Markov chain, we have $\tilde{\sigma}(s) \geq \phi \cdot \sigma(s)$ for every $s$.

\[ 1 - \phi = (1 - \delta)p \]

In every state $s_t$, $\Pr(s_t \to T \text{ in } \tilde{\mathcal{M}}) = \frac{\delta}{\phi} = \frac{1}{\phi} \Pr(s_t \to T \text{ in } \mathcal{M})$.

**Summary:** Comparing $\tilde{\tau}$ in $\tilde{\mathcal{M}}$ with $\tau$ in $\mathcal{M}$,

- expected reward is weakly greater,
- expected penalty is scaled by $\frac{1}{\phi}$

If penalty is $\phi \cdot \sigma(s)$, $\tilde{\tau}$ at least breaks even $\implies \tilde{\sigma}(s) \geq \phi \cdot \sigma(s)$
Summary of Main Result

Incentive Cost

Opportunity Cost

Main Theorem

Loss pair \((a, b)\) is achievable if and only if \(\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}\).

- Principal can always achieve 90% of social surplus while paying back only 50% to users via incentive payments.
- Simple policies that randomize between \textit{laissez-faire} and providing incentives for optimal learning are approx. optimal.
- Worst-case instances comprise “diamonds in the rough” alongside a safe alternative.
“Live as if you were to die tomorrow. Learn as if you were to live forever.”
“Live and learn as if you were to die tomorrow with probability $p$ and to live forever with probability $1 - p$.\"
The foregoing model assumed agents have identical preferences. Han & Kempe (WINE 2015) extended this to a model where

- agents’ preferences over arms are identical
- agents’ exchange rates between money and arm-utility vary
- principal receives noisy signal of exchange rate.

Achievable region can still be characterized.

- “Diamonds in the rough” are still worst-case.
- Time-expansions of OPT still attain every achievable loss pair.
- In time expansion, \( p = \Pr(MYO) \) becomes signal dependent.
Chen, Frazier, and Kempe (COLT 2018) analyzed a model where

- agents’ preferences over arms vary
- arms, agents characterized by attribute vectors
- utility = dot product plus mean-zero sub-Gaussian noise
- pull an arm ⇒ observe noisy attribute vector
- agent attributes never observed

They presented a policy with regret $O(Ne^{2/p} + LN \log^3(T))$, assuming $N$ arms, each favored by at least $p$ fraction of agents.

$L$ denotes “density of near ties”:

$$\Pr(\text{diff btw best, 2nd-best arms } \leq \varepsilon) = O(L\varepsilon) \text{ as } \varepsilon \to 0$$

**Main theme:** exploration comes for free when agents prefer different arms.
Strategic Arms

Arms (e.g. firms) strategize about when/if they are pulled.

Optimal mechanisms are complicated.

- Agents report all private information to center . . .
- which runs Gittins index policy . . .
- and charges VCG payments.

Descending-price mechanisms have small constant price of anarchy. [Kleinberg, Waggoner, & Weyl 2016]

- proof uses deferred value amortization lemma, “smoothness” arguments
- extends to combinatorial domains, e.g. matchings, matroids (cf. follow-up work by Sahil Singla [2018, 2019])
- sequential posted price mechanisms also have constant PoA; analysis combines deferred values with prophet inequalities.
Conclusion

- **Undiscounted terminal decision processes**: versatile model of information acquisition in Bayesian settings
  - ...when time steps are strategic
  - ...when alternatives ("arms") are strategic.
- Optimal policy, absent incentive issues: *Gittins index policy*.
- Analysis tool: *deferred value* and *amortization lemma*.
  - Interfaces cleanly with equilibrium analysis of simple mechanisms, smoothness arguments, prophet inequalities, etc.
  - **Beautiful but fragile**: usefulness vanishes rapidly as you vary the assumptions.
Open questions: incentivizing exploration using money

- **Extend to contextual bandits**
  - No Gittins index theorem for Bayesian contextual bandits!
  - Can we characterize achievable region anyway?
  - Success story: Han & Kempe (2015) solves a very special case

- **Hard constraints** on budget and/or social welfare

- **Non-martingale arms**, e.g. pulling an arm represents one round of training a person or an ML model

- **Non-myopic agents**
  - Repeat customers have some incentive to explore arms . . .
  - but free ridership is still a problem.
  - Is finite population always better than the infinite-population limit? Can we quantify how much better?

- **Combine incentive payments and information design**
  - Agents don’t observe history, as in first half of tutorial.
  - Could small incentive payments yield a huge gain in regret?