Incentivizing and Coordinating Exploration
Part II: Bayesian Models with Transfers

Bobby Kleinberg

Cornell University

EC 2017 Tutorial
27 June 2017

Preview of this lecture

Scope

- Mechanisms with monetary transfers
- Bayesian models of exploration
- Risk-neutral, quasi-linear utility

Preview of this lecture

Scope

- Mechanisms with monetary transfers
- Bayesian models of exploration
- Risk-neutral, quasi-linear utility

Applications

- Markets/auctions with costly information acquisition
    - E.g. job interviews, home inspections, start-up acquisitions


Preview of this lecture

Scope

- Mechanisms with monetary transfers
- Bayesian models of exploration
- Risk-neutral, quasi-linear utility

Applications

- Incentivizing “crowdsourced exploration”
  - E.g. online product recommendations, citizen science.

Preview of this lecture

Scope

- Mechanisms with monetary transfers
- Bayesian models of exploration
- Risk-neutral, quasi-linear utility

Key abstraction: *joint Markov scheduling*

- 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.
  - Role similar to virtual values in optimal auction design.

Application 1: 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]

Application 2: Multi-Armed Bandit

- **One planner**
- **n choices (“arms”)**



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

Strategic issues

Firms compete to hire → inefficient investment in interviews.

Strategic issues

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

Competition $\rightarrow$ sunk cost.

Strategic issues

Firms compete to hire → inefficient investment in interviews.

Competition → sunk cost.

Anticipating sunk cost → 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.

Joint Markov Scheduling

Given n Markov chains, each with . . .

- state set $\mathcal{S}_i$, terminal states $\mathcal{T}_i \subset \mathcal{S}_i$
- transition probabilities
- reward function $R_i : \mathcal{S}_i \to \mathbb{R}$

Design policy π that, in any state-tuple $(s_1, \ldots, s_n)$,

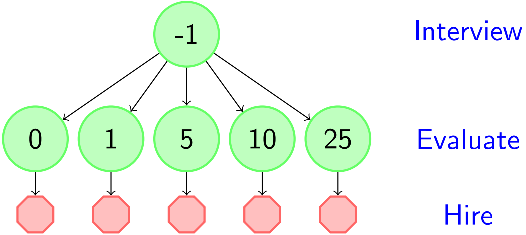
- chooses one Markov chain, i, to undergo state transition,
- receives reward $R(s_i)$

Stop the first time a MC enters a terminal state.

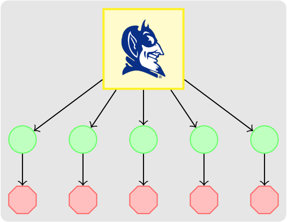
Maximize expected total reward.[1]

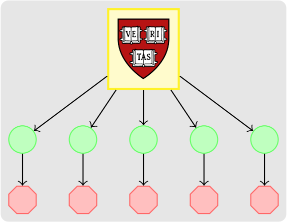
[1] Dumitriu, Tetali, & Winkler, *On Playing Golf with Two Balls*.

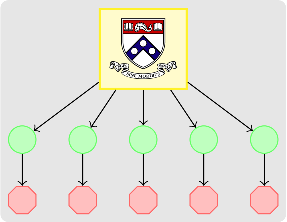
Interview Markov Chain



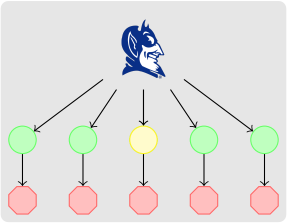
Joint Markov Scheduling of Interviews

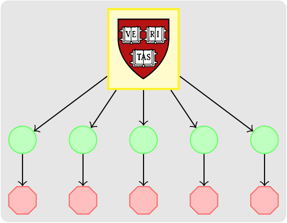


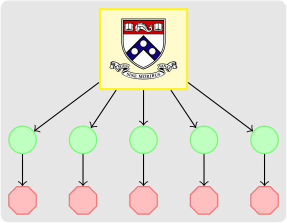




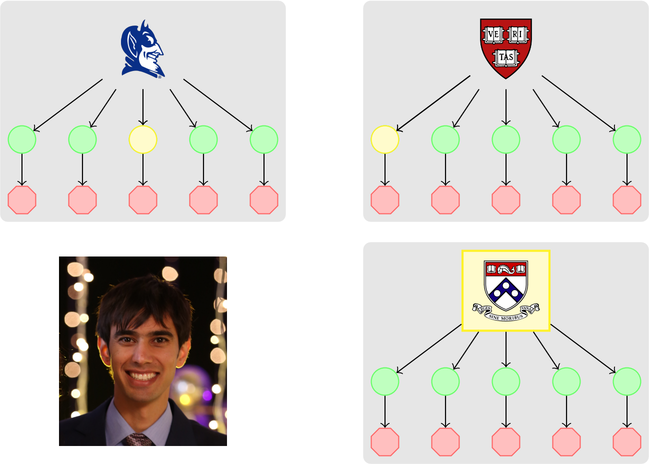
Joint Markov Scheduling of Interviews



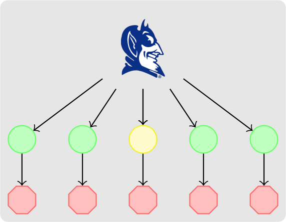


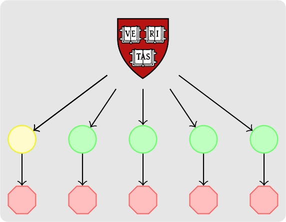


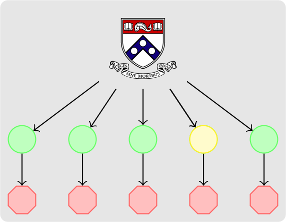
Joint Markov Scheduling of Interviews



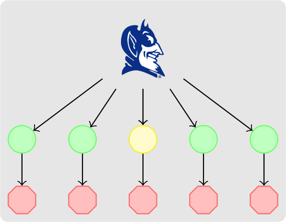
Joint Markov Scheduling of Interviews

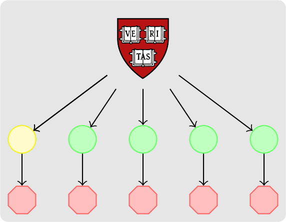


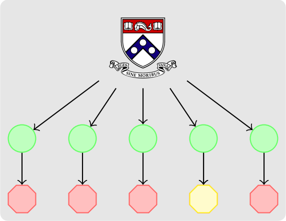




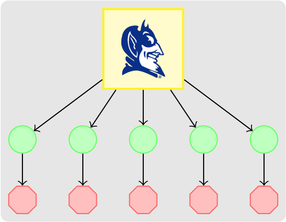
Joint Markov Scheduling of Interviews

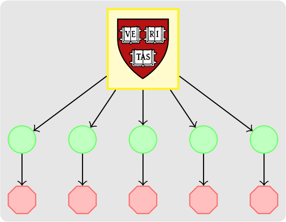


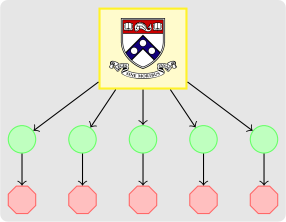




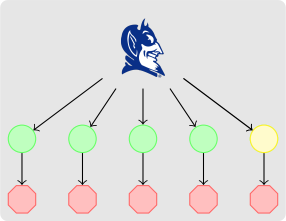
Joint Markov Scheduling of Interviews

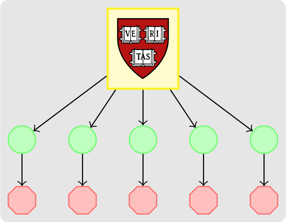


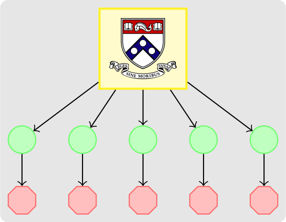




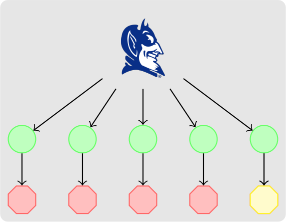
Joint Markov Scheduling of Interviews

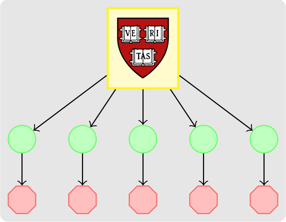


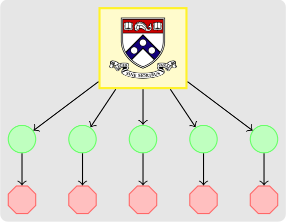




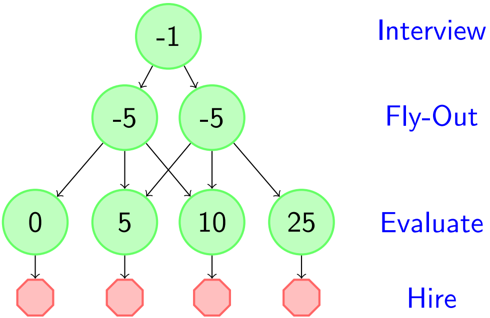
Joint Markov Scheduling of Interviews







Multi-Stage Interview Markov Chain



Multi-Armed Bandit as Markov Scheduling

Markov chain interpretation

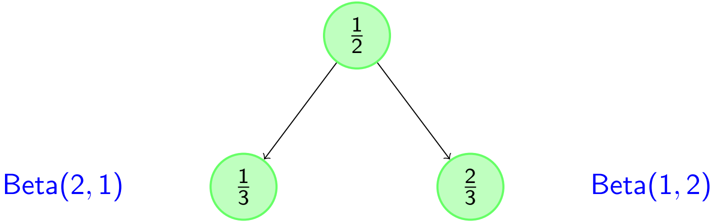
State of an arm represents Bayesian posterior, given observations.

$\text{Beta}(1,1)$ $\frac{1}{2}$

Multi-Armed Bandit as Markov Scheduling

Markov chain interpretation

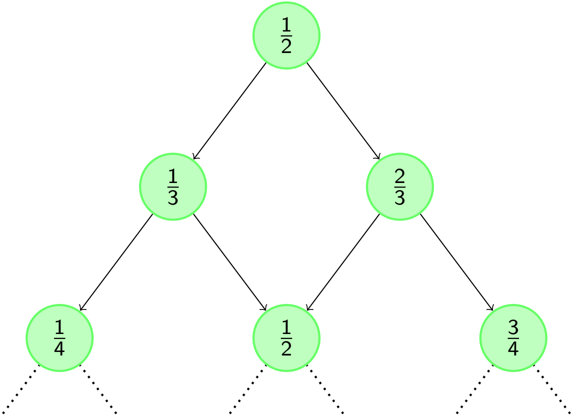
State of an arm represents Bayesian posterior, given observations.



Multi-Armed Bandit as Markov Scheduling

Markov chain interpretation

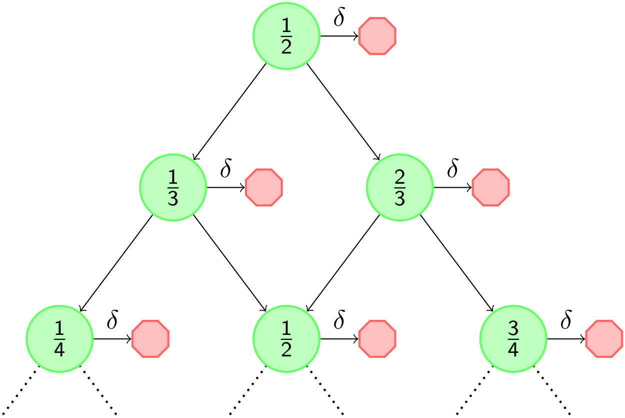
State of an arm represents Bayesian posterior, given observations.



Multi-Armed Bandit as Markov Scheduling

Markov chain interpretation

State of an arm represents Bayesian posterior, given observations.



Part 2:

Solving Joint Markov Scheduling

Naïve Greedy Methods Fail

An example due to Weitzman (1979) . . .

$c_i = 15$

$$v_i = \begin{cases} 100 & \text{w. prob } \frac{1}{2} \\ 55 & \text{otherwise} \end{cases}$$

$c_i = 20$

$$v_i = \begin{cases} 240 & \text{w. prob } \frac{1}{5} \\ 0 & \text{otherwise} \end{cases}$$

- Red is better in expectation and in worst case, less costly.
- Nevertheless, optimal policy starts by trying blue.

Solution to The Box Problem

For each box i, let σ_i be the (unique, if $c_i > 0$) solution to

$$\mathbb{E}\left[(v_i - \sigma_i)^+\right] = c_i$$

where $(\cdot)^+$ denotes $\max\{\cdot, 0\}$.

Interpretation: for an asset with value $v_i \sim F_i$, the fair value of a call option with **strike price** σ_i is c_i.

Optimal policy: Descending Strike Price (DSP)

1. Maintain priority queue, initially ordered by strike price.
2. Repeatedly extract highest-priority box from queue.
3. If closed, open it and reinsert into queue with priority v_i.
4. If open, choose it and terminate the search.

Solution to The Box Problem

For each box i, let σ_i be the (unique, if $c_i > 0$) solution to

$$\mathbb{E}\left[(v_i - \sigma_i)^+\right] = c_i$$

Cost $= 15$

$$\text{Prize} = \begin{cases} 100 & \text{w. prob } \frac{1}{2} \\ 55 & \text{otherwise} \end{cases}$$

$\sigma_{\text{red}} = 70$

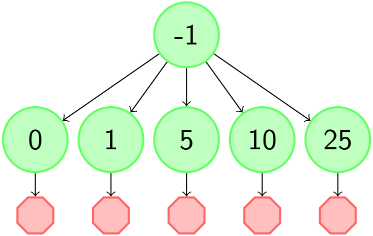
Cost $= 20$

$$\text{Prize} = \begin{cases} 240 & \text{w. prob } \frac{1}{5} \\ 0 & \text{otherwise} \end{cases}$$

$\sigma_{\text{blue}} = 140$

Non-Exposed Stopping Rules

Recall: Markov chain corresponding to Box i has three types of states.



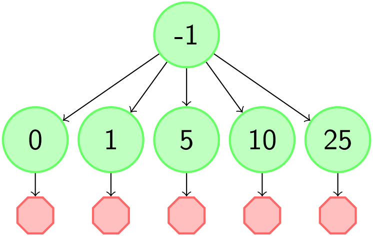
Initial: v_i unknown

Intermediate:
v_i known, payoff $-c_i$

Terminal: payoff $v_i - c_i$

Non-Exposed Stopping Rules

Recall: Markov chain corresponding to Box i has three types of states.



Initial: v_i unknown

Intermediate:
v_i known, payoff $-c_i$

Terminal: payoff $v_i - c_i$

Non-exposed stopping rules

A stopping rule is *non-exposed* if it never stops in an intermediate state with $v_i > \sigma_i$.

Amortization Lemma

Covered call value (of box i)

The *covered call value* is the random variable $\kappa_i = \min\{v_i, \sigma_i\}$.

Amortization Lemma

Covered call value (of box i)

The *covered call value* is the random variable $\kappa_i = \min\{v_i, \sigma_i\}$.

For a stopping rule τ let

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

Inspect   Acquire

Abbreviate as $\mathbb{I}_i$, $\mathbb{A}_i$, when τ is clear from context.

Amortization Lemma

Covered call value (of box i)

The *covered call value* is the random variable $\kappa_i = \min\{v_i, \sigma_i\}$.

Amortization Lemma

For every stopping rule τ, $\mathbb{E}\left[\mathbb{A}_i v_i - \mathbb{I}_i c_i\right] \leq \mathbb{E}\left[\mathbb{A}_i \kappa_i\right]$ with equality if and only if the stopping rule is non-exposed.

Amortization Lemma

Covered call value (of box i)

The *covered call value* is the random variable $\kappa_i = \min\{v_i, \sigma_i\}$.

Amortization Lemma

For every stopping rule τ, $\mathbb{E}\left[\mathbb{A}_i v_i - \mathbb{I}_i c_i\right] \leq \mathbb{E}\left[\mathbb{A}_i \kappa_i\right]$ with equality if and only if the stopping rule is non-exposed.

Proof sketch: If you already hold the asset, adopting the *covered call position* (selling the call option at price c_i) is:

- risk-neutral
- strictly beneficial if the buyer of the option sometimes forgets to “exercise in the money”.

Proof of Amortization

Amortization Lemma

For every stopping rule τ, $\mathbb{E}\left[\mathbb{A}_i v_i - \mathbb{I}_i c_i\right] \leq \mathbb{E}\left[\mathbb{A}_i \kappa_i\right]$ with equality if and only if the stopping rule is non-exposed.

Proof.

$$\begin{aligned}
\mathbb{E}\left[\mathbb{A}_i v_i - \mathbb{I}_i c_i\right] &= \mathbb{E}\left[\mathbb{A}_i v_i - \mathbb{I}_i (v_i - \sigma_i)^+\right] && (1)\\
&\leq \mathbb{E}\left[\mathbb{A}_i \left(v_i - (v_i - \sigma_i)^+\right)\right] && (2)\\
&= \mathbb{E}\left[\mathbb{A}_i \kappa_i\right]. && (3)
\end{aligned}$$

Inequality (2) is justified because $(\mathbb{I}_i - \mathbb{A}_i)(v_i - \sigma_i)^+ \geq 0$.
Equality holds if and only if τ is non-exposed. □

Optimality of Descending Strike Price Policy

Any policy induces an n-tuple of stopping rules, one for each box. Let

$$\tau_1^*, \ldots, \tau_n^* = \{\text{stopping rules for OPT}\}$$
$$\tau_1, \ldots, \tau_n = \{\text{stopping rules for DSP}\}$$

Then

$$\mathbb{E}\left[\text{OPT}\right] \leq \sum_i \mathbb{E}\left[\mathbb{A}_i(\tau_i^*)\kappa_i\right] \leq \mathbb{E}\left[\max_i \kappa_i\right]$$
$$\mathbb{E}\left[\text{DSP}\right] = \sum_i \mathbb{E}\left[\mathbb{A}_i(\tau_i)\kappa_i\right] = \mathbb{E}\left[\max_i \kappa_i\right]$$

because DSP is non-exposed and always selects the maximum κ_i.

Gittins Index and Deferred Value

Consider one Markov chain (arm) in isolation.

Stopping game $\Gamma(\mathcal{M}, s, \sigma)$

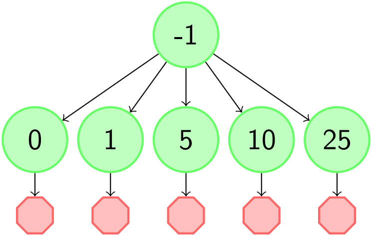
- Markov chain $\mathcal{M}$ starts in state s.
- In a non-terminal state s', you may continue or stop.
- Continue: Receive payoff $R(s')$. Move to next state.
- Stop: game ends.
- In a terminal state, game ends and you pay penalty σ.

Gittins index

The *Gittins index* of (non-terminal) state s is the maximum σ such that the game $\Gamma(\mathcal{M}, s, \sigma)$ has an optimal policy with positive probability of stopping in a terminal state.

Gittins Index and Deferred Value

Consider one Markov chain (arm) in isolation.

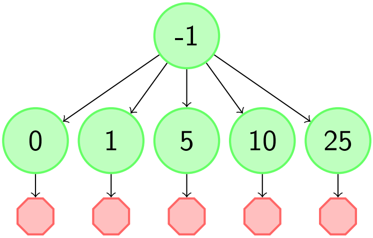


Gittins index

The *Gittins index* of (non-terminal) state s is the maximum σ such that the game $\Gamma(\mathcal{M}, s, \sigma)$ has an optimal policy with positive probability of stopping in a terminal state.

Gittins Index and Deferred Value

Consider one Markov chain (arm) in isolation.



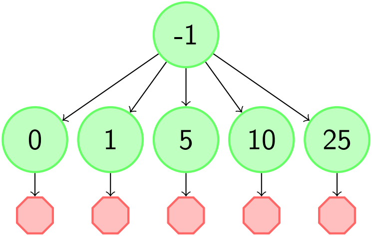
$\sigma(s) = v_i$

Gittins index

The *Gittins index* of (non-terminal) state s is the maximum σ such that the game $\Gamma(\mathcal{M}, s, \sigma)$ has an optimal policy with positive probability of stopping in a terminal state.

Gittins Index and Deferred Value

Consider one Markov chain (arm) in isolation.



$\sigma(s) = \sigma_i$

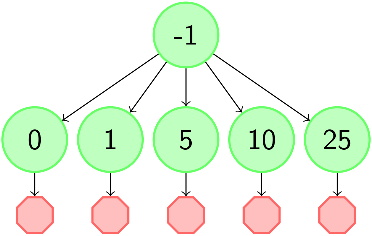
$\sigma(s) = v_i$

Gittins index

The *Gittins index* of (non-terminal) state s is the maximum σ such that the game $\Gamma(\mathcal{M}, s, \sigma)$ has an optimal policy with positive probability of stopping in a terminal state.

Gittins Index and Deferred Value

Consider one Markov chain (arm) in isolation.



$\sigma(s) = \sigma_i$

$\sigma(s) = v_i$

Deferred value

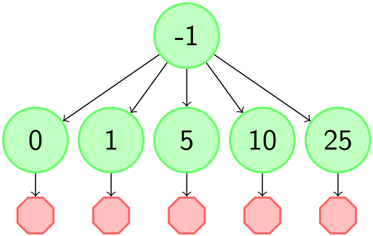
The *deferred value* of Markov chain $\mathcal{M}$ is the random variable

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

where T is the time when the Markov chain enters a terminal state.

Gittins Index and Deferred Value

Consider one Markov chain (arm) in isolation.



$\sigma(s) = \sigma_i$

$\sigma(s) = v_i$

$\kappa = \min\{v_i, \sigma_i\}$

Deferred value

The *deferred value* of Markov chain $\mathcal{M}$ 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.

General 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 $\sigma(s_\tau) > \min\{\sigma(s_t) \mid t < \tau\}$.

For a stopping rule τ, 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.

General Amortization Lemma

Amortization Lemma

If Markov chain $\mathcal{M}$ satisfies AST and NFL, then every stopping rule τ satisfies $\mathbb{E}\left[\sum_{0<t<\tau} R(s_t)\right] \leq \mathbb{E}[\mathbb{A}\kappa]$, 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.

Gittins Index Theorem

Gittins Index Theorem

A joint Markov scheduling 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 Gittins 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 $i^* = \mathrm{argmax}_i\{\kappa_i\}$ into a terminal state unless $\kappa_{i^*} < 0$. Hence

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

whereas amortization lemma implies

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

Joint Markov Scheduling, General Case

Feasibility constraint $\mathcal{I}$: a collection of subsets of $[n]$.

Joint Markov scheduling w.r.t. $\mathcal{I}$: when the policy stops, the set of Markov chains in terminal states must belong to $\mathcal{I}$.[2]

Theorem (Gittins Index Theorem for Matroids)

Let $\mathcal{I}$ be a matroid. A policy for joint Markov scheduling w.r.t. $\mathcal{I}$ is optimal iff, in each state-tuple $(s_1, \ldots, s_n)$, the policy advances $\mathcal{M}_i$ whose state s_i has maximum Gittins index, among those i such that $\{i\} \cup \{j \mid s_j \text{ is a terminal state}\} \in \mathcal{I}$, or stops if $\sigma(s_i) < 0$.

Proof sketch: Same proof as before. The policy described is non-exposed and simulates the greedy algorithm for choosing a max-weight independent set w.r.t. weights $\{\kappa_i\}$.

[2] Sahil Singla, *The Price of Information in Combinatorial Optimization*, contains further generalizations.

Joint Markov Scheduling, General Case

Feasibility constraint $\mathcal{I}$: a collection of subsets of $[n]$.

Joint Markov scheduling w.r.t. $\mathcal{I}$: when the policy stops, the set of Markov chains in terminal states must belong to $\mathcal{I}$.[2]

Box Problem for Matchings

Put "Weitzman boxes" on the edges of a bipartite graph, and allow picking any set of boxes that forms a matching.

Simulating greedy max-weight matching with weights $\{\kappa_i\}$ yields a 2-approximation to the optimum policy.

Simulating exact max-weight matching yields no approximation guarantee. (Violates the non-exposure property, because an augmenting path may eliminate an open box with $v_i > \sigma_i$.)

[2]Sahil Singla, *The Price of Information in Combinatorial Optimization*, contains further generalizations.

Exogenous Box Order

Suppose boxes are presented in order $1, \ldots, n$. We only choose *whether* to open box i, not *when* to open it.

Theorem

There exists a policy for the box problem with exogenous order, whose expected value is at least half that of the optimal policy with endogenous order.

Proof sketch. $\kappa_1, \ldots, \kappa_n$ are independent random variables. Prophet inequality $\Rightarrow$ threshold stop rule τ such that

$$\mathbb{E}[\kappa_\tau] \geq \frac{1}{2}\mathbb{E}[\max_i \kappa_i].$$

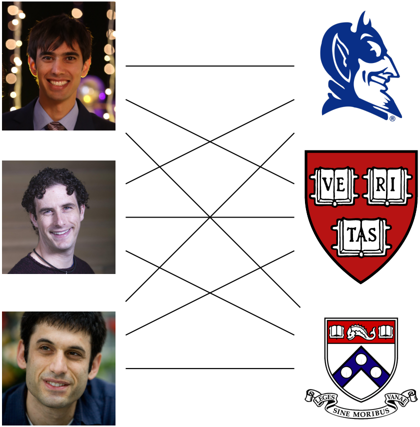
Threshold stop rules are non-exposed: open box if $\sigma_i \geq \theta$, select it if $v_i \geq \theta$.

Part 3:

Information Acquisition in Markets

Auctions with Costly Information Acquisition

- m heterogeneous items for sale
- n bidders: unit demand, risk neutral, quasi-linear utility



Auctions with Costly Information Acquisition

- m heterogeneous items for sale
- n bidders: unit demand, risk neutral, quasi-linear utility
- Bidder i has private type $\theta_i \in \Theta_i$.
- Value of item j to bidder i given $\theta = \theta_i$ is $v_{ij} \sim F_{\theta j}$.

Auctions with Costly Information Acquisition

- m heterogeneous items for sale
- n bidders: unit demand, risk neutral, quasi-linear utility
- Bidder i has private type $\theta_i \in \Theta_i$.
- Value of item j to bidder i given $\theta = \theta_i$ is $v_{ij} \sim F_{\theta j}$.
- Inspection: bidder i must pay cost $c_{ij}(\theta_i) \geq 0$ to learn v_{ij}. Unobservable. Cannot acquire item without inspecting.

Auctions with Costly Information Acquisition

- m heterogeneous items for sale
- n bidders: unit demand, risk neutral, quasi-linear utility
- Bidder i has private type $\theta_i \in \Theta_i$.
- Value of item j to bidder i given $\theta = \theta_i$ is $v_{ij} \sim F_{\theta j}$.
- Inspection: bidder i must pay cost $c_{ij}(\theta_i) \geq 0$ to learn v_{ij}. Unobservable. Cannot acquire item without inspecting.
- Types may be correlated
- $\{v_{ij}\}$ are conditionally independent given types, costs.

Auctions with Costly Information Acquisition

- m heterogeneous items for sale
- n bidders: unit demand, risk neutral, quasi-linear utility
- Bidder i has private type $\theta_i \in \Theta_i$.
- Value of item j to bidder i given $\theta = \theta_i$ is $v_{ij} \sim F_{\theta j}$.
- Inspection: bidder i must pay cost $c_{ij}(\theta_i) \geq 0$ to learn v_{ij}. Unobservable. Cannot acquire item without inspecting.
- Types may be correlated
- $\{v_{ij}\}$ are conditionally independent given types, costs.

Extension

Inspection happens in stages indexed by $k \in \mathbb{N}$. Each reveals a new signal about v_{ij}. Cost to observe first k signals is $c_{ij}^k(\theta_i)$.

Simultaneous Auctions (Single-item Case)

If inspections must happen before auction begins, 2nd-price auction maximizes expected welfare. [Bergemann & Välimäki, 2002]

May be arbitrarily inefficient relative to best sequential procedure.

- n identical bidders: cost $c = 1-\delta$, value $\begin{cases} H & \text{with prob. } \frac{1}{H} \\ 0 & \text{otherwise.} \end{cases}$
- Take limit as $H \to \infty$, $\frac{n}{H} \to \infty$, $\delta \to 0$.
- First-best procedure gets $H(1-c) = H \cdot \delta$.
- For any simultaneous-inspection procedure ...
  - Let $p_i = \Pr(i \text{ inspects})$, $x = \sum_{i=1}^{n} p_i$.
  - Cost is cx. Benefit is $\lesssim H \left(1 - e^{-x/H}\right)$.
  - Difference is maximized at $x \cong H \ln(1/c) \cong H \cdot \delta$.
  - Welfare $\lesssim H \cdot \delta^2$.

Efficient Dynamic Auctions

If a dynamic auction is efficient, it must

- Implement the first-best policy. (DSP or Gittins index)
- Charge agents using Groves payments.

Seminal papers on dynamic auctions [Cavallo, Parkes, & Singh 2006; Crémer, Spiegel, & Zheng, 2009; Bergemann & Välimäki 2010; Athey & Segal 2013] specify how to do this.

(Varying information structures and participation constraints.)

Efficient Dynamic Auctions

If a dynamic auction is efficient, it must

- Implement the first-best policy. (DSP or Gittins index)
- Charge agents using Groves payments.

Seminal papers on dynamic auctions [Cavallo, Parkes, & Singh 2006; Crémer, Spiegel, & Zheng, 2009; Bergemann & Välimäki 2010; Athey & Segal 2013] specify how to do this.

(Varying information structures and participation constraints.)

Any such mechanism requires either:

- agents communicate their entire value distribution
- the center knows agents' value distributions without having to be told.

Efficient dynamic auctions rarely seen in practice.

Descending Auction

Descending-Price Mechanism

Descending clock represents uniform price for all items. Bidders may claim any remaining item at the current price.

Intuition: parallels descending strike price policy.
Bidders with high "option value" can inspect early. If value is high, can claim item immediately to avoid competition.

Descending Auction

Descending-Price Mechanism

Descending clock represents uniform price for all items. Bidders may claim any remaining item at the current price.

Intuition: parallels descending strike price policy.
Bidders with high "option value" can inspect early. If value is high, can claim item immediately to avoid competition.

Theorem

For single-item auctions, any n-tuple of bidders has an n-tuple of "counterparts" who know their valuations. Equilibria of descending-price auction correspond to equilibria of 1^{st}*-price auction among counterparts.*

Descending Auction

Descending-Price Mechanism

Descending clock represents uniform price for all items. Bidders may claim any remaining item at the current price.

Intuition: parallels descending strike price policy.

Bidders with high "option value" can inspect early. If value is high, can claim item immediately to avoid competition.

Theorem

For multi-item auctions with unit-demand bidders, every descending-price auction equilibrium achieves at least 43% of first-best welfare.

Descending-Price Auction: Single-Item Case

Definition (Covered counterpart)

For each bidder i define their *covered counterpart* to have zero inspection cost and value κ_i.

Equilibrium Correspondence Theorem

For single-item auctions there is an expected-welfare preserving one-to-one correspondence

$$\{\text{Equilibria of descending price auction with } n \text{ bidders}\}$$
$$\Updownarrow$$
$$\{\text{Equilibria of } 1^{\text{st}} \text{ price auction with their covered counterparts}\}.$$

Proof of Equilibrium Correspondence

Consider the best responses of bidder i and covered counterpart i' when facing any strategy profile b_{-i}.

Suppose counterpart's best response is to buy item at time $b_i'(\kappa_i)$.

Proof of Equilibrium Correspondence

Consider the best responses of bidder i and covered counterpart i' when facing any strategy profile b_{-i}.

Suppose counterpart's best response is to buy item at time $b'_i(\kappa_i)$.

Bidder i can emulate this using the following strategy b_i:

- Inspect at price $b'_i(\sigma_i)$.
- Buy immediately if $v_i \geq \sigma_i$.
- Else buy at price $b'_i(v_i)$.

Proof of Equilibrium Correspondence

Consider the best responses of bidder i and covered counterpart i' when facing any strategy profile b_{-i}.

Suppose counterpart's best response is to buy item at time $b_i'(\kappa_i)$.

Bidder i can emulate this using the following strategy b_i:

- Inspect at price $b_i'(\sigma_i)$.
- Buy immediately if $v_i \geq \sigma_i$.
- Else buy at price $b_i'(v_i)$.

This strategy b_i is non-exposed, so $\mathbb{E}\left[u_i(b_i, b_{-i})\right] = \mathbb{E}\left[u_i'(b_i', b_{-i})\right].$

Proof of Equilibrium Correspondence

Consider the best responses of bidder i and covered counterpart i' when facing any strategy profile b_{-i}.

Suppose counterpart's best response is to buy item at time $b'_i(\kappa_i)$.

Bidder i can emulate this using the following strategy b_i:

- Inspect at price $b'_i(\sigma_i)$.
- Buy immediately if $v_i \geq \sigma_i$.
- Else buy at price $b'_i(v_i)$.

This strategy b_i is non-exposed, so $\mathbb{E}\left[u_i(b_i, b_{-i})\right] = \mathbb{E}\left[u'_i(b'_i, b_{-i})\right].$

No other strategy $\tilde{b}_i$ is better for i, because

$$\mathbb{E}\left[u_i(\tilde{b}_i, b_{-i})\right] \leq \mathbb{E}\left[\text{covered call value minus price}\right]$$
$$= \mathbb{E}\left[u'_i(\tilde{b}_i, b_{-i})\right] \leq \mathbb{E}\left[u'_i(b'_i, b_{-i})\right].$$

Welfare and Revenue of Descending-Price Auction

Bayes-Nash equilibria of first-price auctions:

- are efficient when bidders are symmetric [Myerson, 1981];
- achieve $\geq 1 - \frac{1}{e} \cong 0.63\ldots$ fraction of best possible welfare in general. [Syrgkanis, 2012]

Our descending-price auction inherits the same welfare guarantees.

Descending-Price Auction for Multiple Items

Descending clock represents uniform price for all items.

Bidders may claim any remaining item at the current price.

Theorem

Every equilibrium of the descending-price auction achieves at least one-third of the first-best welfare.

Remarks:

- First-best policy not known to be computationally efficient.
- Best known polynomial-time algorithm is a 2-approximation, presented earlier in this lecture.

Descending-Price Auction for Multiple Items

Descending clock represents uniform price for all items.

Bidders may claim any remaining item at the current price.

Theorem

Every equilibrium of the descending-price auction achieves at least one-third of the first-best welfare.

Proof sketch: via the *smoothness framework* [Lucier-Borodin '10, Roughgarden '12, Syrgkanis '12, Syrgkanis-Tardos '13].

Descending-Price Auction for Multiple Items

Descending clock represents uniform price for all items.

Bidders may claim any remaining item at the current price.

Theorem

Every equilibrium of the descending-price auction achieves at least one-third of the first-best welfare.

Proof sketch: via the *smoothness framework*.

For bidder i, consider "deviation" that inspects each j when price is at $\frac{2}{3}\sigma_{ij}$ and buys at $\frac{2}{3}\kappa_{ij}$. (Note this is non-exposed.)

One of three alternatives must hold:

- In equilibrium, the price of j is at least $\frac{2}{3}\kappa_{ij}$.
- In equilibrium, i pays at least $\frac{2}{3}\kappa_{ij}$.
- In deviation, expected utility of i is at least $\frac{1}{3}\kappa_{ij}$.

$$\frac{1}{2}p^j + \frac{1}{2}p_i + u_i \geq \frac{1}{3}\kappa_{ij}$$

Descending-Price Auction for Multiple Items

Descending clock represents uniform price for all items.

Bidders may claim any remaining item at the current price.

Theorem

Every equilibrium of the descending-price auction achieves at least one-third of the first-best welfare.

$$\mathbb{E}[\text{welfare of descending price}] = \mathbb{E}\left[\sum_i (u_i + p_i)\right]$$

$$= \mathbb{E}\left[\sum_i u_i + \tfrac{1}{2}\sum_i p_i + \tfrac{1}{2}\sum_j p^j\right]$$

$$\geq \tfrac{1}{3}\,\mathbb{E}\left[\max_{\mathcal{M}} \sum_{(i,j)\in\mathcal{M}} \kappa_{ij}\right] \geq \tfrac{1}{3}\,\text{OPT}$$

where $\mathcal{M}$ ranges over all matchings.

Part 4:

Social Learning

Crowdsourced investigation "in the wild"

amazon

Flipboard

GALAXY ZOO

Crowdsourced investigation "in the wild"

Decentralized exploration suffers from misaligned incentives.

- Platform's goal: Collect data about many alternatives.
- User's goal: Select the best alternative.

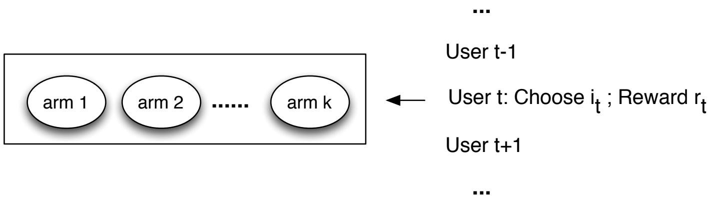
Crowdsourced investigation “in the wild”

Decentralized exploration suffers from misaligned incentives.

- Platform’s goal: **EXPLORE.**
- User’s goal: **EXPLOIT.**

A Model Based on Multi-Armed Bandits

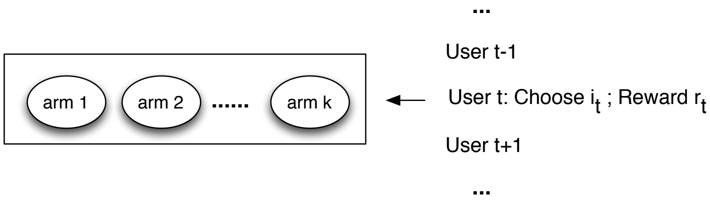
k arms have independent random types that govern their (time-invariant) reward distribution when selected.



Users observe all past rewards before making their selection.

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.

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

User t's goal: maximize r_t

Incentivized Exploration

Incentive payments

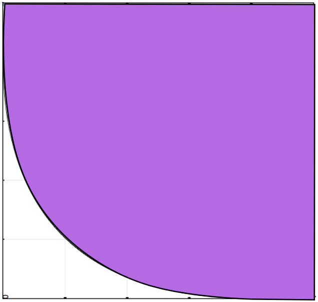
At time t, announce reward $c_{t,i} \geq 0$ for each arm i.
User now chooses i to maximize $\mathbb{E}[r_{i,t}] + c_{i,t}$.

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 π.
- Cost of a policy pulling arm i at time t is $r_t^{\max} - r_{i,t}$, where $r_t^{\max}$ denotes myopically optimal reward.

The Achievable Region



Suppose, for platform's policy π:

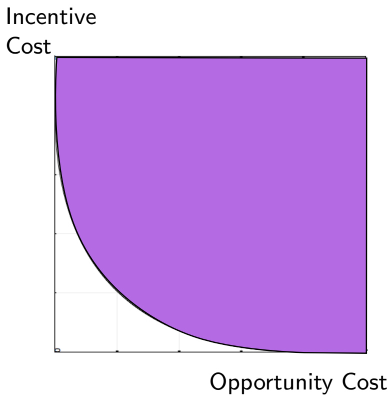
- reward $\geq (1 - a) \cdot \text{OPT}$.
- payment $\leq b \cdot \text{OPT}$.

We say π achieves loss pair (a, b).

Definition

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

The Achievable Region



Suppose, for platform's policy π:

- reward $\geq (1 - a) \cdot \mathrm{OPT}$.
- payment $\leq b \cdot \mathrm{OPT}$.

We say π achieves loss pair (a, b).

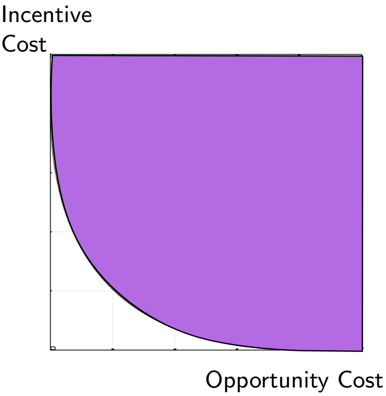
Definition

(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 $\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}$.

The Achievable Region

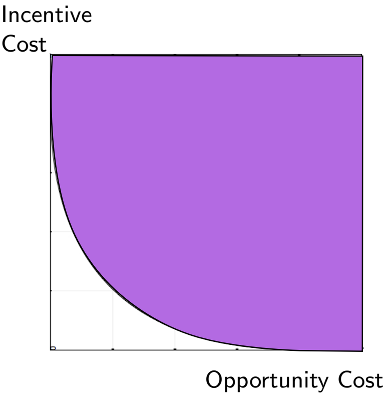


- Achievable region is convex, closed, upward monotone.

Main Theorem

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

The Achievable Region

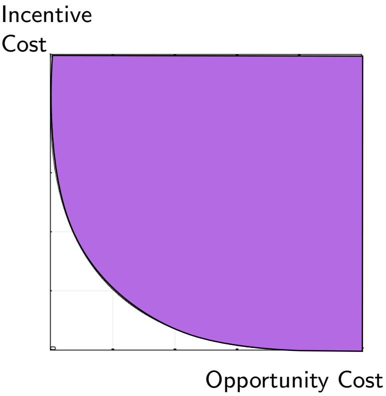


- Achievable region is convex, closed, upward monotone.
- Set-wise increasing in δ.

Main Theorem

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

The Achievable Region

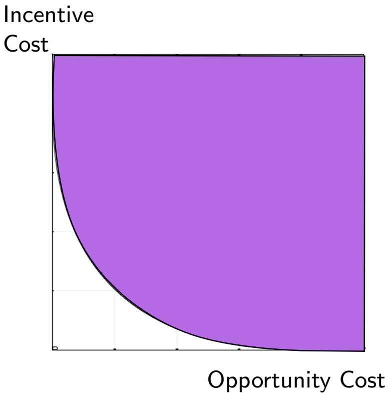


- Achievable region is convex, closed, upward monotone.
- Set-wise increasing in δ.
- (0.25,0.25) and (0.1,0.5) achievable for all δ.

Main Theorem

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

The Achievable Region



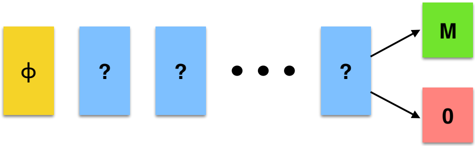
- Achievable region is convex, closed, upward monotone.
- Set-wise increasing in δ.
- (0.25,0.25) and (0.1,0.5) achievable for all δ.

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 $\sqrt{a} + \sqrt{b} \geq \sqrt{1-\delta}$.

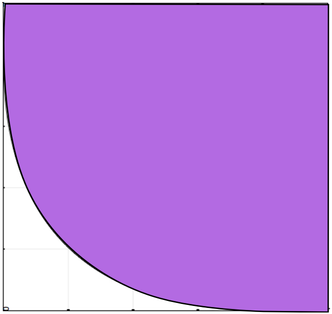
Diamonds in the Rough



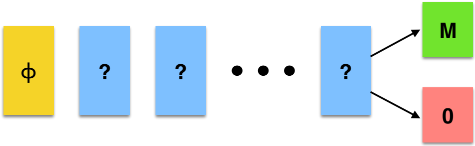
A Hard Instance

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

(Type fully revealed when pulled.)



Diamonds in the Rough



A Hard Instance

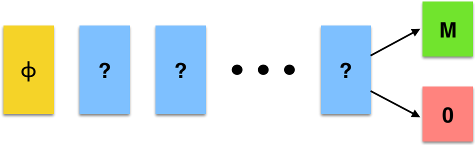
Infinitely many "collapsing" arms
M with prob. $\frac{1}{M}\delta^2$, else 0.

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.

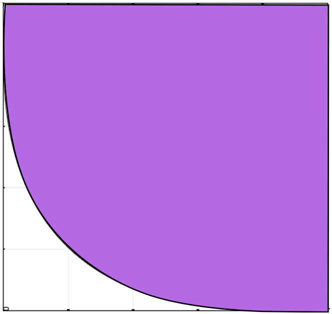
Diamonds in the Rough



A Hard Instance

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

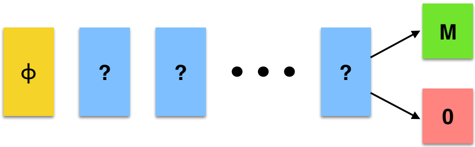
One arm whose payoff is always $\phi \cdot \delta$.



Extreme points of achievable region correspond to:

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

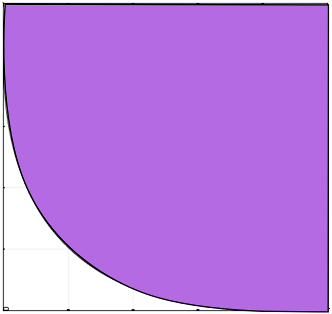
Diamonds in the Rough



A Hard Instance

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

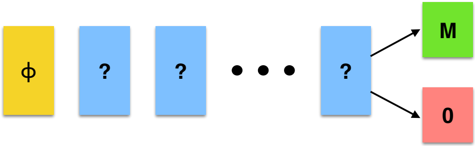
One arm whose payoff is always $\phi \cdot \delta$.



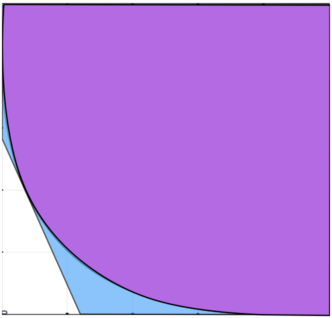
Extreme points of achievable region correspond to:

- OPT: reward ≈ 1, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- MYO: reward ϕ, cost 0. $(a, b) = (1 - \phi, 0)$

Diamonds in the Rough



A Hard Instance



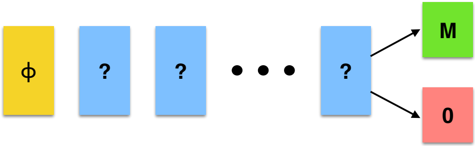
Infinitely many "collapsing" arms
M with prob. $\frac{1}{M}\delta^2$, else 0.

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

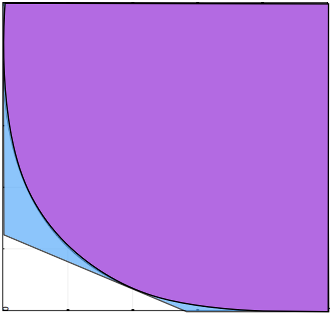
Extreme points of achievable region correspond to:

- OPT: reward ≈ 1, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- MYO: reward ϕ, cost 0. $(a, b) = (1 - \phi, 0)$

Diamonds in the Rough



A Hard Instance



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

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

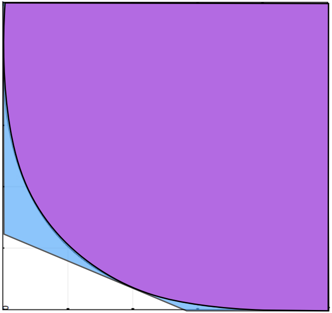
Extreme points of achievable region correspond to:

- OPT: reward ≈ 1, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- MYO: reward ϕ, 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

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

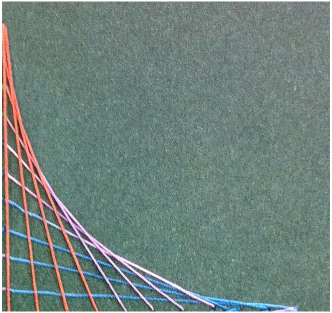


- OPT: reward ≈ 1, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- MYO: reward ϕ, cost 0. $(a, b) = (1 - \phi, 0)$

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 ≈ 1, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- MYO: reward ϕ, cost 0. $(a, b) = (1 - \phi, 0)$

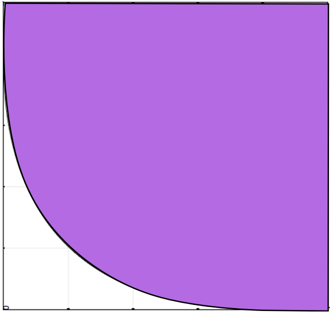
Diamonds in the Rough

The inequality

$$\sqrt{x} + \sqrt{y} \geq \sqrt{1 - \delta}$$

holds if and only if

$$\forall \phi \in (\delta, 1) \quad x + \left(\frac{1 - \phi}{\phi - \delta}\right) y \geq 1 - \phi$$



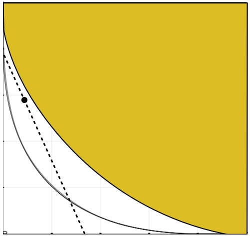
- OPT: reward ≈ 1, cost $\approx \phi - \delta$. $(a, b) = (0, \phi - \delta)$
- MYO: reward ϕ, cost 0. $(a, b) = (1 - \phi, 0)$

Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}$.

Then there is a line through (a, b) outside the achievable region.

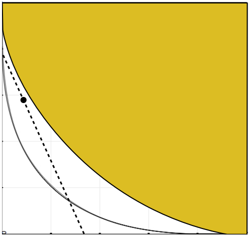


Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}$.

Then there is a line through (a, b) outside the achievable region.



For all achievable x, y,

$$(1 - p)x + py > (1 - p)a + pb$$

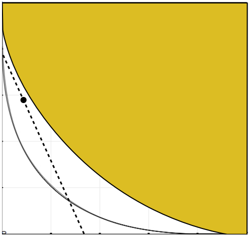
.

Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1 - \delta}$.

Then there is a line through (a, b) outside the achievable region.



For all achievable x, y,

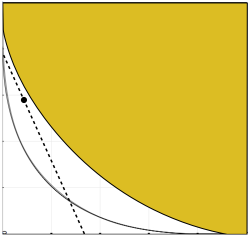
$$x + \left(\frac{p}{1-p}\right) y > a + \left(\frac{p}{1-p}\right) b$$

Lagrangean Relaxation

Proof of achievability is by contradiction.

$$\text{Suppose } (a, b) \text{ unachievable and } \sqrt{a} + \sqrt{b} \geq \sqrt{1-\delta}.$$

Then there is a line through (a, b) outside the achievable region.



For all achievable x, y,

$$x + \left(\frac{p}{1-p}\right) y > a + \left(\frac{p}{1-p}\right) b$$

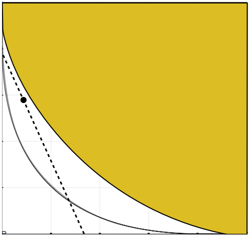
Let $\phi = 1 - (1-\delta)p$, so $p = \frac{1-\phi}{1-\delta}$, $1 - p = \frac{\phi-\delta}{1-\delta}$.

Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1-\delta}$.

Then there is a line through (a, b) outside the achievable region.



For all achievable x, y,

$$x + \left(\frac{1-\phi}{\phi-\delta}\right) y > a + \left(\frac{1-\phi}{\phi-\delta}\right) b$$

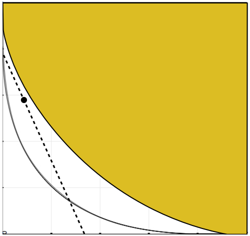
Let $\phi = 1 - (1-\delta)p$, so $p = \frac{1-\phi}{1-\delta}$, $1 - p = \frac{\phi-\delta}{1-\delta}$.

Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1-\delta}$.

Then there is a line through (a, b) outside the achievable region.



For all achievable x, y,

$$x + \left(\frac{1-\phi}{\phi-\delta}\right) y > 1 - \phi$$

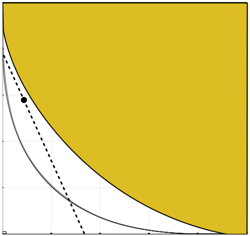
Let $\phi = 1 - (1-\delta)p$, so $p = \frac{1-\phi}{1-\delta}$, $1 - p = \frac{\phi-\delta}{1-\delta}$.

Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1-\delta}$.

Then there is a line through (a, b) outside the achievable region.



For all achievable x, y,

$$(1-x) - \left(\frac{1-\phi}{\phi-\delta}\right) y < \phi$$

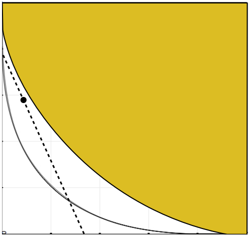
Let $\phi = 1 - (1-\delta)p$, so $p = \frac{1-\phi}{1-\delta}$, $1 - p = \frac{\phi-\delta}{1-\delta}$.

Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1-\delta}$.

Then there is a line through (a, b) outside the achievable region.



For all achievable x, y,

$$(1-x) - \left(\frac{p}{1-p}\right) y < \phi$$

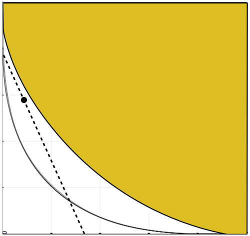
Let $\phi = 1 - (1-\delta)p$, so $p = \frac{1-\phi}{1-\delta}$, $1 - p = \frac{\phi-\delta}{1-\delta}$.

Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1-\delta}$.

Then there is a line through (a, b) outside the achievable region.



For all achievable x, y,

$$(1-x) - \left(\frac{p}{1-p}\right) y < \phi$$

LHS $= \mathbb{E}[\text{Payoff}(\pi) - \frac{p}{1-p}\text{Cost}(\pi)]$, if π achieves loss pair (x, y).

Lagrangean Relaxation

Proof of achievability is by contradiction.

Suppose (a, b) unachievable and $\sqrt{a} + \sqrt{b} \geq \sqrt{1-\delta}$.

To reach a contradiction, must show that for all $0 < p < 1$, if $\phi = 1 - (1-\delta)p$, there exists policy π such that

$$\mathbb{E}[\text{Payoff}(\pi) - \tfrac{p}{1-p}\text{Cost}(\pi)] \geq \phi.$$

For all achievable x, y,

$$(1-x) - \left(\frac{p}{1-p}\right) y < \phi$$

LHS $= \mathbb{E}[\text{Payoff}(\pi) - \frac{p}{1-p}\text{Cost}(\pi)]$, if π achieves loss pair (x, y).

Time-Expanded Policy

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

The difficulty is $\mathsf{Cost}(\pi)$. 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.

Time-Expanded Policy

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

Example. At time 0, suppose myopically optimal arm i has reward r_i and OPT wants arm j with reward $r_j < r_i$.

Pull i with probability p, j with probability $1 - p$.

$$\mathbb{E}[\text{Reward} - \tfrac{p}{1-p}\text{Cost}] = pr_i + (1-p)[r_j - \tfrac{p}{1-p}(r_i - r_j)] = r_j$$

Time-Expanded Policy

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

Example. At time 0, suppose myopically optimal arm i has reward r_i and OPT wants arm j with reward $r_j < r_i$.

Pull i with probability p, j with probability $1 - p$.

$\mathbb{E}[\text{Reward} - \frac{p}{1-p}\text{Cost}] = pr_i + (1-p)[r_j - \frac{p}{1-p}(r_i - r_j)] = r_j$

Keep going like this?
Hard to analyze OPT with unplanned state changes.
Instead, treat unplanned state changes as "no-ops".

Time-Expanded Policy

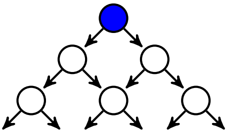
The time-expansion of policy π with parameter p; TE(π, p)

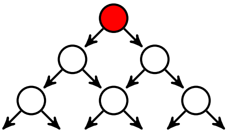
Maintain a FIFO queue of states for each arm, tail is current state.

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 π 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

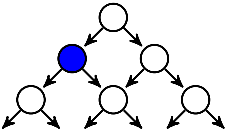
The time-expansion of policy π with parameter p; TE(π, p)

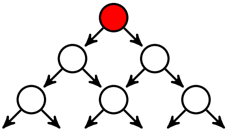
Maintain a FIFO queue of states for each arm, tail is current state.

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 π 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 π with parameter p; TE(π, p)

Maintain a FIFO queue of states for each arm, tail is current state.

At each time t, toss a coin with bias p.

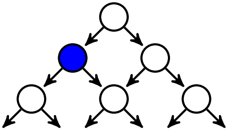
Heads: Offer no incentive payments.

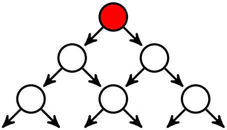
User plays myopically. Push new state into tail of queue.

Tails: Apply π 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 π with parameter p; TE(π, p)

Maintain a FIFO queue of states for each arm, tail is current state.

At each time t, toss a coin with bias p.

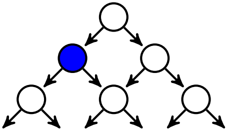
Heads: Offer no incentive payments.

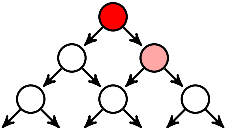
User plays myopically. Push new state into tail of queue.

Tails: Apply π 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.





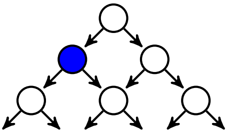
Time-Expanded Policy

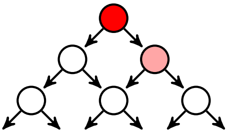
The time-expansion of policy π with parameter p; TE(π, p)

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 π 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 π with parameter p; TE(π, p)

Maintain a FIFO queue of states for each arm, tail is current state.

At each time t, toss a coin with bias p.

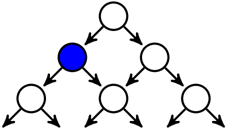
Heads: Offer no incentive payments.

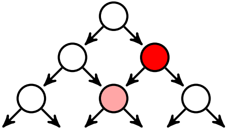
User plays myopically. Push new state into tail of queue.

Tails: Apply π 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 π with parameter p; TE(π, 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 π 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.

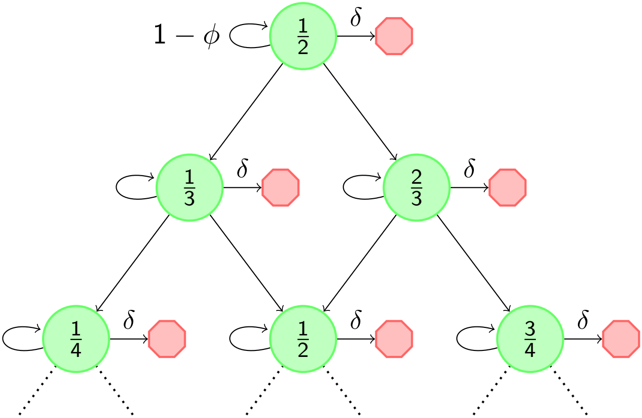
Lagrangean payoff analysis. In a state where MYO would pick i and π would pick j, expected Lagrangean payoff is

$$p r_{i,t} + (1-p)\left[r_{j,t} - \left(\frac{p}{1-p}\right)(r_{i,t} - r_{j,t}) \right] = r_{j,t}.$$

If s is at the head of j's queue at time t, then $\mathbb{E}[r_{j,t} | s] = R_j(s)$.

Stuttering Arms

The "no-op" steps modify the Markov chain to have self-loops in every state with transition probability $(1-\delta)p = 1-\phi$.



Gittins Index of Stuttering Arms

Lemma

Letting $\tilde{\sigma}(s)$ *denote the Gittins index of state* s *in the modified Markov chain, we have* $\tilde{\sigma}(s) \geq \phi \cdot \sigma(s)$ *for every* s*.*

Gittins Index of Stuttering Arms

Lemma

Letting $\tilde{\sigma}(s)$ denote 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 π 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

$$\mathbb{E}[\text{Payoff} - \tfrac{p}{1-p}\text{Cost}] \geq \phi.$$

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

Gittins Index of Stuttering Arms

Lemma

Letting $\tilde{\sigma}(s)$ denote 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 τ such that

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

Let τ' be the equivalent stopping rule for $\tilde{\mathcal{M}}$, i.e. τ' simulates τ on the subset of time steps that are not self-loops.

Gittins Index of Stuttering Arms

Lemma

Letting $\tilde{\sigma}(s)$ denote the Gittins index of state s in the modified Markov chain, we have $\tilde{\sigma}(s) \geq \phi \cdot \sigma(s)$ for every s.

The proof will show

$$\begin{aligned}
\mathbb{E}\left[\sum_{0<t<\tau'} R(\tilde{s}_t)\right] &\geq \mathbb{E}\left[\sum_{0<t<\tau} R(s_t)\right] \\
&\geq \sigma(s) \cdot \Pr(s_\tau \in \mathcal{T}) \\
&\geq \phi \cdot \sigma(s) \cdot \Pr(\tilde{s}_{\tau'} \in \mathcal{T}) > 0.
\end{aligned}$$

By definition of Gittins index, this means $\tilde{\sigma}(s) \geq \phi \cdot \sigma(s)$.

Second line holds by assumption. Prove first, third by coupling.

Gittins Index of Stuttering Arms

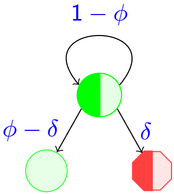
$$\mathbb{E}\left[\sum_{0<t<\tau'} R(\tilde{s}_t)\right] \geq \mathbb{E}\left[\sum_{0<t<\tau} R(s_t)\right]$$

$$\Pr(s_\tau \in \mathcal{T}) \geq \phi \cdot \Pr(\tilde{s}_{\tau'} \in \mathcal{T})$$

Gittins Index of Stuttering Arms

$\mathbb{E}\left[\sum_{0<t<\tau'} R(\tilde{s}_t)\right] \geq \mathbb{E}\left[\sum_{0<t<\tau} R(s_t)\right]$

$\Pr(s_\tau \in \mathcal{T}) \geq \phi \cdot \Pr(\tilde{s}_{\tau'} \in \mathcal{T})$



For $t \in \mathbb{N}$ sample color green vs. red with probability $1-\delta$ vs. δ.
Independently, sample light vs. dark with probability $1-p$ vs. p.

State transitions of $\tilde{\mathcal{M}}$ are:

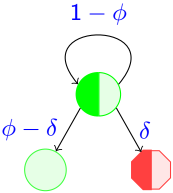
- terminal on red
- self-loop on dark green
- non-terminal $\mathcal{M}$-step on light green.

The light time-steps simulate $\mathcal{M}$.
Let f = monotonic bijection from $\mathbb{N}$ to light time-steps.

Gittins Index of Stuttering Arms

$$\mathbb{E}\left[\sum_{0<t<\tau'} R(\tilde{s}_t)\right] \geq \mathbb{E}\left[\sum_{0<t<\tau} R(s_t)\right]$$

$$\Pr(s_\tau \in \mathcal{T}) \geq \phi \cdot \Pr(\tilde{s}_{\tau'} \in \mathcal{T})$$



At any light green time,

$$\Pr(\text{light red before next light green}) = \delta$$

$$\Pr(\text{red before next light green}) = \delta/\phi.$$

So for all m, conditioned on $\mathcal{M}$ running m steps without terminating,

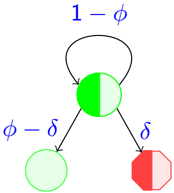
$$\Pr(\tilde{\mathcal{M}} \text{ enters terminal state between } f(m) \text{ and } f(m+1))$$
$$= \phi \cdot \Pr(\mathcal{M} \text{ enters terminal state between } m \text{ and } m+1)$$

implying $\Pr(s_\tau \in \mathcal{T}) \geq \phi \cdot \Pr(\tilde{s}_{\tau'} \in \mathcal{T})$.

Gittins Index of Stuttering Arms

$$\mathbb{E}\left[\sum_{0<t<\tau'} R(\tilde{s}_t)\right] \geq \mathbb{E}\left[\sum_{0<t<\tau} R(s_t)\right]$$

$$\Pr(s_\tau \in \mathcal{T}) \geq \phi \cdot \Pr(\tilde{s}_{\tau'} \in \mathcal{T})$$



Let $t_1 =$ first red step, $t_2 =$ first light red step

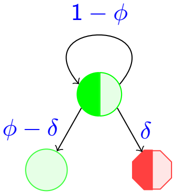
$t_3 =$ first green step when τ' stops

Then $\tau = \min\{t_2, t_3\}$, $f(\tau') = \min\{t_1, t_3\}$.

Gittins Index of Stuttering Arms

$\mathbb{E}\left[\sum_{0<t<\tau'} R(\tilde{s}_t)\right] \geq \mathbb{E}\left[\sum_{0<t<\tau} R(s_t)\right]$

$\Pr(s_\tau \in \mathcal{T}) \geq \phi \cdot \Pr(\tilde{s}_{\tau'} \in \mathcal{T})$



To prove: $\mathbb{E}[\sum_{0<t<\tau'} R(\tilde{s}_t)] \geq \mathbb{E}[\sum_{0<t<\tau} R(s_t)]$

$$\sum_{0<t<\tau'} R(\tilde{s}_t) = \sum_{0<t<t_1} R(\tilde{s}_t) - \sum_{t_3 \leq t < t_1} R(\tilde{s}_t)$$

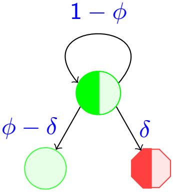
$$\sum_{0<t<\tau} R(s_t) = \sum_{0<f(t)<t_2} R(\tilde{s}_{f(t)}) - \sum_{t_3 \leq f(t) < t_2} R(\tilde{s}_{f(t)})$$

First terms on RHS have same expectation, $R(\tilde{s}_1) \cdot \delta^{-1}$.
Compare second terms by case analysis on ordering of t_1, t_2, t_3.

Gittins Index of Stuttering Arms

$\mathbb{E}\left[\sum_{0<t<\tau'} R(\tilde{s}_t)\right] \geq \mathbb{E}\left[\sum_{0<t<\tau} R(s_t)\right]$

$\Pr(s_\tau \in \mathcal{T}) \geq \phi \cdot \Pr(\tilde{s}_{\tau'} \in \mathcal{T})$



To prove: $\mathbb{E}\left[\sum_{t_3 \leq t \leq t_1} R(\tilde{s}_t)\right] \leq \mathbb{E}\left[\sum_{t_3 \leq f(t) \leq t_2} R(\tilde{s}_{f(t)})\right]$

1. $t_1 \leq t_2 < t_3$: Both sides are zero.
2. $t_1 < t_3 < t_2$: Left side is zero, right side is non-negative.
3. $t_3 < t_1 \leq t_2$: Conditioned on $s = s_{t_3}$, both sides have expectation $R(s) \cdot \delta^{-1}$.

Conclusion

- Joint Markov scheduling: versatile model of information acquisition in Bayesian settings.
  - . . . when alternatives ("arms") are strategic
  - . . . when time steps are strategic.
- First-best policy: Gittins index policy.
- Analysis tool: *deferred value* and *amortization lemma*.
  - Akin to virtual values in optimal mechanism design . . .
  - 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

Algorithmic.

- Correlated arms (cf. ongoing work of Anupam Gupta, Ziv Scully, Sahil Singla)
- More than one way to inspect an alternative (i.e., arms are MDPs rather than Markov chains; cf. [Glazebrook, 1979; Cavallo & Parkes, 2008])
- Bayesian contextual bandits
- Computational hardness of any of the above?

Open questions

Algorithmic.

- Correlated arms (cf. ongoing work of Anupam Gupta, Ziv Scully, Sahil Singla)
- More than one way to inspect an alternative (i.e., arms are MDPs rather than Markov chains; cf. [Glazebrook, 1979; Cavallo & Parkes, 2008])
- Bayesian contextual bandits
- Computational hardness of any of the above?

Game-theoretic.

- Strategic arms ("exploration in markets")
  - Revenue guarantees (cf. [K.-Waggoner-Weyl, 2016])
  - Two-sided markets (patent applic. by K.-Weyl, no theory yet!)
- Strategic time steps ("incentivizing exploration")
  - Agents who persist over time.