Searching for Arms

Daniel Fershtman Alessandro Pavan

Stony Brook Game Theory Festival
July 2019

Motivation

- Experimentation/Sequential Learning
  - central to many problems
- In many cases,
  - *endogenous* set of alternatives/arms
  - *search*
- Tradeoff: **exploring** existing alternatives vs **searching** for new ones

Motivation

Example

- Consumer sequentially explores different alternatives within "consideration set", while expanding consideration set through search
- Firm interviews candidates, while searching for additional suitable candidates to interview
- Researcher splits time on several ongoing projects of unknown return, while also searching for new projects

- Key difference
  - experimentation: directed
  - search: **undirected, stochastic**

This Paper

- Multi-armed bandit problem with **endogenous** set of arms
  - Each period, DM pulls one of existing arms or searches for new ones
  - Search brings (stochastically) new arms of different types
  - Search technology (cost + distribution over new arms) function of past search outcomes
- Show problem is decomposable:
  - optimal policy: **index policy (with special index for search)**
- Extension to problems with **irreversible choice** (based on partial information)
  - Weitzman: special case where set of boxes exogenous and uncertainty resolved after first inspection

Difficulties

- Difficulty 1: opportunity cost of search depends on entire composition of current choice set
  - e.g., profitability of searching for additional candidates depends on observable covariates of current candidates (gender, education, etc.) and past interviews
- Difficulty 2: Search outcome may depend on
  - type and number of arms previously found
  - past search costs
- Difficulty 3: Search competes with its own "descendants" (i.e., with arms discovered through past searches)
  - correlation
- Difficulty 4: treating search as "meta arm" requires decisions within meta arm invariant to info outside meta arm
  - bandit problems with meta arms (e.g., arms that can be activated with different intensities – "super-processes") rarely admit index solution

Literature

- Bandits
  - Gittins and Jones (1974), Rothschild (1974), Rustichini and Wolinsky (1995), Keller and Rady (1999)...
- Bandits with time-varying set of alternatives
  - Whittle (1981), Varaiya, Walrand and Buyukkoc (1985), Weiss (1988), Weber (1994)...
- Sequential search for best alternative (Pandora's problem)
  - Weitzman (1979), Olszewski and Weber (2015), Choi and Smith (2016), Doval (2018)...
- Experimentation before irreversible choice
  - Ke, Shen and Villas-Boas (2016), Ke and Villas-Boas (2018)...

Plan

1. Model
2. Optimal policy
3. Dynamics
4. Proof of main theorem
5. Applications
6. Extensions
    - irreversible choice
    - search frictions
    - multiple search arms
    - no discounting

Model

Model: Environment

- Discrete time: $t = 0, ..., \infty$
- Available "physical" arms in period t: $I_t = \{1, ..., n_t\}$
  (I_0 exogenous)
- At each t, DM
  - pulls arm among I_t
  - **searches** for new arms
  - opts-out: arm $i = 0$ (fixed reward equal to outside option)
- Pulling arm $i \in I_t$
  - stochastic reward $r_i \in \mathbb{R}$
  - transition to new state
- Arms not pulled:
  - zero reward
  - frozen state
- Search
  - costly
  - stochastic set of new arms $I_{t+1} \setminus I_t$

Model: "Physical" Arms

- "State" of physical arm: $\omega^P = (\xi, \theta) \in \Omega^P$:
  - $\xi \in \Xi$: persistent "type"
  - $\theta \in \Theta$: evolving state
- Example:
  - ξ: type of research project/idea (theory, empirical, experimental)
  - $\theta = (\sigma^m)$: history of signals about project's impact
  - r: utility from working on project
- H_{ω^P}: distribution over Ω^P, given ω^P
- Reward: $r(\omega^P)$
- Key assumptions:
  - Arm' state "frozen" when arm not pulled
  - θ drawn independently of calendar time t and of other arms' state, **conditional on arms' types**

Model: Search Technology

- State of *search technology:* $\omega^S = ((c_0, E_0), (c_1, E_1), ..., (c_m, E_m)) \in \Omega^S$
  - m: number of past searches
  - $(c_0, ..., c_m)$: history of past search costs
  - $(E_0, ..., E_m)$: history of types of arms found
    - c_k: cost of k'th search
    - $E_k = (n_k(\xi) : \xi \in \Xi)$, where $n_k(\xi) \in \mathbb{N}$ is number of arms of type ξ found at k'th search
- H_{ω^S}: joint distribution over (c, E), given ω^S
- Search technology **independent of calendar time and of arms' idiosyncratic shocks,** θ
- **Correlation though** ξ

Model: Search Technology

- Search technology:
  - learning about alternatives not yet in consideration set
  - stochastic history-dependence
    - ability/cost of finding new alternatives deteriorating/improving with time
    - possibly limited set of outside alternatives

Model: states and policies

- Overall state: $\mathcal{S} \equiv (\omega^S, \mathcal{S}^P)$
  - $\mathcal{S}^P(\omega^P)$: number of physical arms in state $\omega^P \in \Omega^P$
  - $\mathcal{S}^P \equiv (\mathcal{S}(\omega^P) : \omega^P \in \Omega^P)$ state of physical arms
- Period-t state: $\mathcal{S}_t \equiv (\omega_t^S, \mathcal{S}_t^P)$
- Definition eliminates dependence on calendar time, while keeping track of relevant information
- *Policy* χ describes feasible decisions at all histories
- Policy χ *optimal* if it maximizes expected discounted sum of net payoffs

$$\mathbb{E}^{\chi}\left[\sum_{t=0}^{\infty}\delta^t\left(\sum_{j=1}^{\infty}x_{jt}r_{jt}-c_t y_t\right)|\mathcal{S}_0\right]$$

Plan

1. Model
2. **Optimal policy**
3. Dynamics
4. Proof of main theorem
5. Extensions
   - Search frictions
   - Irreversible choice
6. Applications

Optimal Policy

Indexes for Physical Arms

- Index for "physical" arms:

$$\mathcal{G}^P(\omega^P) \equiv \sup_{\tau>0} \frac{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s r_s | \omega^P\right]}{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s | \omega^P\right]}$$

  where τ is stopping time

- Interpretations:
  - maximal expected discounted reward, per unit of expected discounted time (Gittins)
  - annuity that makes DM indifferent between stopping right away and continuing with option to retire in the future (Whittle)
  - fair charge (Weber)

Index for Search

- Index for search:

$$\mathcal{G}^S(\omega^S) \equiv \sup_{\pi,\tau} \frac{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s (r_s^\pi - c_s^\pi) | \omega^S\right]}{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s | \omega^s\right]}$$

  - τ : stopping time
  - π: rule prescribing choice among arms discovered AFTER t and FUTURE searches
  - r_s^π, c_s^π: stochastic rewards/costs, under rule π

- Interpretation: fair (flow) price for visiting "casinos" found stochastically over time, playing in them, and continue searching for other casinos
- Definition:
  - accommodates for correlation among arms found over time
  - compatible with possibility that search lasts indefinitely and brings unbounded set of alternatives

Optimality of index policy

Definition

Index policy selects at each t "search" iff

$$\mathcal{G}^S(\omega_t^S) \geq \underbrace{\mathcal{G}^*(\mathcal{S}_t^P)}_{\substack{\text{maximal index among} \\ \text{available physical arms}}}$$

otherwise, it selects any "physical" arm with index $\mathcal{G}^*(\mathcal{S}_t^P)$

Theorem 1

Index policy optimal in bandit problem with search for new arms

Implications of Index Policy

- Suppose you want to hire new worker
- Each worker can be ξ =Male of ξ=Female (different processes over signals/rewards)
- Probability search brings Male: .8
- Optimality of searching for new candidates same no matter whether you have 49 M and 1 F, or 25 M and 25 F
- However, value of continuing with current set depends on number of M and F (and past interviews)
- **Maximal index among current arms NOT sufficient statistics for state of current arms when it comes to continuation payoff with current arms.**

Plan

1. Model
2. Optimal policy
3. Dynamics
4. Proof of main theorem
5. Extensions
   - Search frictions
   - Irreversible choice
6. Applications

Dynamics

Dynamics under index policy

- Stationary search technology: $H_{\omega^S} = H^S$ all ω^S
  - given $\mathcal{G}^*(\mathcal{S}_t^P)$, composition of $\mathcal{S}^P \equiv (\mathcal{S}(\omega^P) : \omega^P \in \Omega^P)$ irrelevant for decision to search
  - all physical arms present at t never pulled again (**search=replacement**)
  - Result extends to "Improving search technologies":
    - physical arms required to pass more stringent tests over time
- Deteriorating search technology:
  - existing arms put on hold
  - DM may return to arms present before last search

Plan

1. Model
2. Optimal policy
3. Dynamics
4. Proof of main theorem
5. Extensions
   - Search frictions
   - Irreversible choice
6. Applications

Proof of Main Theorem

Proof of Theorem 1: Road Map

1. Characterization of payoff under index policy
   - representation uses particular "**timing process**" based on **optimal stopping** in indexes definition:
     - physical arms: stop when index drops below its initial value (Mandelbaum, 1986)
     - search: stop when search index and all indexes of **new** arms are smaller than value of search index when search began

2. Dynamic programming
   - payoff function under index policy solves dynamic programming equation

Proof: Step 1

- $(\mathcal{S}_t)_{t\geq 0}$: **state process** under index policy
- $\kappa(v|\mathcal{S}) \in \mathbb{N} \cup \{\infty\}$: minimal time until *all* indexes (search/existing arms/newly found arms) weakly below $v \in \mathbb{R}_+$

  ($\kappa(v|\mathcal{S}) = \infty$ if event never occurs)

Lemma 1

$$\underbrace{\mathcal{V}(\mathcal{S}_0)}_{\substack{\text{payoff under index policy, starting from state } \mathcal{S}_0}} = \int_0^\infty [1 - \underbrace{\mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)}}_{\substack{\text{expected discounted time till all indexes drop weakly below } v}}] dv$$

Proof: Step 2

- Show that $\mathcal{V}(\mathcal{S}_0)$ solves **dynamic programming equation**:

$$\mathcal{V}(\mathcal{S}_0) = \max\{ \underbrace{V^S(\omega^S|\mathcal{S}_0)}_{\substack{\text{value from searching} \\ \text{and reverting} \\ \text{to index} \\ \text{policy thereafter}}}, \underbrace{\max_{\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0^P(\hat{\omega}^P) > 0\}} V^P(\omega^P|\mathcal{S}_0)}_{\substack{\text{value from pulling} \\ \text{physical arm and} \\ \text{reverting to index} \\ \text{policy thereafter}}} \}$$

- Proof uses
  - representation of payoff under index policy from Lemma 1
  - decomposition of overall problem into collection of binary problems where choice is between single arm (possibly search) and auxiliary fictitious arm with fixed reward

Plan

1. Model
2. Index policy
3. Dynamics
4. Proof of main theorem
5. Applications
6. Extensions
    - irreversible choice
    - search frictions
    - multiple search arms
    - no discounting

Applications

Dynamic Matching on a Platform

- Fershtman and Pavan (2017): platform dynamically matches agents
  - Shocks to match quality
  - Gradual learning about attractiveness
- Set of agents in Fershtman and Pavan (2017) exogenously fixed
- Many markets:
  - platforms solicit buyers/sellers in response to past outcomes/bids
- Above results allow to study joint dynamics of
  - bidding
  - matching
  - solicitation
- Interesting distortions in solicitation dynamics (due to mkt power + private info)
  - initial phase with excessive solicitation
  - subsequent phase with insufficient solicitation

Design of Search Engines

- Representative buyer uses search engine to identify product to purchase
- Search brings set of sponsored and organic links
- Clicking on a link brings additional information
- GSP auction
  - sellers compete by submitting bids
  - higher bids: higher positions
  - payments linked to clicks
- Theory permits to
  - endogenize click through rates (CTR)
  - characterize firms' value for being on different positions/pages
  - auction design
    - how many products per page?
    - payments

Selling an Asset (e.g., your house)

- Arms: buyers
- ξ: offer's bid (endogenously controlled by buyers)
- $\theta = (s^m)$: history of signals about cost of serving the buyer
- Buyers arrive to mkt according to Poisson
- Seller's search increases Poisson's rate
- Theory sheds light on
  - eq. bidding
  - solicitation process

Plan

1. Model
2. Index policy
3. Dynamics
4. Proof of main theorem
5. Applications
6. Extensions
    - irreversible choice
    - search frictions
    - multiple search arms
    - no discounting

Extensions

Extension 1: Irreversible Choice

- Consider following extension with **irreversible choice**
- In each period, DM can
  - search for new alternatives
  - experiment with existing ones
  - **irreversibly select one alternative from those found from past searches**
- Arm of type ξ must be pulled at least $M_\xi \geq 0$ times before DM can irreversibly commit to it (Weitzman: $M_\xi = 1$ all ξ)
- Flow-payoff from irreversibly selecting arm in state ω^P: $R(\omega^P)$

Extension 1: Irreversible Choice

- $\hat{\omega}^P \succeq \omega^P$ iff $\omega^P = (\xi, \sigma, m)$ and $\hat{\omega}^P = (\xi, \sigma, \hat{m})$ with $\hat{m} \geq m$
  - that is, $\theta = (\sigma, m)$ contains information on number of times "m" arm has been activated

Definition

Arm of type ξ satisfies **"better-later-than-sooner"** property if, for any $\omega^P = (\xi, \sigma, m)$, with $m \geq M_\xi$, for any $\hat{\omega}^P \succeq \omega^P$, either $R(\hat{\omega}^P) \geq R(\omega^P)$ or $R(\omega^P), R(\hat{\omega}^P) \leq 0$.

- Remark: Weitzman special case

Theorem

Suppose all arm types satisfy "better-later-than-sooner" property. Then index policy optimal.

Extension 1: Irreversible Choice

- **Fictitious environment** with no irreversible choice:
  - for any physical arm in state ω^P found though search, or pulled in period t, "auxiliary" arm generated at t with fixed reward $R(\omega^P)$ also "found"
  - Auxiliary arms remain in same state forever and do not generate other auxiliary arms
  - Pulling auxiliary arm corresponding to arm j equivalent to choosing arm j (once pulled, it will be pulled forever)
- Given state ω^P, index of each physical arm

$$\hat{\mathcal{G}}^P(\omega^P) \equiv \sup_{\pi,\tau} \frac{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s \tilde{r}_s | \omega^P\right]}{\mathbb{E}\left[\sum_{s=0}^{\tau-1} \delta^s | \omega^P\right]}$$

  - rule π specifies selection over primitive and auxiliary arms
  - $\tilde{r}_s$: period-s reward (can coincide with $R(\hat{\omega}^P)$ in case period-s arm is auxiliary)
- Index for search as before - but search adjusted to include discovery of auxiliary arms
- Index policy optimal in fictitious environment
- Difficulty: Recasting problem this way possible only if auxiliary arms corresponding to past states of same arm *never selected*
  - guaranteed by "'better-later-than-sooner" property

Extension 2: Search frictions

- Main theorem extends to settings where **pull** of an arm occupies **arbitrary number of periods** (before a different action may be taken)
- Relative length of time in which pulling arms is interrupted for search can be made arbitrarily small (by re-scaling payoffs and adjusting discount factor)
- Hence analysis extends to settings where
  - search and experimentation "virtually" in parallel

Conclusions

- Experimentation with **endogenous** set of alternatives determined by past searches
- Optimal policy: **index policy**
  - "physical" arms: Gittins (1979) index
  - "search" arm: special index (accounts for selection from new arms found)
- Constant, or improving, search technology: *search=replacement*
- Otherwise,
  - existing arms put on hold and resumed later
- Irreversible selection among endogenous alternatives:
  - "better-later-than-sooner" property: index policy optimal
- Applications:
  - mediated matching
  - design of search engines
  - R&D and patenting

Conclusions

THANKS!

Policy: formal definition

- Period-t decision: $d_t \equiv (x_t, y_t)$
  - $x_{it} = 1$ if "physical" arm i pulled; $x_{it} = 0$ otherwise
  - $y_t = 1$ if search; $y_t = 0$ otherwise
  - Sequence of decisions $d = (d_t)_{t=0}^{\infty}$ *feasible* if, for all $t \geq 0$:
    - $x_{jt} = 1$ only if $j \in I_t$
    - $\sum_{j \in I_t} x_{jt} + y_t = 1$
- Given sequence of feasible decisions $(d_t)_{t \geq 0}$, process $(\mathcal{S}_t)_{t \geq 0}$ generates natural filtration $\{\mathcal{F}_t\}_{t \geq 0}$
- Rule χ governing feasible decisions $(d_t)_{t \geq 0}$ is a **policy** iff sequence of decisions $\{d_t^{\chi}\}_{t \geq 0}$ under χ is $\{\mathcal{F}_t^{\chi}\}_{t \geq 0}$-adapted

Go back

Recursive characterization of index for search

- Index of search arm can be re-written as

$$\mathcal{G}^S(\omega^S) = \frac{\mathbb{E}^{\chi^*}\left[\sum_{s=0}^{\tau^*-1} \delta^s (r_s - c_s)|\omega^S\right]}{\mathbb{E}^{\chi^*}\left[\sum_{s=0}^{\tau^*-1} \delta^s|\omega^s\right]},$$

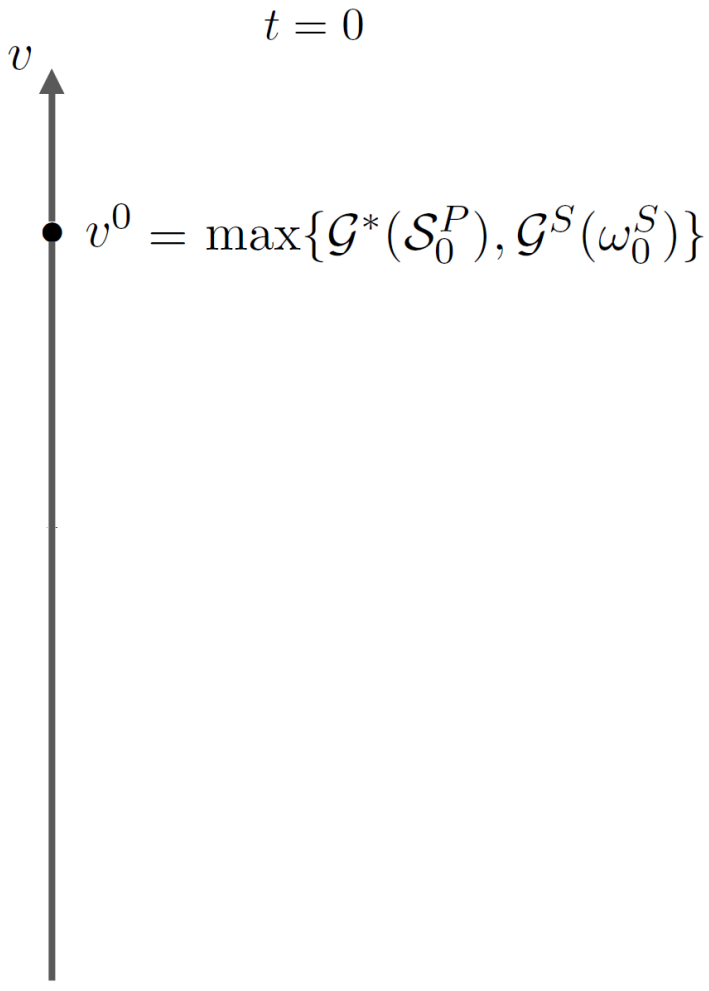
where χ^* is index policy and τ^* is first time $s \geq 1$ at which index of search and indexes of all physical arms obtained through search fall below value of search index at $s = 0$.

Go back

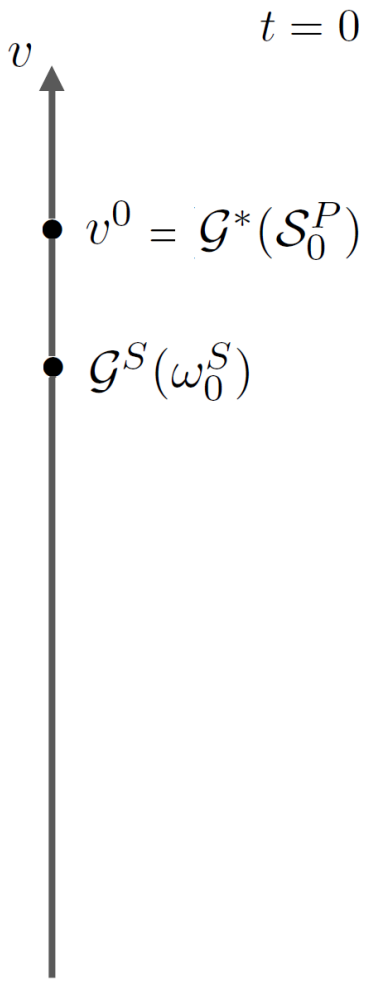
Proof of Lemma 1

- $v^0 = \max\{\mathcal{G}^*(\mathcal{S}_0^P), \mathcal{G}^S(\omega_0^S)\}$
- t^0: first time all indexes (including search) **strictly below** v^0 ($t^0 = \infty$ if event never occurs)
- $\eta(v^0|\mathcal{S}_0)$: discounted sum of rewards, net of search costs, till t^0 (includes rewards from newly arrived arms)
- $v^1 = \max\{\mathcal{G}^*(\mathcal{S}_{t^0}^P), \mathcal{G}^S(\omega_{t^0}^S)\}$ (note: $t^0 = \kappa(v^1|\mathcal{S}_0)$)
- ...
- $\eta(v^i|\mathcal{S}_0)$: net rewards between $\kappa(v^i|\mathcal{S}_0)$ and $\kappa(v^{i+1}|\mathcal{S}_0) - 1$
- Stochastic sequence of values $(v^i)_{i\geq 0}$, times $(\kappa(v^i|\mathcal{S}_0))_{i\geq 0}$, and discounted net rewards $(\eta(v^i|\mathcal{S}_0))_{i\geq 0}$

Proof of Lemma 1



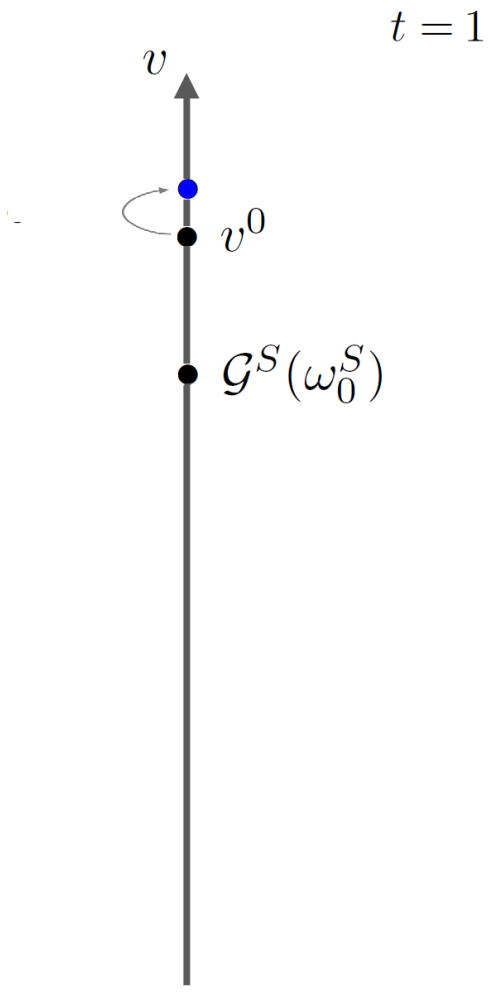
Proof of Lemma 1



$v^0 = \text{index of arm}$

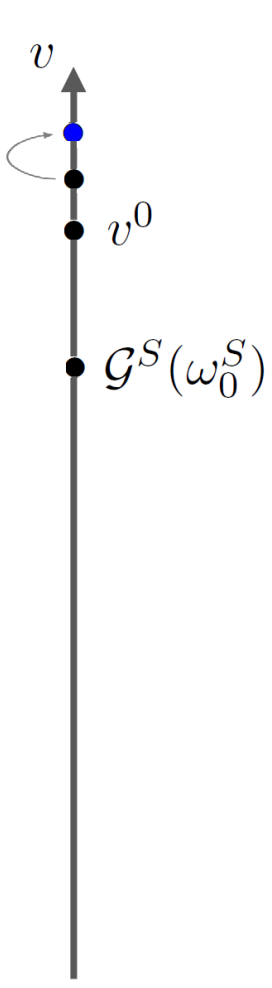
$\kappa(v^0|\mathcal{S}_0) = 0$

Proof of Lemma 1



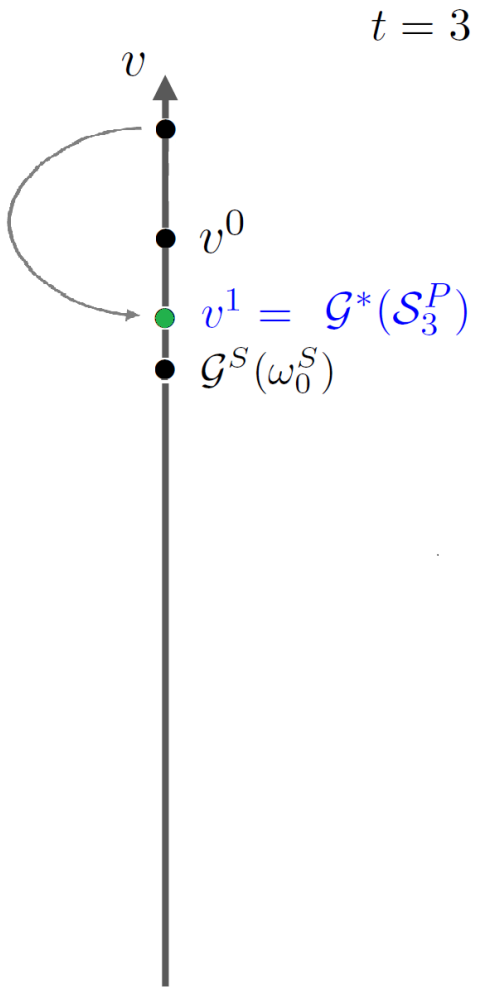
$$\kappa(v^0|\mathcal{S}_0)=0$$

Proof of Lemma 1



$$\kappa(v^0|\mathcal{S}_0) = 0$$

Proof of Lemma 1

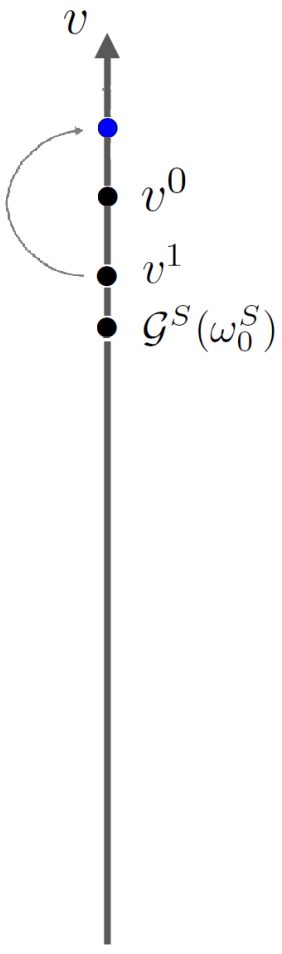


$$\kappa(v^0|\mathcal{S}_0) = 0$$

$$t^0 = \kappa(v^1|\mathcal{S}_0) = 3$$

Proof of Lemma 1

$t = 4$

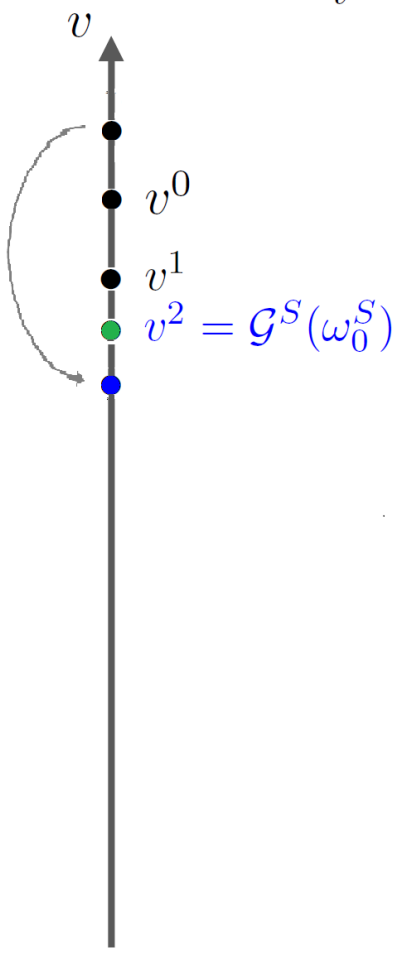


$$\kappa(v^0|\mathcal{S}_0) = 0$$

$$t^0 = \kappa(v^1|\mathcal{S}_0) = 3$$

Proof of Lemma 1

$t = 5$

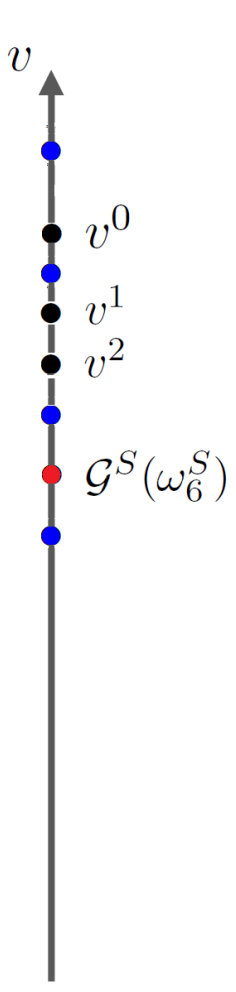


$$\kappa(v^0|\mathcal{S}_0) = 0$$

$$t^0 = \kappa(v^1|\mathcal{S}_0) = 3$$

$$t^1 = \kappa(v^2|\mathcal{S}_0) = 5$$

Proof of Lemma 1

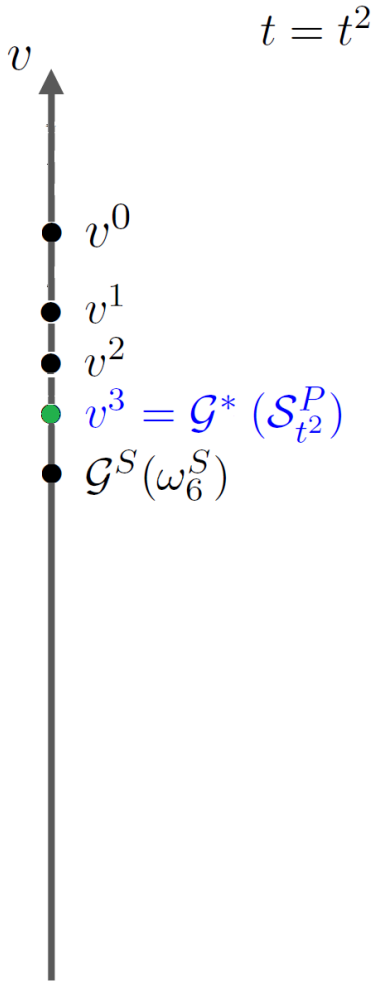


indexes of new arms

new index for search

$$\kappa(v^0|\mathcal{S}_0) = 0$$
$$t^0 = \kappa(v^1|\mathcal{S}_0) = 3$$
$$t^1 = \kappa(v^2|\mathcal{S}_0) = 5$$

Proof of Lemma 1



$$\kappa(v^0|\mathcal{S}_0) = 0$$

$$t^0 = \kappa(v^1|\mathcal{S}_0) = 3$$

$$t^1 = \kappa(v^2|\mathcal{S}_0) = 5$$

$$t^2 = \kappa(v^3|\mathcal{S}_0)$$

Proof of Lemma 1

- (Average) payoff under index policy:

$$\mathcal{V}(\mathcal{S}_0) = (1-\delta)\mathbb{E}\left[\sum_{i=0}^{\infty} \delta^{\kappa(v^i)}\eta(v^i)|\mathcal{S}_0\right].$$

- Starting at $\kappa(v^i)$, optimal stopping time in index defining v^i is $\kappa(v^{i+1})$
  - if v^i is index of physical arm, $\kappa(v^{i+1})$ is first time its index drops below v^i
  - if v^i is index of search arm, $\kappa(v^{i+1})$ is first time search index + index of **all arms** discovered after $\kappa(v^i)$ drop below v^i

- Hence, v^i = expected discounted sum of net rewards, per unit of expected discounted time, from $\kappa(v^i)$ until $\kappa(v^{i+1})-1$:

$$v^i = \frac{\mathbb{E}\left[\eta(v^i)|\mathcal{F}_{\kappa(v^i)}\right]}{\mathbb{E}\left[1-\delta^{\kappa(v^{i+1})-\kappa(v^i)}|\mathcal{F}_{\kappa(v^i)}\right]/(1-\delta)}$$

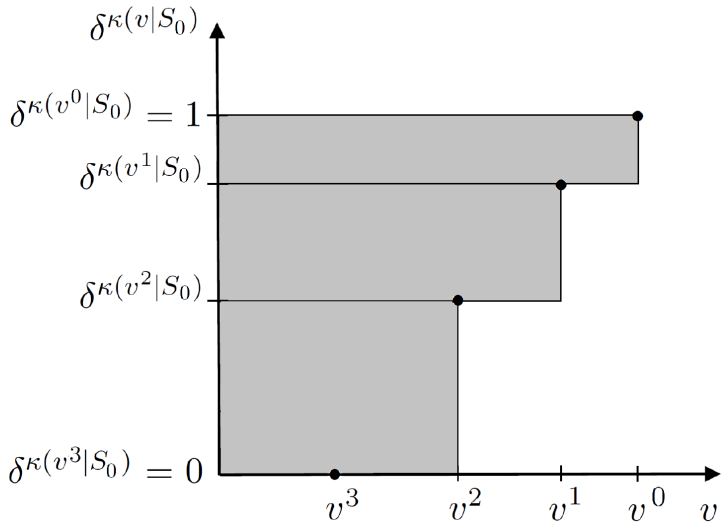
- Same true if multiple arms and/or search have index equal to v^i at $\kappa(v^i)$

Proof of Lemma 1

- Plugging in expression for v^i,

$$\mathcal{V}(\mathcal{S}_0) = \mathbb{E}\left[\sum_{i=0}^{\infty} v^i \left(\delta^{\kappa(v^i)} - \delta^{\kappa(v^{i+1})}\right) | \mathcal{S}_0\right]$$

$\sum_{i=0}^{\infty} v^i \left(\delta^{\kappa(v^i|\mathcal{S}_0)} - \delta^{\kappa(v^{i+1}|\mathcal{S}_0)}\right)$



- Therefore,

$$\mathcal{V}(\mathcal{S}_0) = \mathbb{E}\left[\int_0^{\infty} v \mathrm{d}\delta^{\kappa(v)} | \mathcal{S}_0\right] = \int_0^{\infty} \left(1 - \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)}\right) \mathrm{d}v$$

Go back

Proof of DP

- Want to show that $\mathcal{V}(\mathcal{S}_0)$ solves **dynamic programming equation**:

$$\mathcal{V}(\mathcal{S}_0) = \max\{ \underbrace{V^S(\omega^S|\mathcal{S}_0)}_{\substack{\text{value from searching} \\ \text{and reverting} \\ \text{to index} \\ \text{policy thereafter}}}, \underbrace{\max_{\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0^P(\hat{\omega}^P) > 0\}} V^P(\omega^P|\mathcal{S}_0)}_{\substack{\text{value from pulling} \\ \text{physical arm and} \\ \text{reverting to index} \\ \text{policy thereafter}}} \}$$

Auxiliary arms

- $e(\omega_M^A)$: state with single **auxiliary** arm yielding fixed reward M
- Note: $\kappa(v|\underbrace{\mathcal{S}_0 \vee e(\omega_M^A)}_{\mathcal{S}_0 + \text{auxiliary arm}}) = \begin{cases} \kappa(v|\mathcal{S}_0) & \text{if } v \geq M \\ \infty & \text{otherwise} \end{cases}$
- From Lemma 1, payoff from index policy when auxiliary arm added:

$$\begin{aligned} \mathcal{V}(\mathcal{S}_0 \vee e(\omega_M^A)) &= \int_0^\infty [1 - \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0 \vee e(\omega_M^A))}]\mathrm{d}v \\ &= M + \int_M^\infty [1 - \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)}]\mathrm{d}v \\ &= \mathcal{V}(\mathcal{S}_0) + \int_0^M \mathbb{E}\delta^{\kappa(v|\mathcal{S}_0)}\mathrm{d}v \end{aligned}$$

Auxiliary arms

$$\underbrace{D^S(\omega^S|e(\omega^S)\vee e(\omega_M^A))}_{\text{loss from starting with search given only search + auxiliary arm}} \equiv \underbrace{\mathcal{V}(e(\omega^S)\vee e(\omega_M^A))}_{\text{value under index policy given only search + auxiliary arm}} - \underbrace{V^S(\omega^S|e(\omega^S)\vee e(\omega_M^A))}_{\text{value of searching and reverting to index policy given only search + auxiliary arm}}$$

$$= \begin{cases} 0 & \text{if } M \leq \mathcal{G}^S(\omega^S) \\ > 0 & \text{if } M > \mathcal{G}^S(\omega^S) \end{cases}$$

Similarly, for physical arm in state ω^P:

$$D^P(\omega^P|e(\omega^P)\vee e(\omega_M^A)) = \begin{cases} 0 & \text{if } M \leq \mathcal{G}^P(\omega^P) \\ > 0 & \text{if } M > \mathcal{G}^P(\omega^P) \end{cases}$$

Dynamic Programming

Lemma 2

$\mathcal{V}(\mathcal{S}_0)$ solves dyna*mic programming equation (hence index policy optimal)*

Proof that $\mathcal{V}$ solves Bellman eq

- From Lemma 1: $D^S(\omega^S|e(\omega^S) \vee e(\omega^A_M)) = \begin{cases} 0 & \text{if } M \leq \mathcal{G}^S(\omega^S) \\ > 0 & \text{otherwise} \end{cases}$
- Can show ("tedious"): $D^S(\omega^S|\boldsymbol{\mathcal{S}_0}) = \int_0^{\mathbf{v^0}} D^S(\omega^S|\mathbf{e(\omega^S)} \vee \mathbf{e(\omega^A_M)})\mathrm{d}\mathbb{E}\delta^{\kappa(M|\mathcal{S}_0^P)}$
- Hence: $D^S(\omega^S|\mathcal{S}_0) = 0$

  $\iff D^S(\omega^S|e(\omega^S) \vee e(\omega^A_M)) = 0,\ \forall M \in [0, \max\{\mathcal{G}^*(\mathcal{S}_0^P), \mathcal{G}^S(\omega^S)\}]$

  $\iff \mathcal{G}^*(\mathcal{S}_0^P) \leq \mathcal{G}^S(\omega^S)$

  loss from starting with search $= 0$ iff search has largest index, and > 0 otherwise
- Similarly, $D^P(\omega^P|\mathcal{S}_0) = 0 \iff \mathcal{G}^P(\omega^P) = \mathcal{G}^*(\mathcal{S}_0^P) \geq \mathcal{G}^S(\omega^S)$
- Hence, $\mathcal{V}(\mathcal{S}_0) = \max\left\{V^S(\omega^S|\mathcal{S}_0), \max_{\omega^P \in \{\hat{\omega}^P \in \Omega^P : \mathcal{S}_0^P(\hat{\omega}^P) > 0\}} V^P(\omega^P|\mathcal{S}_0)\right\}$. ■

Validation

- Assumption: For any $\mathcal{S}$, and policy χ,

$$\lim_{t\to\infty} \delta^t \mathbb{E}^{\chi}\left[\sum_{s=t}^{\infty} \delta^s \left(\sum_{j=1}^{\infty} x_{js} r_{js} - c_s y_s\right) | \mathcal{S}\right] = 0$$

  - Solution to DP equation coincides with value function
  - Assumption satisfied if rewards/costs uniformly bounded
  - Also compatible with unbounded rewards/costs. E.g., arms are sampling processes, with rewards drawn from Normal distribution with unknown mean

Go back