On Radical Extensions to Multi-Armed Bandits and to Notions of Indexation

Kevin Glazebrook

Department of Management Science, Lancaster University.

Acknowledgements: Jake Clarkson, Chris Kirkbride (Lancaster), David Hodge (Nottingham), John Gittins (Oxford), Richard Weber (Cambridge), Kyle Lin, Roberto Szechtman (Naval Postgraduate School), EPSRC.

Talk Outline

1. Introduction, History
2. **Problem 1:** Dynamic Resource Allocation
3. **Problem 2:** Optimal Two-Speed Search
4. **Problem 3:** Intelligent Intelligence Gathering and Analytics

A Little History

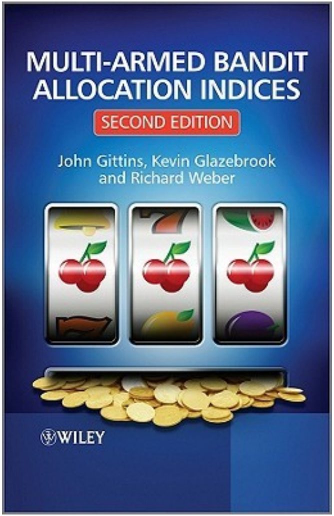
What are we talking about?

Models and methods for the dynamic allocation of a single key resource among a collection of stochastic reward generating projects (bandits) which are competing for it. Solutions which make use of simple project-based measures (indices) to guide decision-making;

Three historic papers

- JC Gittins and DM Jones (1974) "A dynamic allocation index for the sequential design of experiments", North-Holland, Amsterdam (Presented at EHS, Budapest, 1972)
- P. Whittle (1988) "Restless bandits: Activity allocation in a changing world", J.Appl. Prob
- DP Bertsimas and J Niño-Mora (1996) "Conservation laws, extended polymatroids and multi-armed bandit problems: A polyhedral approach to indexable systems", Maths of OR.

Multi-Armed Bandit Allocation Indices
Gittins, Glazebrook & Weber (2011)



Problem 1: Dynamic Resource Allocation

with Hodge, Kirkbride, Minty

- N stochastic reward generating/cost incurring projects are driven by the application of some divisible resource.
- At each decision epoch (state transition) an action $\mathbf{a} = (a_1, a_2, \ldots, a_N)$ is applied to the system.
- Admissible actions:

  $$A = \left[\mathbf{a}; a_n \in \{0, 1, \ldots, S\}, 1 \leq n \leq N, \text{ and } \sum_{n=1}^{N} a_n \leq R\right].$$
- System state: $\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{N}^N$.
- Project n: $\begin{cases} \text{Reward rate earned,} & d_n(x_n), \\ \text{Transition rates,} & q_n\left(x_n' \mid a_n, x_n\right) \end{cases}$
- Construct a policy for resource allocation to maximise the average return per unit time from all projects.
- $S = 1, R < N$: Whittle's RB Model
- $S = 1, R = 1$ and no resource $\Rightarrow$ no transition: Gittins' MAB
- Take $S = R$ in what follows.

Problem 1: Index Policies (1)

Optimisation Goal:

$$D^{opt} = \max_{\mathbf{u}} \sum_{n=1}^{N} D_n(\mathbf{u}) \qquad \text{(admissible policies)}$$

$$\downarrow$$

Lagrangian Relaxation (LR):

$$D(W) = \max_{\mathbf{u}} \sum_{n=1}^{N} \{D_n(\mathbf{u}) - WR_n(\mathbf{u})\} + WR$$

(constraint $\sum a_n \leq R$ abandoned)

$$D(W) \geq D^{opt}, \;\; W \in \mathbb{R}^+$$

$\min_{W} D(W)$ achieved at W^*. ("Soft" problem)

$$\downarrow$$

Projectwise Decomposition:

$$D(W) = \sum_{n=1}^{N} D_n(W) + WR, \quad \text{where}$$

$$D_n(W) = \max_{u_n} \{D_n(u_n) - WR_n(u_n)\} \qquad \text{(problem } P(n, W))$$

Problem 1: Index Policies (2)

(Full) Indexability:

Project n is **fully indexable** if there exist stationary policies $\{u_n(W); W \in \mathbb{R}^+\}$ such that

(a) $u_n(W)$ is optimal for $P(n, W)$, and

(b) $u_n(x_n, W)$ is decreasing in W $\forall$ x_n

$\downarrow$

Indices:

If project n is **fully indexable**, define **indices**

$$W_n(a_n, x_n) = \inf\{W; u_n(x_n, W) \leq a_n\} \qquad \text{(index as fair charge)}$$

$\downarrow$

Index Solution to LR:

If all K projects are **fully indexable** the above Lagrangian Relaxation is solved by the policy $\mathbf{u}(W)$ such that $\forall$ $\mathbf{x}$

$$\mathbf{u}(W, \mathbf{x}) = \mathbf{a} \iff W_n(a_n - 1, x_n) > W \geq W_n(a_n, x_n), \ \forall \ n.$$

In words: accumulate resource at each project until the fair charge for adding further resource falls below the prevailing charge W.

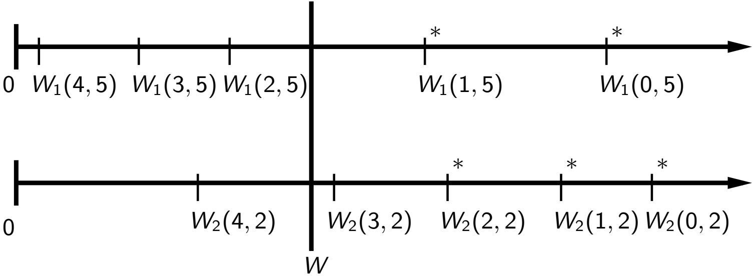
Problem 1: Index Policies (3)

Index heuristic for the original problem:

Increase resource levels at the projects in decreasing order of the appropriate indices/fair charges until the resource constraint is violated.

Example:

Let $N = 2,\ \mathbf{x} = (5, 2)$

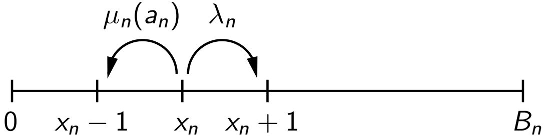


Optimal action for Lagrangian Relaxation: $(a_1, a_2) = (2, 4)$.

If $S = R = 5$, greedy index heuristic $*$ chooses $(a_1, a_2) = (2, 3)$.

Problem 1: A Queueing Control Example

- A team of R servers provides service at N stations. Station n has finite waiting room of size B_n. Completed services at station n earn a return d_n. Arrivals at full stations are lost.
- How to dynamically allocate the R servers among the stations to maximise the aggregate return rate?
- Dynamics at station n with a_n servers:



  Service rate $\mu_n(a_n)$ is strictly increasing and strictly concave in a_n.
  Reward rate: $d_n(x_n) = d_n \lambda_n I(x_n < B_n)$
- Station n is **fully indexable**.

Problem 1: Queueing Control (continued)

Analysis of problems with $N = 2,\ S = R = 25,\ d_1 = d_2 = 1,$

$$\mu_n(a_n) = a_n \mu_n (a_n + \nu_n)^{-1},\ n = 1,\ 2$$

and a range of choices for $\lambda_1,\ \lambda_2,\ \mu_1,\ \mu_2,\ \nu_1,\ \nu_2,\ B_1,\ B_2$.

	MIN	LQ	MED	UQ	MAX	#problems
Greedy Index	0.0023	0.0148	0.0235	0.0336	0.1199	5250
Optimum Static	17.9544	23.4628	25.4526	27.4720	33.8567	5250

Percentage reward rate deficit compared to optimum

Problem 1: (Full) Indexability

- Indexability is guaranteed for Gittins' MAB model;
- Indexability is **NOT** guaranteed for Whittle's RB model. Usually established (when true) using direct arguments for particular models. General approaches based on conservation laws/polyhedral ideas espoused by Niño-Mora (2001);
- (Full) indexability for the general DRA model has only been established for projects with birth-death dynamics exhibiting diminishing returns as the resource increases. DP-based proofs are tough.
  - Numerical tests for full indexability may be available;
  - Full indexability may be available locally but not globally - Hodge and Glazebrook (2011);
  - Can use policy improvement to explore good (but sub-optimal) solutions to the Lagrangian relaxation which have an indexable structure - Glazebrook et al. (2014).
  - See Graczová and Jacko (2014).

Problem 1: Performance of Policies

- Gittins' index policies are optimal for Gittins' MAB model, index-based performance bounds on general policies available;
- Strong empirical performance of Whittle's index policy observed widely;
- Under mild conditions, Whittle's index policy is optimal for Whittle's RB model in a limit as the amount of resource (R) and the number of bandits (N) scale in proportion - Weber and Weiss (1990); see also Verloop (2016);
- Polyhedral approaches to performance bounds based on Niño-Mora's indexability work sometimes available for RBs - e.g. Glazebrook et al (2009);
- Other forms of asymptotic optimality have been established for specific (queueing) RB models - Glazebrook et al. (2009);
- Weber and Weiss (1990) asymptotic optimality results extend to the general DRA model - Hodge and Glazebrook (2015).

Problem 2: Optimal Two-Speed Search

with Clarkson, Lin

Hiding probability (prior): p_i

$$\sum_{i=1}^{N} p_i = 1$$

Fast Search: $t_{i,f}$, $q_{i,f}$

$$0 < t_{i,f} < t_{i,s}$$

Slow Search: $t_{i,s}$, $q_{i,s}$

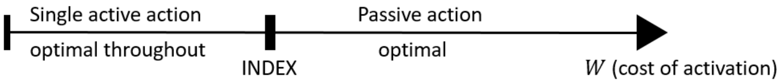
$$0 < q_{i,f} < q_{i,s} < 1$$

Location (Box) i $(1 \leq i \leq N)$

Goal: Determine a policy to minimise the expected time to find the object.

Problem 2: Some History

- Single-speed problem first solved in 1962 (Blackwell): search box with maximal $p_i' q_i / t_i$ with p_i' the current posterior;
- Kelly (1979) argues that the single-speed problem can be modelled as a MAB, with Blackwell's policy the Gittins index policy;
- The two-speed problem can be modelled as a **family of alternative superprocesses**, a variant of the MAB in which bandits have several active actions. Notion of **strong indexability** (Whittle, 1980): for given (box, state)



- If within-box subsequences of search modes $A_i = \{a_{i,n}; n \in \mathbb{Z}^+\}$ are pre-specified for each box i, optimal policy is a Gittins index policy with indices $G_i(\bullet, A_i)$, $1 \leq i \leq N$.

Problem 2: Main Result

Theorem

(a) If any box j satisfies

$$\frac{q_{j,s}}{t_{j,s}} \geq \frac{q_{j,f}}{t_{j,f}} \quad (j \in \mathcal{S})$$

then an optimal search sequence exists where box j is only searched slowly.

(b) If any box j satisfies

$$\frac{q_{j,f} \cdot (1 - q_{j,s})}{t_{j,f}} \geq \frac{q_{j,s}}{t_{j,s}} \quad (j \in \mathcal{F})$$

then an optimal search sequence exists where box j is only searched fast.

Problem 2: Main Result

Theorem

(a) If any box j satisfies

$$\frac{q_{j,s}}{t_{j,s}} \geq \frac{q_{j,f}}{t_{j,f}} \quad (j \in \mathcal{S})$$

then an optimal search sequence exists where box j is only searched slowly.

(b) If any box j satisfies

$$\frac{q_{j,f} \cdot (1 - q_{j,s})}{t_{j,f}} \geq \frac{q_{j,s}}{t_{j,s}} \quad (j \in \mathcal{F})$$

then an optimal search sequence exists where box j is only searched fast.

- Proof of (a) makes extensive use of stochastic coupling;

Problem 2: Main Result

Theorem

(a) If any box j satisfies

$$\frac{q_{j,s}}{t_{j,s}} \geq \frac{q_{j,f}}{t_{j,f}} \quad (j \in \mathcal{S})$$

then an optimal search sequence exists where box j is only searched slowly.

(b) If any box j satisfies

$$\frac{q_{j,f} \cdot (1 - q_{j,s})}{t_{j,f}} \geq \frac{q_{j,s}}{t_{j,s}} \quad (j \in \mathcal{F})$$

then an optimal search sequence exists where box j is only searched fast.

- Proof of (b) makes extensive appeal to structure of Gittins index policies;

Problem 2: Main Result

Theorem

(a) If any box j satisfies

$$\frac{q_{j,s}}{t_{j,s}} \geq \frac{q_{j,f}}{t_{j,f}} \quad (j \in \mathcal{S})$$

then an optimal search sequence exists where box j is only searched slowly.

(b) If any box j satisfies

$$\frac{q_{j,f} \cdot (1 - q_{j,s})}{t_{j,f}} \geq \frac{q_{j,s}}{t_{j,s}} \quad (j \in \mathcal{F})$$

then an optimal search sequence exists where box j is only searched fast.

- Above result yields an easily computed upper bound on $\frac{V_B - V^*}{V^*}$ where V_B is the expected search time of BSM and V^* is the optimal expected search time.

Problem 2: Main Result

Theorem

(a) If any box j satisfies

$$\frac{q_{j,s}}{t_{j,s}} \geq \frac{q_{j,f}}{t_{j,f}} \quad (j \in \mathcal{S})$$

then an optimal search sequence exists where box j is only searched slowly.

(b) If any box j satisfies

$$\frac{q_{j,f} \cdot (1 - q_{j,s})}{t_{j,f}} \geq \frac{q_{j,s}}{t_{j,s}} \quad (j \in \mathcal{F})$$

then an optimal search sequence exists where box j is only searched fast.

- Write $\mathcal{H} = \{1, 2, \ldots, N\} \setminus \{\mathcal{S} \cup \mathcal{F}\}$. How to search $\mathcal{H}$-boxes?

Problem 2: Building Intuition

- Notion of **immediate benefit** from search mode $\bullet \in \{s, f\}$ given by $\text{IB}_\bullet = q_\bullet / t_\bullet$;
- Notion of **future benefit** from search mode $\cdot$ given by

$$\text{FB}_\bullet = \frac{d}{dx}\left\{\frac{1-p}{p(1-q_\bullet)^{x/t_\bullet}+1-p}\right\}\Bigg|_{x=0} = \frac{-p(1-p)\log(1-q_\bullet)}{t_\bullet};$$

Problem 2: Building Intuition

- Notion of **immediate benefit** from search mode $\bullet \in \{s, f\}$ given by $\mathrm{IB}_\bullet = q_\bullet / t_\bullet$;
- Notion of **future benefit** from search mode $\cdot$ given by

$$\mathrm{FB}_\bullet = \frac{d}{dx}\left\{\frac{1-p}{p(1-q_\bullet)^{x/t_\bullet} + 1 - p}\right\}\Bigg|_{x=0} = \frac{-p(1-p)\log(1-q_\bullet)}{t_\bullet};$$

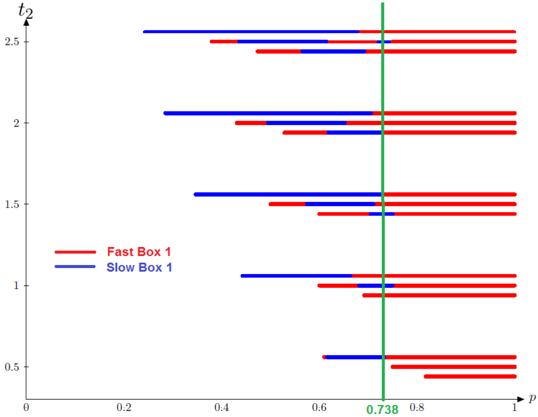
- For $\mathcal{S}$-boxes we have $\mathrm{IB}_s \geq \mathrm{IB}_f$ and $\mathrm{FB}_s \geq \mathrm{FB}_f$ while for $\mathcal{F}$-boxes we have $\mathrm{IB}_f \geq \mathrm{IB}_s$ and $\mathrm{FB}_f \geq \mathrm{FB}_s$;
- For $\mathcal{H}$-boxes we have $\mathrm{IB}_f \geq \mathrm{IB}_s$. Trade off IB-advantage of fast against (possible) FB-advantage of slow:

$$\alpha = \left(\frac{\mathrm{IB}_f}{\mathrm{IB}_s} - 1\right) > 0; \quad \beta = \frac{\mathrm{FB}_s}{\mathrm{FB}_f} - 1.$$

Natural choice of threshold satisfies
$\tilde{p}\alpha = (1-\tilde{p})\beta \Rightarrow p \geq \tilde{p} = \frac{\beta}{\alpha+\beta}$ search $\mathcal{H}$-box fast.

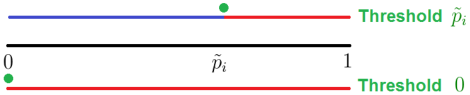
Problem 2: (Approximate Invariance)

- Two-box problem. Box 1 has two search modes with $q_{1,f} = 0.4$, $t_{1,f} = 1$, $q_{1,s} = 0.64$ and $t_{1,s} = 1.7$. Box 2 has one mode. We take $q_2 \in \{0.3 \text{ (upper)}, 0.6 \text{ (middle)}, 0.9 \text{ (lower)}\}$ and $0.5 \leq t_2 \leq 2.5$;
- Proposed threshold $\tilde{p} = 0.738$ for this example



Problem 2: A Two-Speed Heuristic Policy

- For each $i \in \mathcal{H}$, **consider** searching box i fast if p_i is above the threshold $\tilde{p}_i$.
- For p_i lower than the threshold $\tilde{p}_i$, try both policies searching all fast and all slow



- Choice of two thresholds above which we consider fast: 0 or $\tilde{p}_i$.
- Leads to up to $2^{|\mathcal{H}|}$ policies, let that with the lowest expected search time be the *best threshold* (BT) policy.

Problem 2: Numerical Results: $N = 4$

Results reported as percentage over optimal value estimate.

Simulated $N \times 1,000$ pairs of boxes, using a pre-selected 5 priors representing a scenario.

Table: Test with $N = 4$ and $|\mathcal{H}| = 4$.

	Uniform Prior		
Metric/Policy	DR	BSM	BT
Mean	1.42	0.007	0.006
75th Percentile	2.11	0	0
95th Percentile	6.59	0.042	0.034
	One Box Dominates Prior		
Metric/Policy	DR	BSM	BT
Mean	1.09	0.043	0.010
75th Percentile	1.42	0	0
95th Percentile	5.01	0.271	0.043

- DR: Detection Rate, all boxes in $\mathcal{H}$ searched fast.
- BSM: The best of the $2^{|\mathcal{H}|}$ single-mode policies.
- BT: Best Threshold, our only two-speed heuristic.

Problem 2: Numerical Results: $N = 8$

Results reported as percentage over optimal value estimate.

Simulated $N \times 1,000$ pairs of boxes, using a pre-selected 5 priors representing a scenario.

Table: Test with $N = 8$ and $|\mathcal{H}| = 8$.

	Uniform Prior		
Metric/Policy	DR	BSM	BT
Mean	2.24	0.004	0.004
75th Percentile	3.45	0	0
95th Percentile	5.43	0.008	0.007
	One Box Dominates Prior		
Metric/Policy	DR	BSM	BT
Mean	2.06	0.017	0.006
75th Percentile	2.98	0	0
95th Percentile	4.98	0.023	0.009

- DR: Detection Rate, all boxes in $\mathcal{H}$ searched fast.
- BSM: The best of the $2^{|\mathcal{H}|}$ single-mode policies.
- BT: Best Threshold, our only two-speed heuristic.

Problem 3: Intelligent Intelligence Gathering and Analysis
with Kirkbride, Marshall, Szechtman

Scenario: Copious amounts of data possibly related to an intelligence question are available from a number of sources of unknown quality/relevance. Analytical capability is limited as is the time available. A processor makes an initial estimate of the value/relevance of individual items drawn from the sources. Only items of high value/relevance should be passed on for analysis.

Proposal: Model as a Multi-Armed Bandit Allocation (MABA) model with finite horizon (T), a variant of the MAB in which any M (typically $\ll T$) of the T observed rewards are claimed/realised. Rewards are claimed (or not) immediately after observation. Goal is to maximise aggregate reward claimed. Bayesian formulation.

Problem 3: MABA Model

- Active source/bandit transitions:
  - Pre-Activation State: x (sufficient statistic for source quality)

    $\downarrow$
  - Reward (r) sampled from $\{p(\bullet \mid x), \bullet \in \Sigma\}$, Σ a finite connected subset of $\mathbb{N}$. Reward r is claimed or not.

    $\downarrow$
  - Post-Activation State: $X(x, r)$ (new value of sufficient statistic)
- Non-active source/bandits do not generate rewards nor change state.
- Bandits are activated one at a time over horizon T. No more than M may be claimed.

Problem 3: MABA Analysis

Approach: Relax MABA to MABA*. In MABA*, any number of sources/bandits may be activated at $t = 0, 1, \ldots, T - 1$. Each reward sampled may be claimed or not.
Constraints for MABA*:

$$E(\text{total activations}) \leq T, \quad E(\text{total rewards claimed}) \leq M.$$

Plainly, value of MABA* $\equiv V^* \geq V \equiv$ value of MABA.

Idea: Solve MABA* by means of Lagrangian relaxation. This induces a decomposition of the problem by source/bandit. Need to solve a thresholding/stopping problem for each source/bandit. Stopping problem has an index solution for any given reward threshold.

Problem 3: MABA Analysis (Contd.)

Consider a source/bandit which has been activated at $0, 1, \ldots, t-1$ and which is in state x.
Consider <u>stopping times</u> on bandit activation from this point:

$$\tau = \min\left(\min\left[s; s \geq t \text{ and } X(s) \in \omega_s\right]; T\right).$$

For given reward threshold C, we have associated <u>index</u>

$$\omega_t(x, C) = \max_{\tau} \frac{E\left\{\sum_{s=t}^{\tau-1}(r_s - C)^+ \mid x\right\}}{E(\tau \mid x)}$$

Result: There exist W^*, C^* such that MABA* is solved as follows: at all epochs t activate all sources/bandits k for which $\omega_{kt}\{X_k(t), C^*\} \underset{(-)}{\geq} W^*$ and claim all rewards $\underset{(-)}{\geq} C^*$.

Remark: Construct a 'single arm activation' version of the above policy for MABA* and thereby develop heuristics for P in the form of admissible approximations to it.

Some Open Issues

- **Problem 1:** Major unresolved issues concerning (full) indexability and index policy performance;
- **Problem 2:** Game theoretic versions, two-speed search on a graph, multi-speed search;
- **Problem 3:** Most effective construction of admissible approximations to index policy, competing approaches and formulations.