Asymptotic optimality of multi-action restless bandits

David Hodge
Kevin Glazebrook

The University of Nottingham
&
Lancaster University

November 26th 2010
YEQT IV, EURANDOM, Eindhoven
Multi-armed bandits

History
Multi-armed bandits date back to times long before the term was coined.

What are they?
- A collection of $n$ reward-generating objects;
- Rewards are incurred in continuous time;
- **Action/Decision**: which objects to activate at each timestep?
- Reward rates depend on current state and action;
- Markovian dynamics also depend on whether a state is active or passive;

Applications? **Everywhere in stochastic control!**
- Natural, obvious, direct uses in queues, and machine maintenance;
- Also in financial decision making;
- A very wide variety of MDPs.
The problem

To optimally determine a dynamic policy of activation decisions, at each system state, which bandit to activate and leave all other bandits passive. Passive $\Rightarrow$ no change in state!

What does optimally mean above?

- Discounted rewards (over infinite horizon);
- Long-run average rewards.

Examples

- Drug trials – which drug to use on the next patient?
- Single server queue with holding costs – which class to serve next?
Optimality of Gittins

Theorem

The solution, \( \pi \), maximizing

\[
V_\pi = E_\pi \left[ \sum_{t=0}^{\infty} \beta^t R_{j(t)}(x_j(t)(t)) \mid x(0) = x \right],
\]

is characterized by index functions \( \mathcal{I}^j(\cdot) \) for each bandit \( j \in \{1, \ldots, n\} \).

Optimal policy \( \pi \) acts on bandit \( j \) at time \( t \) if

\[
\mathcal{I}^j(x_j(t)) = \max_{1 \leq i \leq n} \mathcal{I}^i(x_i(t))
\]

Note:

- One active bandit at each time;
- Passive bandits are fixed.
Subsidy problem approach (primarily Whittle)
Various proofs from Gittins, Jones, Weber, Whittle

The retirement option

- Introduce a new bandit with fixed constant reward $W$;
- Equivalent to a reward $W$ for passivity;
- Characterize the value function in terms of $W$;
- Identify the value function as a solution to the original DP, for appropriate $W$.

Optimality?

- When only one active choice, yes!
- More than one active bandit, no! (Sometimes yes)
Restless bandits

What are they?

- Passive bandits can evolve;
- Passive bandits reward rates now matter (previously could be reassigned and neglected);
- We consider discrete state space restless bandits.

How much harder?

- Tsitsiklis & Papadimitriou showed PSPACE-hard. This is (probably!) worse than NP-Hard.

Applications?

Far too many to list!
Whittle approach for restless bandits

What’s been tried?
- $W$-subsidy approach still applies;
- Equivalent to rewarding $W$ for being passive;
- (or $-W$ if minimizing some costs)
- Index policies no longer necessarily optimal.
- Conjecture of asymptotic result...false! (Weber & Weiss 1990)

How do indices arise?
- Introduce passivity reward $W$;
- Bandits become independent;
- Lagrangian relaxation attains optimum (with $W$);
- Index = Fair charge = $W$ value at which optimal policy changes;
- Indexability: passive set monotone increasing in $W$. 
Model

- Define a bandit on a finite state space \( \{1, 2, \ldots, k\} \);
- Take \( n \) copies of this bandit;
- **Two** actions: active or passive for each bandit;
- Reward rate \( g(i, a) \) in state \( i \) under action \( a \);
- Long-run average reward objective;
- \( m \) of \( n \) bandits can be activated with \( m \approx \alpha n \), \( \alpha \in (0, 1) \);
- Different Markovian evolution matrices for active or passive.

Conjecture

If the bandits are indexable then the policy which, in each state, activates the \( m \) indices with current highest value, achieves asymptotically optimal reward per bandit as \( n \to \infty \) with \( m/n \to \alpha \).  

False! (rarely and by very little)
Overview

- Two problems: hard constraint \( m = \alpha n \), relaxed constraint \( \mathbb{E}m = \alpha n \);
- Inequalities:

\[
R_{\text{ind}}^{(n)}(\alpha) \leq R_{\text{opt}}^{(n)}(\alpha) \leq R_{\text{rel}}^{(n)}(\alpha) = nr(\alpha);
\]

- Inequality 2 is a per bandit (i.e. \( \div n \)) equality – relaxing \( m = \alpha n \) to \( \mathbb{E}m = \alpha n \) doesn’t improve reward per bandit;
- Indexability is not sufficient for 1 to be an order \( n \) equality;
- Indexability plus global attraction of a fluid limit differential equation \( \Rightarrow \) asymptotic optimality.
Weber & Weiss provide a (hard sought) counterexample above. Constructing an indexable bandit not satisfying the differential equation condition on four states.

Theorem

Global attraction of a unique solution to the derived fluid limit differential equation in two and three dimensions is guaranteed.

Question: What happens if we extend the action space?

More than just active, 1, or passive, 0, …

- Does indexability still make sense?
- What constraints are natural?
- Do we have asymptotic optimality?

Before we address these we ask ‘What more has been shown?’
### Intervening years – application areas

#### Areas with an interest – 1990 to present

- ADP/LP relaxations: Exploration v Exploitation (Powell)
- Bandwidth allocation
- Complexity (Papadimitriou & Tsitsiklis)
- Maintenance (Glazebrook)
- Military applications: primarily target selection
- Network optimization
- PCLs, high-level abstract indexability (Niño-Mora)
- Revenue management: esp. retail (Caro & Gallien)
- Optimal search: e.g. the Cow-path problem
- Sensor management
- Warranties (Glazebrook)
- More general resource allocation (Glazebrook, Niño-Mora)

Around 100 references from works in a wide variety of areas.
More general resource allocation
Multi-action bandits

Model
- Multiple levels of activity;
- Extended Markovian dynamics;
- Varying resource consumption;
- More general resource constraints.

Summary
- Niño-Mora: very general, gives heuristics with knapsack concerns;
- Glazebrook, Hodge, Kirkbride:
  - Indexability of multi-action restless bandits – server pools & replenishment;
  - Performance evaluation of index heuristics;
  - Indexability under state dependent resource consumption.
Multi-action asymptotic framework

Model

- Define a bandit on a finite state space \( \{1, 2, \ldots, k\} \);
- Take \( n \) copies of this bandit;
- Many actions: \( a \in \{0, 1, 2, \ldots, A\} \) for each bandit;
- Reward rate \( g(i, a) \) in state \( i \) under action \( a \);
- Long-run average reward objective;
- \( m \) units of activity to use across \( n \) bandits – i.e. \( m \cong \beta n, \beta \in (0, A) \);
- Different Markovian evolution matrices depending on action \( a \).
What does indexability mean?

### Multi-action finite state restless bandit

- Decouple bandits with $W$-passivity relaxation (equivalently mean usage constraint);
- We’re talking state-wise monotonicity of bandit optimal policy in a $W$-passivity relaxation;
- In a given state $x$:  
  - at high $W$ we use a low action,
  - at low $W$ we use a high action;
- Given $x$, we see $W$-values at which the optimal policy transitions between actions $a$;
- $I(x, a) \equiv I_x(a) =$ value of $W$ at which optimal policy is indifferent between $a$ and $a - 1$;
- $\forall x$, $I_x(1) \geq I_x(2) \geq I_x(3) \geq \ldots \geq I_x(A)$ (indexability).
Asymptotic optimality of greedy index policy

New result

**Theorem**

If we take \( n \) copies of an indexable restless bandit (as previously described), and if the fluid limit differential equation for the proportion of bandits in each state has a single-point limit set, then the greedy multi-action index policy agrees with both the strict resource constraint and relaxed constraint problems in average reward per bandit:

\[
\lim_{n \to \infty} \frac{R_{\text{ind}}^{(n)}(\beta)}{n} = \lim_{n \to \infty} \frac{R_{\text{opt}}^{(n)}(\beta)}{n} = r(\beta).
\]
Stage 1: Establish that $R_{opt}^{(n)}(\beta) \sim R_{rel}^{(n)}(\beta)$ – difference is $o(n)$

You can modify the Weber & Weiss argument:

- **Bright idea:** Consider the evolution of $n$ bandits under the optimal relaxed policy;
- **Zoom in on a single bandit** and observe its equilibrium $\pi$ on $\{1, 2, \ldots, k\}$;
- **Now make rational** ($\mathbb{Q}$) assumptions, incl. $n$ such that $n\pi_i \in \mathbb{N}$;
- **Now start $n$ bandits** from $x^* \in \{1, 2, \ldots, k\}^n$ mirroring $\pi$;
- **The relaxed optimal policy** will use exactly $\beta n$: use that policy for fixed time $\delta$. A suboptimal, feasible(!), policy for the hard constraint which almost achieves $r(\beta)$ per bandit.

**Theorem**

*This establishes that asymptotically the strict $m = \beta n$ and $\mathbb{E}m = \beta n$ problems have the same reward per bandit.*
Stage 2: Evaluate the greedy index policy

- Space scaling $\Rightarrow z^{(n)} \in [0, 1]^k$ with jumps of size $1/n$;
- Time scaling $\Rightarrow$ rates of $z^{(n_1)} \sim$ rates of $z^{(n_2)}$ for all $n_1, n_2$;
- For a known set of indices $\mathcal{I}_x(a)$ the evolution of $z^{(n)}$ under the index policy can be compared with a ‘piecewise not-quite-linear’ $k$-dimensional differential equation:
  \[
  \frac{dz}{dt} = \sum_{i,j} z_i \phi_i(z, \lambda_{ij}(\cdot))e_{ij}.
  \]
- ‘$\|z^{(n)}(t) - z(t)\|$ is small’ (same mean rewards);
- Idea: Identify the relaxed single-bandit equilibrium $\pi$ from earlier as a stationary point!
- Indexability $\Rightarrow$ uniqueness of stationary point.
Motivating areas

Direct:
- Many flows models in communication networks;
- Large scale bandit problems.

Indirect:
- Theoretical justification that greedy index-based heuristics are strong;
- Motivation to study approaches to NP-Hard bandit problems via approximations with index-interpretations;
- Problems in the many diverse areas mentioned earlier now may have a much closer class of problems with known asymptotically optimal policies.
Open questions

Where now?

- Can we quantify suboptimality in counterexamples? (Likely yes!) How large suboptimality?
- Infinite bandit state spaces?
Thank you