RISK-SENSITIVE AND RISK-NEUTRAL MULTI-ARMED BANDITS: OPTIMALITY OF INDEX POLICIES

Joint Work with
Eric V. Denardo and Haerchurl Park

Minerva Summer School, Ein Gedi – June 5, 2007
OUTLINE

• The standard (risk-neutral) Multi-Armed Bandit Problem (MABP) and its optimal index-solution
  – Model
  – Review of previous approaches

• Outline a new constructive derivation of an optimal index-solution

• Risk-seeking and risk averse MABP

Contributions:
A constructive proof of optimality of index policies for classic risk-neutral MABP – Efficient algorithm
  Extension to risk sensitive model – Risk averse and risk seeking

Note: Some technical issues will be ignored
MULTI-ARMED BANDITS (MABP)

- There are $K$ bandits – each is a Markov chains with rewards

- The decision-process:
  1. Observe the current state of each chain.
  2. Select one of the chains and play it once.
     a. The chain that was played evolves for one transition. Total stopping is possible.
     b. A reward is gained (it depends on the state of the pulled bandit and the outcome).
     c. States of Chains that were not played are unchanged.
  3. Return to Step 1.

Traditionally, the goal is to maximize expected discounted reward. This is the risk-neutral case.
**SOLUTION OF MABP (risk-neutral)**

**COMPUTATIONAL DIFFICULTY:**
If one has 10 states in each of 20 chains, one has to specify what to do in \(10^{20}\) situations (multi-states).

But, the size of the data is \(O[10^2 \times 20]\).

**The key result (Gittins and Jones [1974]):**
Each state of each bandit (Markov chain with rewards) has an *index* that depends *solely* on that chain’s data. It is optimal to play a bandit whose current state has the largest index.

Index policies are *priority policies.*
Method #1 - Gittins and Jones [1974]: For each Markov chain $k$, state $i$ of that chain and stopping time $\tau$ that starts at $i$, let

$$I(\tau, i) = \frac{E_i[\text{discounted return till time } \tau]}{E_i[1 + \beta + \beta^2 + \ldots + \beta^{\tau}]}$$

The **index of** $i$ is the optimal ratio -- it is optimal to select the bandit whose state’s index is highest.

Arguments – probabilistic, using stopping times.

Method #2 - Whittle [1982]: Introduced an (artificial) option retiring at any moment and earning a retirement reward of $M$. When $M$ is very large, one retires immediately. When $M$ is very small, one never retires. Whittle used parametric analysis in $M$, to show that state-dependent break-even points are indices that determine priority.
**Method #3 - Kathehakis and Veinott [1987]:**
Considered a “restart” problem and used an interchange argument of a stopping time.

**Method #4 - Weiss [1988] and others:**
Used interchange arguments for stopping times.

**Method #5: El Karoui and Karatzas [1993]:**
Used martingales.

**Method #6 - Tsitsiklis [1994]:** Used an induction on the number of states -- nonconstructive an applies to continuous time.


**REMARK:** Most of these proofs do not focus on constructing algorithms. All use probabilistic ideas.
NEW APPROACH

• Replaces index-comparison with domination of pairs of computable characteristics
• Replaces (real) probability theory (stopping times) with linear algebraic arguments relying on Nonnegativity and Dynamic Programming
• Develops a fast algorithm that is based on Tsitsiklis’s inductive argument.
• Applies to a risk-sensitive problems.
**MARKOV DECISION PROBLEMS**

At epochs 1, 2, …, :

1. Observe the state of the Markov chain, say state $s \in S$.

2. Select an action $k$ from a finite set $D(s)$.
   
   a) Stopping occurs with probability $P^k_s$ and reward $Z^k_s$ is earned.
   
   b) Transition to state $t$ occurs with probability $P^k_{st}$ and a reward $Z^k_{st}$ is earned.

3. Return to Step 1.

**Notes:** “Stopping” is not traditional.

“Rewards” can be $>0$, $<0$, $=0$.

A (stationary nonrandomized) policy $\delta$ selects for each state $s$ an action $\delta(s)$ in $D(s)$.

$(\nu^\delta)_s$ - the expected utility (infinite horizon; conditions for convergence).
Let $0 \leq \beta \leq 1$ be the discount rate. We then have

$$(v^{\delta,n})_s = r_s^{\delta(s)} + \sum_t Q_{st}^{\delta(s)} v^{\delta,n-1}_t$$

$$(v^{\delta,0})_s = 0$$

where for each $s$ and $k$

$$r_s^k = P_s^k z_s^k + \sum_t P_{st}^k z_{st}^k$$

$$Q_{st}^k = \beta P_{st}^k$$

Remarks:

- Q’s are nonnegative, in fact, substochastic.
- Discounting is eliminated by enlarging stopping probabilities.
MABP AS MDP

- We assume that the **states** of the bandits are disjoint and $b(i)$ denotes the bandit of $i$.

- **Data:** $z_i, z_{ij}, P_i, P_{ij}$ (no need to specify bandit!)

- MABP is a MDC has **multi-states** $s=\{i_1, i_2, \ldots, i_K\} \in S$

- Refer: $i \in s$ or $i=s_k$

- Actions consist of selecting a bandit. When observing $s$ and selecting $k$, the transition is to the multi-state $t=s_k \leftarrow u$ -- it coincides with $s$ except that $s_k \rightarrow u$.

- **Rewards** and **transition-probabilities** are then

  \[
  z^k_s = z_{s_k}, \quad P^k_s = P_{s_k}
  \]

  \[
  z^k_{s,t} = z_{s_k t_k}, \quad P^k_{s,t} = P_{s_k t_k}
  \]

  where $t=s_k \leftarrow u$
For multi-state $s \in S$, action $k$ and $i = s_k$, let

\[ r_s^k = P_i z_i + \sum_j P_{ij} z_{ij} \triangleq r_i \]

\[ Q_{st}^k = \beta P_{ij} \triangleq Q_{ij} \]
TWO-PERIOD INTERCHANGE

Suppose $i$ and $j$ are available. It is better to pull $b(i)$ followed by pulling $b(j)$ over pulling $b(j)$ followed by pulling $b(i)$ means:

$$r_i + \sum_u Q_{iu} [r_j + \sum_w Q_{jw} V_{uw}] > r_j + \sum_w Q_{jw} [r_i + \sum_u Q_{iu} V_{wu}]$$

For each $i$, let

$$\alpha_i = \sum_u Q_{iu} \leq 1$$

We say that $i$ dominates $j$, written $i \triangleright j$, if:

$$[r_i + \alpha_i r_j > r_j + \alpha_j r_i] \iff r_i [1 - \alpha_j] > r_j [1 - \alpha_i]$$

**Lemma:** $\triangleright$ is transitive

**Proof.** $[i \triangleright j] \iff \left[ \frac{r_i}{1 - \alpha_i} > \frac{r_j}{1 - \alpha_j} \right]$ (nondegeneracy)
Theorem 1:
(a) There is an undominated state.
(b) If \( i \) is undominated, it is optimal to pull \( b(i) \) whenever \( i \) is part of the observed multi-state.

Proof: Part (a) is immediate from the transitivity of \( \succ \).

For part (b), consider a multi-state which contains \( i \). Let \( \pi \) be an optimal policy which does not select \( b(i) \) at some \( s \) that contains \( i \).
We will show that pulling \( b(i) \) at each such state and then following \( \pi \) is also an optimal policy.
By the product form of the set of optimal policy it then follows that selecting \( b(i) \) at each state that contains \( i \) is an optimal action.
Let $C$ the (nonempty) set of multistates $s$ with $i \in s$ and $\pi(s) \neq b(i)$.

Let $\sigma$ be the (history-remembering deterministic) policy that acts like $\pi$ except that it selects $b(i)$ when $s \in C$ is observed for the first time and then “mimicks” $\pi$ till $b(i)$ is used.

Also set:

$(\alpha^\pi)_s \equiv \alpha_v$ where $v \in s$ and $\pi(s)=b(v)$

$(r^\pi)_s \equiv r_v$ where $v \in s$ and $\pi(s)=b(v)$

$(\Delta^{\pi,i})_s \equiv r_i [1 - (\alpha^\pi)_s ] - [1-\alpha_i] (r^\pi)_s$

**Lemma:** If $i$ is undominated, $\Delta^{\pi,i}(s) \geq 0$
Comparison Lemma (interchange):

\[(v^\sigma)_C - (v^\pi)_C = (I - Q^\pi_C)^{-1} (\Delta^{\pi,i})_C\]

**Proof:** Linear algebra

**Conclusion:** If \( i \) is undominated and \( s \) contains \( i \), selecting \( b(i) \) at \( s \) is an optimal action.

Use the product form of the set of optimal policies – (though \( \sigma \) is not a stationary policy).
Theorem 2: If \( i \) is undominated, it is possible to eliminate \( i \) by changing the data of bandit \( b(i) \) to:

\[
\begin{align*}
    r_j &\leftarrow r_j + \frac{Q_{ji}}{1 - Q_{ii}} r_i \\
    Q_{jt} &\leftarrow Q_{jt} + \frac{Q_{ji}}{1 - Q_{ii}} Q_{it}
\end{align*}
\]

Proof:

**Intuition:** Pull \( b(i) \) till transition occurs to another state.

**Formal arguments:** Rely on the functional-equation characterization of optimal policies for MDP.

**Algorithm:** Iteratively eliminate states and update data.

**EFFORT:** \( \sum_k |N_k|^3 \)!
Record the terminal values of $r(i)$ and $\alpha(i)$ of each state $i$ when eliminated.

**Theorem 3:** When observing a multi-state, it is optimal to select the $b(i)$ any undominated state $j$ of that multi-state *(with respect to terminal data).*
**RISK-SEEKING MDP**

**Risk-seeking:** \( u(x) = \exp(\lambda x) \)

\[ u(w+x) = [\exp(\lambda w)] u(x) \]

Stopping with probability \( p_1 \) gaining \( A \) and continuing with probability \( p_2 \) gaining \( B \) and future utility \( v \):

\[ \text{expected utility=} [p_1\exp(\lambda A)] + [p_2\exp(\lambda B)]v \]

We then have

\[ (v^{\delta,n})_s = r^\delta(s) + \sum_t Q^{\delta(s)}_{st} v^{\delta,n-1}_t \]

\[ (v^{\delta,0})_s = 1 \]

where for each \( s \) and \( k \)

\[ r^k_s = P^k_s \exp[\lambda z^k_s] \geq 0 \]

\[ Q^k_{st} = P^k_{st} \exp[\lambda z^k_{st}] \geq 0 \]
RISK-AVERSE MDP

Risk-AVERSE: \( u(x) = -\exp(-\lambda x) \)
\( u(w + x) = [\exp(-\lambda w)] u(x) \)

Stopping with probability \( p_1 \) gaining \( A \) and continuing with probability \( p_2 \) gaining \( B \) and future utility \( v \):

\[
\text{expected utility} = -p_1[\exp(-\lambda A)] + p_2[\exp(-\lambda B)]v
\]

We then have

\[
(v^{\delta,n})_i = r_i^{\delta(i)} + \sum_j Q_{ij}^{\delta(i)} v_{j,n-1}
\]

\[
(v^{\delta,0})_i = -1 \quad \Leftarrow \text{!!!!}
\]

where for each \( s \) and \( k \)

\[
r_s^k = -P_s^k \exp[-\lambda z_s^k] \leq 0
\]

\[
Q_{st}^k = P_{st}^k \exp[-\lambda z_{st}^k] \geq 0
\]
RISK-SENSITIVE MDP

• All three cases have returns $r_s$’s and nonnegative transition rates $Q_{st}$’s
  – Risk Neutral (RN): the $Q_{st}$’s are substochastic
  – Risk Seeking (RS): $r_s \geq 0$
  – Risk Averse (RA): $r_s \leq 0$

DR79 (MP) and DR06 (SICOP) consider MDP under RA and RA. Conditions are determined that allow one to ignore the “final reward $\pm 1$” and verify that stationary optimal policies:
  (i) exist
  (ii) are computable by LP
  (iii) are characterized by a corresponding functional equation

Condition for RA/RS: every/some $Q^6$ is transient
RISK-SENSITIVE MABP

\[ r_s^k \Rightarrow r_i = \pm P_i \exp[-\lambda z_i] \propto 0 \]

\[ Q_{st}^k \Rightarrow Q_{ij} = P_{ij} \exp[\lambda z_{ij}] \geq 0 \]

where \( s \in S \) and \( i = s_k \)
DOMINATION - revisited

Recall domination $\succ$:

$$[i \succ j] \iff r_i[1-\alpha_j] > r_j[1-\alpha_i] \quad (\star)$$

**GENERALLY:** The relation is not transitive

**Example:** Consider $i\rightarrow(3,4)$, $j\rightarrow(.2,.5)$ and $u\rightarrow(-4,6)$

$3[1-.5] > .2[1-4] \text{ and } .2[1-6] > (-4)[1-.5]$

while

$3[1-6]<(-4)[1-4]$,

**Lemma:** $\succ$ is transitive if either all $r_i$'s or all $[1-\alpha_i]$'s are one-sided ($\geq 0$ or $\leq 0$)

**Remark:** zeroes of one-sided variables can be perturbed
SOLUTION OF MABP

• Risk-neutral : $1 - \alpha_i \geq 0$
• Risk-seeking : $r_i \geq 0$
• Risk-averse : $r_i \leq 0$

So, domination can be made transitive.

**Theorem 1:** Under either of the three cases: There is an undominated state.

If $i$ is undominated, it is optimal to pull $b(i)$ whenever $i$ is part of the observed multi-state.
**Theorem 2:** If i is undominated, it is possible to eliminate i by changing the data of bandit b(i) to:

\[
\begin{align*}
    r_j &\leftarrow r_j + \frac{Q_{ji}}{1 - Q_{ii}} r_i \\
    Q_{jt} &\leftarrow Q_{jt} + \frac{Q_{ji}}{1 - Q_{ii}} Q_{it}
\end{align*}
\]

*Q_{ii} < 1* is guaranteed

**Algorithm:** Iteratively eliminate states and update data.

**EFFORT:** \[\sum_k |N_k|^3\]

**Theorem 3:** When observing a multi-state, it is optimal to select the b(i) any undominated state j of that multi-state (*with respect to terminal data*).
COMPUTING THE VALUE (INFINITE HORIZON EXPECTED UTILITY)

- An alternative to elimination is triangularization
- Once triangularized, there is no return to a state of a bandit that was visited
- So the big MDC becomes acyclic and terminates in finite time.
- This allows for computing $v(s)$ for each multistate $s$ in time that is linear in the total number of local states (for initial state or distribution)
REFERENCES


OUTLINE

• The standard (risk-neutral) Multi-Armed Bandit Problem (MABP) and its optimal index-solution
  – Model
  – Review of previous approaches
• Risk-seeking and risk-averse MABP
• Outline a simple derivation of an optimal index-solution of the three variants of MABP

Contribution:
  – Simple constructive proof of optimality of index policies for classic risk-neutral MABP
  – Extension to risk sensitive model – risk averse and risk seeking
  – Efficient algorithm
TRANSIENT AND SUBSTOCHASTIC MATRICES

- The matrices $Q$ on which we will focus are *nonnegative*, i.e., each entry is nonnegative.
- A square matrix $Q$ is said to be **transient** if $\lim_{n \to \infty} Q^n = 0$.
- A square matrix is **substochastic** if it is nonnegative and if each of its rows sums to 1 or less.

**Fact.** Let $Q$ be a nonnegative square matrix. The following are equivalent: (i) $Q$ is transient, (ii) $\rho(Q) < 1$, (iii) there exists $x > 0$ with $Qx << x$.

If $Q^\delta$ is transient

$$v^\delta = \lim_{n \to \infty} \sum_{t=0}^{n-1} (Q^\delta)^t r^\delta = (I - Q^\delta)^{-1} r^\delta$$
\[ \nu^\delta = \lim_{n \to \infty} \left[ \sum_{t=0}^{n-1} (Q^\delta)^t r^\delta - (Q)^n 1 \right] = (I - Q^\delta)^{-1} r^\delta \]

\[ \nu^\delta = \lim_{n \to \infty} \left[ \sum_{t=0}^{n-1} (Q^\delta)^t r^\delta + (Q)^n 1 \right] = (I - Q^\delta)^{-1} r^\delta \]
TERMINOLOGY

- Risk Neutral: linear utility (expected profit)
- Risk Averse: $u(x) = -\exp\{-\lambda x\}$
- Risk Seeking: $u(x) = \exp\{\lambda x\}$
- Risk sensitive
- Multi-Armed Bandits: A sequential decision problem
- Index Policies/Priority Policies: A class of structured decision rules