Restless Bandit Index Policies for Solving Constrained Sequential Estimation Problems

Sofía S. Villar

Postdoctoral Fellow
Basque Center for Applied Mathematics (BCAM)

Lancaster University
November 5th, 2012
The objective of the talk

Multi-armed Restless Bandit Problems (OR)

Optimal Sequential Estimation Problems (Statistics)

⇒ Novel Solution to existing Applied Problems in Modern Sensor Systems
Problem 1: Multi-target tracking (I)

A set of moving targets to be tracked during a period of time:
A set of flexible (and error-prone) sensors to track them
P1: How to best estimate the position of all targets over time by selecting the beams’ direction to produce sensors’ measurements?
Problem 2: Finding elusive targets

An error prone sensor must find an object which hides/exposes itself when sensed/not sensed.

P2: How should we select when to activate the sensor to maximize the chances of finding the object?
Outline

Optimal Sequential Estimation Problems

Multi-armed Restless Bandit Problems

Some Results

Further Work
Let the position of some target $n$ at time $t$, be a random variable $x_{n,t}$ evolving over time:

$$x_{n,t} = x_{n,t-1} + \omega_{n,t}, \quad \omega_{n,t} \sim N(0, q_n) \quad t \geq 1 \quad (1)$$

Let the measurement of target’s $n$ position at time $t$: $y_{n,t}$, be such that

$$y_{n,t} = x_{n,t} + \nu_{n,t}, \quad \nu_{n,t} \sim N(0, r_n) \quad (2)$$

$$\mathbb{E}(x_{n,0}) = \mu \text{ and } \mathbb{V}(x_{n,0}) = 0.$$ 

Q: From the observed sequence of measurements $\rightarrow$ produce estimates of unobservable variables, i.e., $y_{n,t} \Rightarrow \hat{x}_{n,t}$? 

Can these estimates be more precise than those based on a single measurement alone?
The Kalman Filter Algorithm

Let \( v_{n,t} \) be the variance of the sequential estimator \( \hat{x}_{n,t} \)

A priori estimate of \( x_{n,t} \) from (1) (Predictive Step)

\[
(\hat{x}_{n,t})^P = \hat{x}_{n,t-1} \tag{3}
\]
\[
(v_{n,t})^P = v_{n,t-1} + q_n. \tag{4}
\]

A posteriori estimate of \( x_{n,t} \) (Updating Step)

\[
(\hat{x}_{n,t})^U = (1 - k^*)\hat{x}_{n,t-1} + k^* y_{n,t} \tag{5}
\]
\[
(v_{n,t})^U = \frac{v_{n,t-1} + q_n}{v_{n,t-1}/r_n + q_n/r_n + 1}. \tag{6}
\]

Where \( k^* \triangleq \frac{(v_{n,t-1})^P}{(v_{n,t-1})^P + r} \) is the optimal Kalman Gain, and it is optimal in the sense that it minimizes \( (v_{n,t})^U \)
• The Kalman Filter algorithm is defined for models more general than equations (1) and (2).

• Widely used for time series analysis, in fields such as signal processing and econometrics.

• Theoretical properties: minimum Mean Squared Error (MSE) estimators.

**Key Issue:** If there are $N$ targets and $N$ sensors → the Kalman filter yields the minimum MSE targets’ locations estimators at every $t$.

If sensors fall below $N$, at each time for some targets there will be a prediction without update, therefore resulting in a higher total predictive error.

**Room for optimizing while estimating!**
Outline

Optimal Sequential Estimation Problems

Multi-armed Restless Bandit Problems

Some Results

Further Work
Markov Decision Problems (MDPs) are a natural way of formulating optimal sequential estimation problems.

**State space:** \( X_t \in X \longrightarrow \) the sufficient statistic for the \( N \) targets’ state upon which to make decisions at time \( t \).

**Action Set:** \( a_t \in A_t(X_t) \longrightarrow \) the feasible control set applicable to the system at time \( t \). If binary \( a_t \in \{0, 1\} \longrightarrow \) Bandit

**Transition Dynamics:** \( X_{t+1} \sim \mathbb{P}(\cdot | X_t, a_t) \) A probability measure following from the corresponding Filter algorithm.

**One-period net rewards/costs:** \( R(X_t, a_t) / C(X_t, a_t) \) representing the effect in terms of the objective of the system of the selected actions.

**Multi-armed Bandit Problems** are a special class of constrained MDPs.
Multitarget tracking and Multi-armed Bandit Problems

Elements

- **Arms:** $N$ (Moving targets)

- **Action Set:** $a_{n,t} \in \{0, 1\}$
  with $a_{n,t} = 1$ if target $n$ is observed in $t$ and 0 otherwise.

- **State space:** $v_{n,t} \in V_n \triangleq [0, \infty)$

- **One-period Dynamics:**
  
  $$
  v_{n,t} = \begin{cases} 
  \phi_n(v_{n,t-1}, 0) & \triangleq v_{n,t-1} + q_n, \\
  \phi_n(v_{n,t-1}, 1) & \triangleq \frac{v_{n,t-1} + q_n}{v_{n,t-1}/r_n + q_n/r_n + 1},
  \end{cases}
  $$
  if $a_{n,t} = 0$

  or

  $$
  \phi_n(v_{n,t-1}, 0) = v_{n,t-1} + q_n,
  \phi_n(v_{n,t-1}, 1) = \frac{v_{n,t-1} + q_n}{v_{n,t-1}/r_n + q_n/r_n + 1},
  $$
  if $a_{n,t} = 1$,

- **Time periods:** $t = 0, 1, \ldots$ (Infinite horizon)

- **One-period costs:**
  
  $$
  C_n(v_{n,t}, a_{n,t}) \triangleq d_n\phi_n(v_{n,t}, a_{n,t}) + h_n a_{n,t}
  $$

  $h_n \geq 0$: *measurement cost* of measuring target $n$

  $d_n > 0$: relative importance of tracking precision of target $n$. 

Optimal Sequential Estimation
A Constrained MDP’s optimal policy

- The estimation rule yielding the total minimum (discounted) variance solves MDP (7)

$$V^*(V_0) \triangleq \min_{\pi \in \Pi(M)} \mathbb{E}_{V_0}^{\pi} \left[ \sum_{n=1}^{N} \sum_{t=0}^{\infty} \beta^t C^{(n)}(v_t, a_t) \right]$$

(7)

$\mathbb{E}_{V_0}^{\pi}[.]$ denotes expectation under $\pi$ conditioned on $V_0$.

$\Pi(M)$ contains all the possible solutions that observe at most $M$ targets at each $t$.

$\beta$ is a discount factor in $[0, 1)$.

- The solution to (7) $\pi^*$ is stationary deterministic policy solving the traditional dynamic programming (DP) equations

$$V^*(V_0) \triangleq \min_{a \in A(V_0)} \left\{ C(V_0, a) + \beta \mathbb{E}_{V_1} V^*(V_1) \right\} , V_0, V_1 \in [0, \infty)^N$$
(1) The **cardinality** of the state space $\mathbf{V}$ determines the computation and storage requirements of the DP approach.

$\mathbf{V} \subseteq \mathbb{R}^N \rightarrow$ infinite size. Existence of closed-form solutions.

If $|V_n| = k \rightarrow$ its size is $\approx k^N$. Worse for infinite horizon.

= Practical implementation of $\pi^*$ for realistic scenarios is hindered by the curse of dimensionality.

(2) Mathematically based Heuristic solution Methodology $\rightarrow$

Design **tractable** and **near-optimal** priority-index policies.
Bandit Indexation
A Mathematically based Methodology

Gittins ’74  First showed that the optimal solution to the **Classic** version (Bandits’ state remain frozen if not activated) admitted an expression that breaks the curse of dimensionality.

For each arm, an index function (depending on its current state) exists and recovers its optimal policy (**Indexability**).

The index function assigns a **priority** value (**index**) to each arm to being activated, depending on its state.

**Heuristic Index Policy**  Allocate available resources at each $t$ to activate $M$ arms with highest priority values.

Whittle 88’  Proposed a relaxation and a Lagrangian approach for the **Restless** case (Bandits’ state evolve even if not activated)

Index functions were not ensured to always exist.

The resulting heuristic was not guaranteed to be **optimal**.
Solving Sequential Estimation Problems as MDPs
Research Challenges

- Multi-armed Bandit reformulations of the resulting MDPs are restless $\rightarrow$ Predicting and Updating $\neq$ Predicting

Q: Index existence? Index computation? Tractable?

- Niño-Mora 01: Discrete-State Sufficient Indexability Conditions (SIC). If SIC hold $\rightarrow$ Index exists and it is computable through a tractable algorithm.

  Continuous state space $\rightarrow$ Extended SIC for infinite state?

- Optimal sequential estimation models $\rightarrow$ complex nonlinear dynamics for the state variable (Möbius Transformations)

Q: How to address these complications to establish extended SIC?

- If Whittle heuristic exists, it is not necessarily optimal.

Q: Assessing Suboptimality Gap?
Outline

Optimal Sequential Estimation Problems

Multi-armed Restless Bandit Problems

Some Results

Further Work
Whittle Index for Multitarget Tracking

The marginal profit of pulling an arm

**Definition:** Single-arm problem is indexable if \( \exists \) an index function \( \lambda^*(\nu) \) such that \( \forall \lambda \in \mathbb{R} \) and state \( \forall \nu \in V^{0,1} \), it is optimal to observe the target in state \( \nu \) iff \( \lambda^*(\nu) \geq \lambda \).

**One-arm problem:**

\[
V^*(\nu) \triangleq \min_{\pi \in \Pi} \mathbb{E}_\nu^\pi \left[ \sum_{t=0}^{\infty} \beta^t d_n \phi_n(\nu_n, a_n, t) + h_n a_n, t \right],
\]

(8)

SIC for the single-target problem hold \( \rightarrow \) Niño-Mora and Villar (2009) the single-arm tracking problem is indexable under the \( \beta \)-discounted criterion.

The Whittle index can be computed as follows:

\[
\lambda^*(\nu) = d_n \left( \phi_n(\nu_n, 0) - \phi_n(\nu_n, 1) \right) - h_n + \beta \left( V^*(\phi(\nu, 0) - V^*(\phi(\nu, 1)) \right)
\]
The Whittle index for different discount factors $\beta$ in a target instance of $r = 1$, and $q = 5$.
Whittle Index Policy: Benchmarking results

Base instance:
\( \beta = 0.99, \ T = 10^4, \)  
\( M = 1, \ N = 4, \)
where \( \forall n \in N, \)  
\( q_n \equiv 10, \ r_n \equiv 1, \)  
\( d_n \equiv 1, \ h_n \equiv 0, \) and  
\( v_{n,0} = 0. \)

Modify it by letting:  
\( q_1 \in \{0.5, 1, \ldots, 10\}. \)

Myopic policy \((\beta = 0)\)  
TEV policy, index \( \nu_t \)  
Whittle policy, index \( \lambda^*(\nu) \)  
Lower Bound on \( V^*_D(\nu) \)
Outline

Optimal Sequential Estimation Problems

Multi-armed Restless Bandit Problems

Some Results

Further Work
Further Work
Current and Future Work

• Other **Applied Problems.** Some of them:

  **Sensor Networks** Search of a reactive (elusive) target.
  **Wireless Cellular Systems** Allocation of jobs under partial observability of channel conditions.
  **Clinical Trials** Design of Multi-arm multi-stage trials.

• **Methodological:** Optimality Conditions for the Whittle Heuristic.

• **Computational:** Use properties of Möbius transformations to improve on index computation method (by truncation).
Thanks for the attention! 😊
Thanks for the attention! 😊


