Surveillance in an Abruptly Changing World via Multiarmed Bandits

Vaibhav Srivastava
Department of Mechanical & Aerospace Engineering
Princeton University

December 15, 2014

Joint work with: Paul Reverdy and Naomi Leonard

IEEE Conference on Decision and Control
Los Angeles, CA

Incomplete Literature Review

Environmental Monitoring and Surveillance

Multi-armed Bandit Problems

Stochastic Multi-armed Bandits

- $N$ options with unknown mean rewards $m_i$
- the obtained reward is corrupted by noise
- distribution of noise is known $\sim \mathcal{N}(0, \sigma^2)$
- can play only one option at a time

Objective: maximize expected cumulative reward until time $T$
Spatially Embedded Gaussian Multi-armed Bandits

- reward at option $i \sim \mathcal{N}(m_i, \sigma_i^2)$
- prior on rewards $\mathbf{m} \sim \mathcal{N}(\mathbf{\mu}_0, \Sigma_0)$
- spatial structure captured through $\Sigma_0$, e.g., $\sigma_{ij}^0 = \sigma_0 \exp(-d_{ij}/\lambda)$
- value of option $i$ at time $t$: $Q_i^t = \mu_i^t + \sigma_i^t \Phi^{-1}(1 - \frac{1}{Kt})$

- Inference Algorithm:
  \[ \Lambda_t \mu_t = r_t \phi_t / \sigma_s^2 + \Lambda_{t-1} \mu_{t-1} \]
  \[ \Lambda_t = \phi_t \phi_t^T / \sigma_s^2 + \Lambda_{t-1}, \quad \Sigma_t = \Lambda_t^{-1}, \]

-Spatially Embedded Gaussian Multi-armed Bandits

- the mean rewards switches to unknown values at unknown times
- the switched rewards may have the same correlation scale
- the number of switches until time $T$ is upper bounded by $\zeta_T$

-Sliding-window UCL algorithm

- estimate mean using observations at times $\{(t - t_w)^+ + 1, \ldots, t\}$
- selects the arm $i$ with the maximum value of
  \[ Q_i^{t,t_w} := \mu_i^{t,t_w} + \sigma_i^{t,t_w} \Phi^{-1}(1 - \frac{1}{K \min\{t_w, t\}}), \]
- an adaptation of the frequentist algorithm by Garivier and Moulines

---

Stochastic Multi-armed Bandits

- $N$ options with unknown mean rewards $m_i$
- the obtained reward is corrupted by noise
- distribution of noise is known $\sim \mathcal{N}(0, \sigma^2)$
- can play only one option at a time

Objective: maximize expected cumulative reward until time $T$

Equivalently: Minimize the cumulative regret

Cum. Regret = $\sum_{t=1}^{T} (m_{\text{max}} - m_i)$

$m_{\text{max}}$ = max reward $i_t = \text{arm picked at time } t$

Prototypical example of exploration-exploitation trade-off

---

Gaussian multiarmed bandits with abrupt changes

Sliding-Window Approach: Description

- the mean rewards switches to unknown values at unknown times
- the switched rewards may have the same correlation scale
- the number of switches until time $T$ is upper bounded by $\zeta_T$

Sliding-window UCL algorithm

- estimate mean using observations at times $\{(t - t_w)^+ + 1, \ldots, t\}$
- selects the arm $i$ with the maximum value of
  \[ Q_i^{t,t_w} := \mu_i^{t,t_w} + \sigma_i^{t,t_w} \Phi^{-1}(1 - \frac{1}{K \min\{t_w, t\}}), \]
- an adaptation of the frequentist algorithm by Garivier and Moulines
Gaussian multiarmed bandits with abrupt changes

Block Allocation Strategy: Description of algorithm

Block SW-UCL achieves same order of performance as SW-UCL

Analysis of Sliding-Window UCL algorithm
- For $\zeta_T = O(T^{\nu})$, $\nu \in [0, 1)$ and $t_w = \left\lceil \sqrt{T \log T} \right\rceil$
  \[
  \mathbb{E}[n^T_i] \leq O\left( T^{\frac{1+\nu}{2}} \sqrt{\log T} \right);
  \]
- For $\zeta_T \leq \lambda T$, for some $\lambda \in [0, 1)$, and $t_w = \left\lceil \sqrt{-\log \lambda} \right\rceil$
  \[
  \mathbb{E}[n^T_i] \leq O\left( T \sqrt{-\log \lambda} \right).
  \]

Gaussian multiarmed bandits with abrupt changes

Block Allocation Strategy: Description of algorithm

- Block allocation to reduce travel cost
- Divide sampling times into frames $\{1, \ldots, L+1\}$
- $L$-th frame ends at $2^k w$, $k_w$ equivalent of width of time-window
- $k$-th frame subdivided in blocks on length $k \in \{1, \ldots, L\}$
- $(L+1)$-th frame contains times $\{2^{k+1} w, \ldots, T\}$
- $(L+1)$-th frame subdivided in blocks on length $k_w$

Block SLiding-Window UCL algorithm

At beginning of $r$-th block in $k$-th frame, i.e., at time $\tau_{kr}$
- Performs the estimation using the observations collected in the time-window $\{\tau_{kr} - 2^k w + 1, \ldots, \tau_{kr}\}$
- Selects the arm $i$ with the maximum value of
  \[
  Q^{\tau_{kr}, kw}_i := \mu_i^{\tau_{kr}, kw} + \sigma_i^{\tau_{kr}, kw} \phi^{-1}\left(1 - 1/K \min\{2^k w, \tau_{kr}\}\right),
  \]
  for the duration of the block

Block SW-UCL achieves same order of performance as SW-UCL

Gaussian multiarmed bandits with abrupt changes

Numerical Illustration
- Environment: $5 \times 5$ square grid
- Reward at optimal $m^* = 10$
- Reward at other arms $m_j = m_i \exp(-0.3 d_{ij})$, $d_{ij} =$ distance
- Assumed correlation scale $\rho_{ij} = \exp(-0.3 d_{ij})$
- $\sigma^2 = 1$ and $\sigma^2 = 10$
- Number of changes $\zeta_T = \lceil \sqrt{T} \rceil$

Expt number of selections of suboptimal arms
- Black line: SWUCL
- Red line: Adaptive SWUCL
- Green line: Block SWUCL

Expt number of transitions among arms

Frame structure
- Blocks
- Last Frame
- Blocks Last Frame
- Frame structure
How important is the correlation scale

Conclusions and Future Directions

Conclusions
- A multiarmed bandit framework for surveillance problems
- Arrival on events of interest \( \Rightarrow \) Abrupt changes in reward surface
- Exploration-Exploitation trade-off and role of correlation scale
- Block allocation to reduce travel cost

Future Directions
- Extension to multiple vehicles
- Environmental partitioning strategies catered to addressing exploration-exploitation trade-off
- Extensions to continuously changing environments