The Stochastic Multi-Armed Bandit Problem: In Neuroscience, Ecology, and Engineering

Vaibhav Srivastava
CYber Physical Human systEmS Research (CYPHER) Lab
Department of Electrical & Computer Engineering
Michigan State University
http://www.egr.msu.edu/~vaibhav

April 23, 2017
6th Midwest Workshop on Control and Game Theory
University of Michigan

Stochastic Multi-armed Bandit Problems

- 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 \( T \)

Multi-armed Bandit Problem

Mersereau, Rusmevichientong, and Tsitsiklis. A structured MAB problem and the greedy policy. IEEE TAC, 2009

Human Decision-Making in Multi-armed Bandit Tasks

Cohen et al. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. 2007

Outline

- Stochastic Multi-armed Bandit Problems
- Modeling Human Performance in Multi-armed Bandit Tasks
  - Features of human decision-making
  - Upper Credible Limit algorithm
  - Data from experiments with human participants
- Animal Foraging and Multi-armed Bandit Problems
- Distributed Decision-Making in Multi-armed Bandit Problems
- Conclusions and Future Directions

Literature

Multi-armed Bandit Problem

Mersereau, Rusmevichientong, and Tsitsiklis. A structured MAB problem and the greedy policy. IEEE TAC, 2009

Human Decision-Making in Multi-armed Bandit Tasks

Cohen et al. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. 2007


Stochastic Multi-armed Bandit Problems

- **N** options with unknown mean rewards \( m_i \)
- The obtained reward is corrupted by noise
- Distribution of noise is known \( \sim N(0, \sigma_i^2) \)
- Can play only one option at a time

**Objective:** Maximize expected cumulative reward until time \( T \)

**Equivalently:** Minimize the cumulative regret

\[
\text{Cumulative Regret} = \sum_{t=1}^{T} (m_{\text{max}} - m_i).
\]

\( m_{\text{max}} = \text{max mean reward} \quad i_t = \text{arm picked at time } t \)

Best possible performance and State of the art

**Lai-Robbins Bound**
- Cumulative Regret \( \geq K_{\text{min}} \log T \), \( T = \text{Horizon length} \)

**Upper confidence bound algorithm (Auer et al.'00)**
- Play each option once, then at each time \( t \) pick arm

\[
\arg\max_i \tilde{m}_i + C_i^t \quad \text{frequentist estimator uncertainty measure}
\]

- Cumulative Regret \( \leq K_{\text{ucb}} \log T \) for bounded rewards

**Bayesian UCB algorithm (Kauffman et al.'12)**
- At each time \( t \) pick arm with maximum \( \left( 1 + \frac{1}{T} \right) \)-upper confidence bound

- Cumulative Regret \( \leq K_{\text{min}} \log T \) for Bernoulli rewards

Outline

- **Stochastic Multi-armed Bandit Problems**
- **Modeling Human Performance in Multi-armed Bandit Tasks**
  - Features of human decision-making
  - Upper Credible Limit algorithm
  - Data from experiments with human participants
- **Animal Foraging and Multi-armed Bandit Problems**
- **Distributed Decision-Making in Multi-armed Bandit Problems**
- **Conclusions and Future Directions**

Features of Human Decision Making in MAB Tasks

- **Cohen, McClure & Yu ’07**
  - Familiarity with the environment

- **Wilson et al. ’11, ’14, Cohen et al. ’07**
  - Information bonus:

\[
\frac{Q_i}{\text{value}} = \frac{\Delta R_i}{\text{reward gain}} + \frac{A}{\text{info bonus}} \cdot \frac{\Delta I_i}{\text{info gain}}
\]

- Decision noise: Directed v/s random exploration
- Information bonus and noise are sensitive to task horizon

- **Acuta and Schrater ’10**
  - Humans learn environment correlation structure

Specifications of major HAUV components.

Table 1.

<table>
<thead>
<tr>
<th>Component</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camera</td>
<td>520 nm (green) LED, also provides four range beams</td>
</tr>
<tr>
<td>Depth sensor</td>
<td>Sound Metrics 1.8 MHz DIDSON</td>
</tr>
<tr>
<td>IMU sensor</td>
<td>Keller pressure</td>
</tr>
<tr>
<td>Thrusters</td>
<td>6, rotor-wound</td>
</tr>
<tr>
<td>Processor</td>
<td>Doppler velocity RDI 1200 kHz Workhorse;</td>
</tr>
<tr>
<td>Optional tether</td>
<td>150 m long, 5 mm diameter</td>
</tr>
<tr>
<td>Camera footprint</td>
<td>thrusters, comprised of 1m x 0.45 m x 0.1m</td>
</tr>
<tr>
<td>Sonar footprint</td>
<td>150 m x 50 m x 10 m</td>
</tr>
<tr>
<td>Ship-hull inspection</td>
<td>Hover et al. '12</td>
</tr>
</tbody>
</table>

Data from experiments with human participants.

- Features of human decision-making
- Upper Credible Limit algorithm
- Data from experiments with human participants
Spatially Embedded Gaussian Multi-armed Bandits

- reward at option $i \sim N(m_i, \sigma_i^2)$
- prior on reward surface $m \sim N(\mu_0, \Sigma_0)$
- spatial structure captured through $\Sigma_0$, e.g., $\sigma_0^0 = \sigma_0 \exp(-d_{ij}/\lambda)$

Inference Algorithm: Kalman Filter
- $\phi_t = \text{indicator vector of the arm selected at time } t$
- $r_t = \text{reward obtained at time } t$

Posterior Precision: $\Lambda_t = \phi_t \phi_t^T + \Lambda_{t-1}$, $\Sigma_t = \Lambda_t^{-1}$

Posterior Mean: $\mu_t = r_t \phi_t + \Lambda_{t-1} \mu_{t-1}$


The UCL Algorithm

Upper Credible Limit (UCL) Algorithm
- value of option $i$ at time $t = (1 - \frac{1}{K_t})$-upper credible limit:
  $$Q^i_t = \mu^i_t \Phi^{-1}(1 - \frac{1}{K_t})$$

- $\mu^i_t = \text{posterior mean}$
- $(\sigma^i_t)^2 = \text{posterior variance}$
- pick option with maximum value $Q^i_t$ at each time

- for uninformative priors:
  $$\text{Cumulative Regret} \leq K_{\text{wul}} \log T + o(\log T)$$

Stochastic UCL Algorithm and Human Decision-Making

Stochastic Upper Credible Limit (UCL) Algorithm:
- pick option $i$ with probability $\propto \exp(Q^i_t/v_t)$, $v_t = \nu/\log(t)$
- similar performance can be established

Does stochastic UCL algorithm explain human experiment data?
Bandit Experiment with Human Subjects: What Accounts for Expertise?

- Data from Amazon Mechanical Turk bandit experiment
- 10 × 10 spatial grid of N = 100 options
- Given T = 90 trials: insufficient time to explore whole space
- Global mean reward ≈ 30, Maximum mean reward = 60

Stochastic UCL as a model for human subjects

- Prior: \( \mathcal{N}(\mu_0 1_N, \Sigma_0) \), \( (\Sigma_0)_{ij} = \sigma_0^2 \exp(-d_{ij}/\lambda) \)
- Model parameters: \( (\mu_0, \lambda, \sigma_0, \nu) \)

Outline

- Stochastic Multi-armed Bandit Problems
- Modeling Human Performance in Multi-armed Bandit Tasks
  - Features of human decision-making
  - Upper Credible Limit algorithm
  - Data from experiments with human participants
- Animal Foraging and Multi-armed Bandits
- Distributed Decision-Making in Multi-armed Bandit Problems
- Conclusions and Future Directions

Animal Foraging and Multiarmed Bandits

- decision mechanisms underlying such search models?

Optimal foraging theory

- patch to visit?
- patch residence time?
- maximize benefit rate
- foraging path in patch?

Spatial Multi-armed Bandits

- arm to select?
- duration at arm?
- minimize transitions
- path within arm?


MAB Problem with Transition Costs

- reward associated with arm $i$: $N(m_i, \sigma_i^2)$
- each transition from arm $i$ to arm $j$ costs $c_{ij}$

Block allocation strategy:
- at allocation round $(k, r)$

$$Q_i^{kr} = m_i^{kr} + \frac{\sigma_i}{\sqrt{n_i^{kr}}} \Phi^{-1}(1 - \frac{1}{K_{tr}})$$

- pick arm $i^* = \arg\max Q_i^{kr}$ for the next block

Cumulative Regret $\leq K_{buc}\log T + o(\log T)$

$\mathbb{E}[\# \text{ transitions to arm } i] \leq K_{tran}\log T + o(\log T)$

Satisficing in the Mean Reward

$(\mathcal{M}, \delta)$ – satisficing multiarmed bandit problem

- satisfaction in mean reward at time $t$ by the variable $s_t$, defined as

$$s_t = \begin{cases} 1, & \text{if } m_i \geq \mathcal{M} = \text{satisfaction threshold} \\ 0, & \text{otherwise.} \end{cases}$$

- arm $i$ is $\delta$-sufficing in mean reward if

$$\mathbb{P}[S_t = 1] \geq (1 - \delta) = \text{sufficiency threshold}$$

- expected satisficing regret at time $t$

$$R_t = \begin{cases} \Delta^M_t, & \text{if } \mathbb{P}[S_t = 1] \leq 1 - \delta, \\ 0, & \text{otherwise.} \end{cases}$$

$$\Delta^M = \max \{0, \mathcal{M} - m_i\}$$

The Distributed MAB Problem

- $N$ options with unknown mean rewards $m_i$
- $M$ decision-making agents with a connected communication graph
- Each agent can play only one option at a time
- Rewards corrupted by Gaussian noise $N(0, \sigma_i^2)$
- No interference/collisions among agents

Goal: Distributed algorithms that maximize total expected cumulative reward

Anantharam et al. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. Trans on Auto Ctrl, 1987
Anandikumar et al. Distributed algorithms for learning and cognitive medium access with logarithmic regret. J. Sel. Areas Comm., 2011
Kar, Poor, & Cui. Bandit problems in networks: Asymptotically efficient distributed allocation rules. Conf. on Decision & Cont., 2011
Shahrampour, Rakhlin, and Jadabaie. Multi-Armed Bandits in Multi-Agent Networks. ICASSP, 2017
### Distributed Estimation of Mean Rewards via Running Consensus

- $\hat{n}_i(t)$: $k$-th agent’s estimate of # selections of arm $i$
- $\hat{s}_i(t)$: estimate of average sum of reward from arm $i$
- Update estimates via running consensus

For agent $k$ at time $t$: $\phi^k(t) = \text{indicator vector of chosen arm}$, $r^k(t) =$ reward

$$\hat{n}_i(t+1) = \sum_{j \in \text{nbhd}(k)} \rho_{ij}(\hat{n}_j(t) + \phi^j(t))$$

$$\hat{s}_i(t+1) = \sum_{j \in \text{nbhd}(k)} \rho_{ij}(\hat{s}_j(t) + r^i(t)\phi^j(t))$$

Estimate of mean reward from arm $i$: $\hat{\mu}_i(t) = \frac{\hat{s}_i(t)}{\hat{n}_i(t)}$

---

**Outline**

- Stochastic Multi-armed Bandit Problems
- Modeling Human Performance in Multi-armed Bandit Tasks
  - Features of human decision-making
  - Upper Credible Limit algorithm
  - Data from experiments with human participants
- Animal Foraging and Multi-armed Bandit Problems
- Distributed Decision-Making in Multi-armed Bandit Problems
- Conclusions and Future Directions

---

**Conclusions and Future Directions**

**Conclusions**

- UCL algorithm for the correlated MAB problem
- Stochastic UCL algorithm as a model for human decision-making
- Human performance depends critically on assumptions on correlation scale
- Satisficing as a model of bounded rationality
- Cooperative UCB for collective decision-making in MAB problems
- A metric to order nodes in terms of their performance

**Future Directions**

- Reducing the communication burden
- Interference in reward among agents and strategic decision-making
- Analysis of social decision-making data
Thanks for your attention!

Questions?