Multi-Armed Bandit Strategies for Non-Stationary Reward Distributions and Delayed Feedback Processes

Larkin Liu

Loblaw Digital
AISC Original Author Series Presentation

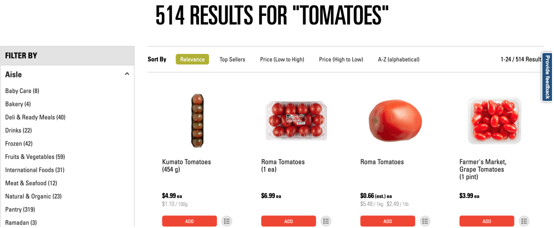
July 29, 2019

Overview

1. Online Grocery Ordering System
2. eCommerce Grocery Business Case
3. ϵ-Greedy Strategy
4. UCB-1 Strategy
5. Thompson Sampling
6. Non-Stationary Reward Functions
7. Delayed Feedback Scenario
8. Adaptive Greedy Strategy
9. Experimental Simulation Results

Online Grocery Ordering System

- Customers have the capability to order online and pick it up at the nearest store of their choice.
- Ideally, items searched for on the website should be in stock at the time of pickup.
- Multiple algorithms run to ensure that the items shown to customers online are available at time of pick-up. (Such models utilize ARIMA models to forecast demand, and historical averages to forecast available inventory).



eCommerce Grocery Business Case

- **The assortment problem:** Ensuring that as many items are available at the time of pick up as possible. The two main factors that affect it are demand and replenishment.
- **Fill Rate:** The percentage of online items sold that were available at the time of pick-up. We wish to select the model that provides the highest fill rate.

AB Testing

AB Testing: Two models are run in parallel to determine which one performs the best based on some metric. Ideally, the best algorithm is found with minimal tests spent on exploration.

Multi-Armed Bandit Strategy

Reward Function

Provided a policy π, the expected reward, V_t from taking action k_t can be expressed as,

$$V_t(\pi) = \sum^{k} Q(k_t) P(k_t|\pi) \tag{1}$$

The objective of the Multi-Armed Bandit Strategy is to minimize the expected regret, defined as,

$$R_T(\pi) = \sum_{t=0}^{T} \left[V_t(\pi^*) - V_t(\pi) \right] \tag{2}$$

We denote the optimal value at time t as V_t^*, and under the optimal policy π^* by selecting the optimal arm from K arms.

ϵ-Greedy Strategy

Algorithm 1 ϵ-Greedy

$Q = \emptyset$
for $t = 0 \rightarrow T$ **do**
  for $k = 1 \rightarrow K$ **do**
    Compute $\widehat{\mu_t^k}$ from Ω
  end for
  $k = \text{argmax}_k \widehat{\mu_t^k}$
  Play k with probability $1 - \epsilon$, else play another arm with probability p_K
  $Q_k \leftarrow Q(k)$
end for

Where p_k is,

$$p_K = \frac{\epsilon}{K - 1} \tag{3}$$

- The lower bound on the expected regret of ϵ-greedy is proportional to T linearly in the infinite time horizon.

UCB-1 Strategy

Algorithm 2 UCB1 Strategy

$Q = \emptyset$
for $t = 0 \to T$ **do**
    for $k = 1 \to K$ **do**
        Compute $\widehat{\mu_t^k}$
    end for
    Play $k_t = \text{argmax}_k \, m_t^k$
    $Q \leftarrow Q(k)$
end for

We seek to maximize m_t^k where,

$$m_t^k = \mu_t^k + \sqrt{\frac{2 \log t}{n(k)}}$$

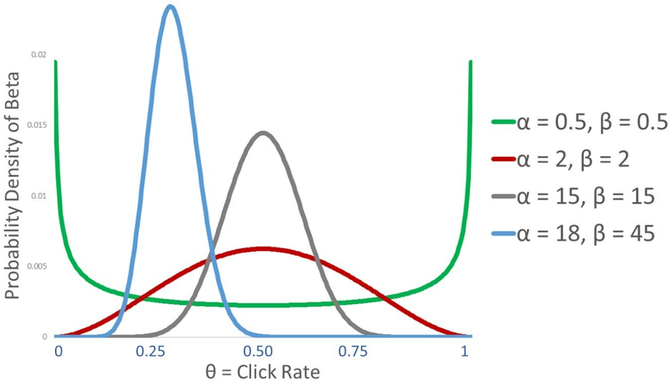
UCB1 Regret Bound (Auer, 2002)

$$\mathbb{E}[R_T(\pi)] \geq 8 \sum_{k:\mu_t^k < \mu_t^*} \left(\frac{\log n(k)}{\mu_t^k - \mu_t^*} \right) + \left(1 + \frac{\pi^2}{3} \right) \left(\sum_{k=1}^{K} \mu_t^k - \mu_t^* \right) \tag{5}$$

Thompson Sampling

- (Thompson, 1933) introduced a Bayesian method framework for implementing MAB strategies.

$$P(k) = \int \mathbb{I}\left[\mathbb{E}[Q(k)] = \max_{K} Q(k)\right] P(\theta|R)d\theta \quad (6)$$



Thompson Sampling

- Recently rediscovered Thompson sampling is advantageous for its simplicity. It requires sampling from a distribution and playing each arm respective to the maximum payout from that proposed distribution,
- A Bayesian approach, parameters are estimated from a previous observation window.

Algorithm 3 Thompson Sampling

```
Q = ∅
for t = 0 → T do
    Estimate(θ)
    for k ∈ {1, ..., K} do
        Compute \widehat{μ_t^k}
        Sample Q(k_t) ∼ \widehat{θ}_k
    end for
    k_t = argmax_k Q(k_t)
    Q ← Q(k_t)
end for
```

Intermission

Intermission

Non-Stationary Reward Functions

We define the expected reward μ_t^k as the expected value of the reward from playing arm k at time t,

$$\mu_t^k = \mathbb{E}[Q(k_t)] \tag{7}$$

In the stationary, and non-stationary case, it can be expressed as,

$$\mu^k = f(\theta^k) \tag{8}$$

$$\mu_t^k = f(t, \theta^k) \tag{9}$$

Delayed Feedback Scenario

- *Delayed-feedback* occurs when the reward process does not immediately return an reward value upon playing of a selected arm, similarly noted in (Chapelle, 2011).
- We specify the number of stores as N, and the number of arms, or *assortment algorithms*, as K. Each arm is denoted as being the k^{th} arm, where $k \in \{1, .., K\}$. In our simulation, we specify N = 100, and K = 10.

$$\widehat{\mu_t^k} = \frac{1}{Nt} \sum_{t \in \Omega} \sum_{n=1}^{N} \frac{1}{\gamma} \sum_{i=1}^{\gamma} 1(k, \pi) Q(k_t) \tag{10}$$

Adaptive Greedy Strategy

Algorithm 4 AG1

```
Q = ∅
for t = 1 → T do
    for k = 1 → K do
        Compute \widehat{μ_t^k} from Ω_r
    end for
    k = argmax_k \widehat{μ_t^k}
    for 1 → n_t^* do
        Play k
        Q ← Q(k)
    end for
    for k' = 1 → K do
        if k' ≠ k then
            for 1 → n_t do
                Play k'
                Q ← Q(k')
            end for
        end if
    end for
end for
```

Adaptive Greedy Strategy

- We create an adaptive greedy strategy in which the time frame to estimate parameters $\widehat{\mu_t^k}$ rely on a fixed time window. Forgetting the history of longer time epochs.
- The number of optimal arm plays for each player at time epoch t is n_t^*. With exploration parameter ϵ,

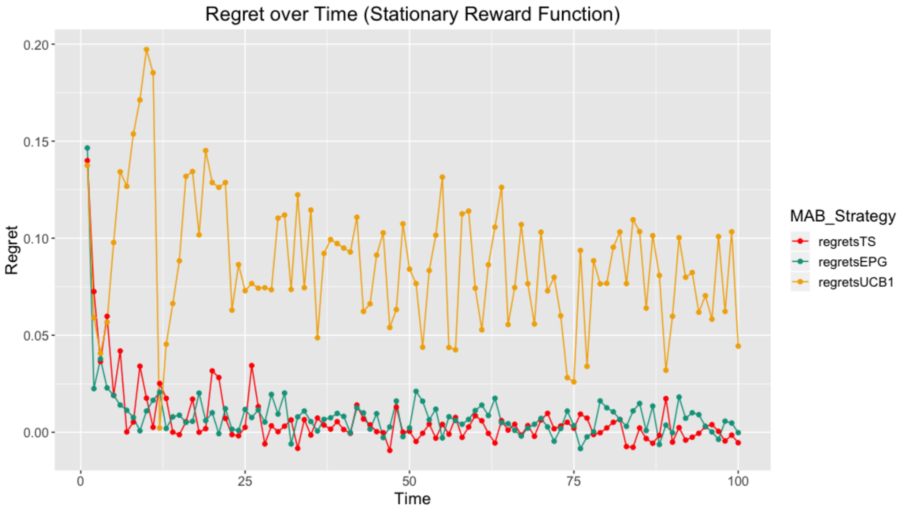
$$n_t^* = \left\lfloor N(1-\epsilon) \right\rfloor \tag{11}$$

And for non-optimal arms,

$$n_t = \left\lceil N \frac{\epsilon}{K-1} \right\rceil \tag{12}$$

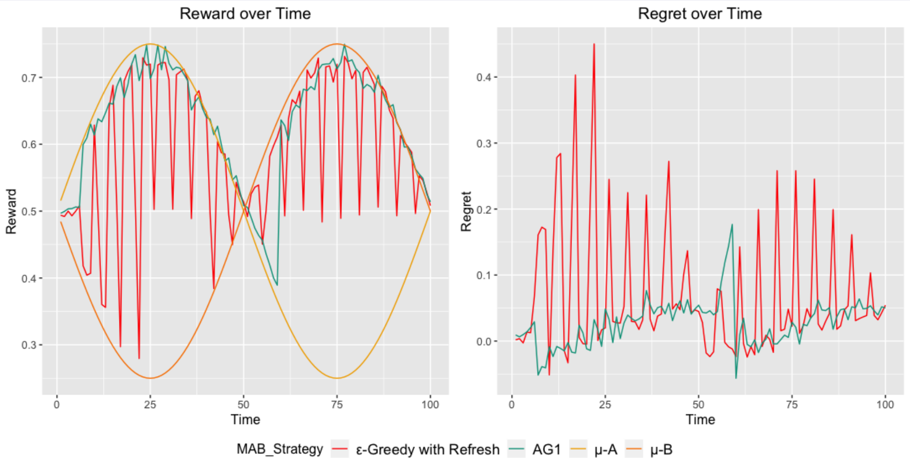
Experimental Simulation Results

Regret over time



Non-Stationary Comparison

Comparative Regret and Reward



Non-Stationary Comparison

- We observe that the cumulative regret when using AG1 is minimized when compared to traditional approaches. (K = 10)

MAB Strategy	Cumulative Regret	Cumulative Reward
ϵ^*-greedy	7.121	59.16
TS^*	6.241	59.82
AG1	2.558	63.71

For Further Reading

Auer, Peter and Cesa-Bianchi, Nicolò and Fischer, Paul
Finite-time Analysis of the Multiarmed Bandit Problem.
Machine Learning, 235–256, 2002.

Chapelle, Oliver and Li, Lihong
An Empirical Evaluation of Thompson Sampling.
Advances in Neural Information Processing Systems 24, 2249–2257, 2011.

Liu, Larkin and Downe, Richard and Reid, Josh
Multi-Armed Bandit Strategies for Non-StationaryReward Distributions and Delayed Feedback Processes.
Canadian Operational Research Society Annual Conference, 2019.

Thompson, William.
On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples
Biometrika, 285–294, 1933.