Threshold Bandit, With and Without Censored Feedback

Jacob Abernethy, Kareem Amin, Ruihao Zhu

NIPS 2016

Previous work

- A sequential decision making
- Select an arm i and receive a stochastic reward R_i^t
  - Independently and identically distribution
- Action is a function of past observed (action, reward) pairs

Previous work

- Optimal solution
  - Pull the arm with highest expected reward
  - Identical in each round
- UCB(Upper Confidence Bound) algorithm
  - Maximization over rewards estimated from previous data
  - Bias each estimation according to its uncertainty
  - Mean + confidence interval

Threshold bandit

- A piece of side information c_t
- Choose one arm $i \in [K]$, produce a value X_i^t
  - Survival function $F_i(x) \sim \Pr(X_i^t \geq x)$
- Binary reward $\mathrm{R}_i^t = \mathbb{I}[X_i^t \geq c^t]$
- Expected payoff $\mathbb{E}\left[R_i^t\right] = F_i(c^t)$
- The optimal arm is related the value of c^t
- c^t is drawn from $M = \{1,2, \dots, m\}$
  - Observe at the beginning of each round

Applications

- Packet Delivery with Deadlines
  - Ship a stream of packages from source to destination
  - Each package is supplied with a delivery deadline
  - Select a transportation route with the highest probability of on-time arrival
  - Uncertain: faster with higher volatility

Applications

- Supplier Selection
  - Produce a product satisfying specific quality demands
  - Select one of several suppliers to contract out the work
  - Uncertain: capabilities and variabilities of the products from each supplier

Applications

- Dark Pool Brokerage
  - Buy/sell various sized bundles of shares
  - Establish financial exchanges that match buyers and sellers
  - Execute the transaction if there is suitable liquidity
  - Choose the dark pool with highest probability of completion
  - Uncertain: the successful rate of transaction

Benchmark

- Compare with the best policy instead of the best arm
  - Optimal policy may incorporate with the threshold
- Regret
  - $Reg(T) = \mathbb{E}[\sum_{t=1}^{T} (\max_{i \in [n]} R_i^t - R_{I^t}^t)]$
- Previous work
  - Static optimal solution
  - E.g., adversary bandit

Feedback

- Uncensored feedback
  - Observe the sample X_i^t regardless of the threshold
  - E.g., learn the travel time of a package regardless of the deadline having been met
- Censored feedback
  - Observe a null value when exceeding the threshold
  - Observe X_i^t otherwise
  - E.g., learn the product quality when it is rejected by the customer

A New Perspective on UCB

- Potential function
  - The current number of plays of each arm
- UCB algorithm
  - Redefine the score
  - Play arm $I^t = \text{argmax}_i \, \hat{\mu}_i^t + \zeta(N_i^t, \delta)$
- The regret is bounded by potential functions
- Bound the number of pulls N_i^t following the standard techniques

Uncensored feedback

- Estimate the survival function via empirical distribution
  - $\hat{F}_i^t(j) = \frac{\sum_{\tau=1}^{t-1} \mathbb{I}[X_{I^\tau}^\tau \geq j, I^\tau = i]}{N_i^t}$
  - The number of times that arm i is selected and exceeds the threshold j
  - Converge to F_t^i at the exponential rate
- DKWUCB algorithm
  - Play arm $I^t = \text{argmax}_i \hat{F}_i^t(c^t) + \zeta(N_i^t, \delta)$

Uncensored feedback

- Redefine potential function
  - The number of pulls of each arm when it is not optimal
  - Track the accumulation of the estimation gap

- Previous analysis in UCB
  - The number of pulls of a bad arm i is $O(\frac{\log(T)}{\Delta_i^2})$
  - The regret suffered by any such pull is $O(\Delta_i)$
  - The contribution of arm i to total regret is $O(\frac{\log(T)}{\Delta_i})$

Uncensored feedback

- Current analysis
  - Bound the difference between the empirical mean estimator and the true mean of the selected arm
  - Uniform convergence rate
    - For any error ϵ, there exists N such that $|f_n(x) - f(x)| < \epsilon, \forall n > N$
    - Accuracy after enough rounds
  - Improve the bound with a constant factor

Censored feedback

- For a given threshold value, all the feedbacks from larger threshold values are useful
  - $\hat{F}_i^t(j) = \frac{\sum_{\tau=1}^{t-1} \mathbb{I}[\min\{X_{I^\tau}^\tau, c^t\} \geq j, I^\tau = i]}{N_i^t(j)}$
  - Realization and threshold exceed the threshold j
- Error bound estimator
  - Define the martingale sequence
    - The summation
    - The gap between expected reward and real reward for each arm
  - By Azuma's inequality

Martingale Sequence

- $\mathbb{E}[Y_{n+1}|Y_1, \ldots, Y_n] = Y_n$
- The conditional expected value of the next observation = the value of last observation
  - Given all the past observations
- Knowledge of past events never helps predict the mean of the future winnings
- Only the current event matters

Martingale Sequence

- Betting game
- Y_n: A gambler's cumulative fortune after n tosses of a fair coin
  - Win 1 if the coin comes up heads
  - Lose 1 if it's tails
- The gambler's conditional expected fortune after the next trial
  - Given the history
  - Equal to his present fortune

Azuma's inequality

- When the values of martingales that have bounded differences
- High probability event
- $\Pr[X_n > t] \leq \exp(\frac{-t^2}{2n})$
- Provide a bound similar in form to the Chernoff bound
  - Without assuming independence

The order of threshold values

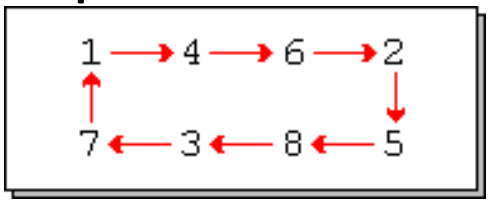
- Adversarial setting
  - Threshold values arrive in a non-decreasing order
  - $1,1 \ldots, 1,2, \ldots 2,3, \ldots, m$
  - Only the samples observed within the same stage can help to inform decision making
  - Play m independent copies of bandits
  - Reduce to stochastic MAB problems
  - Regret scales with m

The order of threshold values

- Optimistic setting
  - Threshold values arrive in a non-increasing order
  - $m, \dots, m, m-1, \dots, m-1, \dots, 1, \dots, 1$
  - Make full use of samples
  - Every sample observed in the previous rounds is useful in later rounds
  - Reduce to the problem with uncensored feedback

Cyclic Permutation Setting

- Threshold values are a cyclic permutation order of $1,2,\ldots,m$



- Divide the time horizon into epochs of length Km
- Initialize: pull each arm once for each threshold value
- Arm elimination process
  - Leverage information across threshold values
  - Utilize observations from higher thresholds to estimate mean payoffs for lower thresholds
  - Information does not flow in the other direction

Arm elimination process

- Once all but one arm has been eliminated, proceed to pull the single remaining arm for the given threshold value
- For any threshold values below j
  - The best arm has been determined
  - Play that arm constantly
- For any threshold values greater or equal to j
  - Explores all arms uniformly
- All sub-optimal arms can be detected after $O(\log T)$ epochs with high probability

Algorithm

Algorithm 1 KMUCB

Input: A set of arms $1, 2, \ldots, K$.
Initialization: $L_j \leftarrow [K]\ \forall j \in M, k \leftarrow 1, j \leftarrow 1$
for epoch $k = 1, 2, \ldots, T/Km$ **do**
  $\text{count}[j'] \leftarrow 0\ \forall j' \in M$
  for t from $(t_{k-1}+1)$ to t_k **do**
    Observe $c^t = j'$ and set $\text{count}[j'] \leftarrow \text{count}[j'] + 1$
    if $j' < j$ **then**
      $I^t \leftarrow$ index of the single arm remaining in $L_{j'}$
    else
      $I^t \leftarrow \text{count}[j']$.
    end if
  end for
  if $j \leq m$ **and** $\max_{i' \in [K]} \hat{F}_{i'}^{t_k}(j) - \hat{F}_i^{t_k}(j) \geq \sqrt{\frac{16 \log(Tk)}{(m-j+1)k}}\ \ \forall i \in L_j \setminus \{\arg\max_{i' \in [K]} \hat{F}_{i'}^{t_k}(j)\}$ **then**

$$L_j \leftarrow \left\{ \arg\max_{i' \in [K]} \hat{F}_{i'}^{t_k}(j) \right\}, \qquad j \leftarrow j+1$$

  end if
end for

Number of pulls for threshold j

Stick to the single best arm

Uniform explore
guaranteed by cyclic permutation

Estimation is accurate enough

Remove all suboptimal arms and increment threshold

Proof Sketch

- The probability that optimal arm for threshold j is eliminated
  - $O\left(\frac{1}{T}\right)$
- The number of times a suboptimal arm is pulled for threshold j
  - Under the condition that real optimal arm is not eliminated
  - Within $O(\log T)$ epochs
- Bounded by criteria in elimination

Summary

- Using potential function in analysis may be of independent interest
- The elimination process is different from the standard UCB work
- Similar to current problem
  - Only get feedback when reward does not exceed the threshold
  - A reverse direction
  - Threshold (demand) is uncertain
  - No side information

Thank you!