Censored Semi-Bandits: A Framework for Resource Allocation with Censored Feedback

Arun Verma
IIT Bombay, India
v.arun@iitb.ac.in

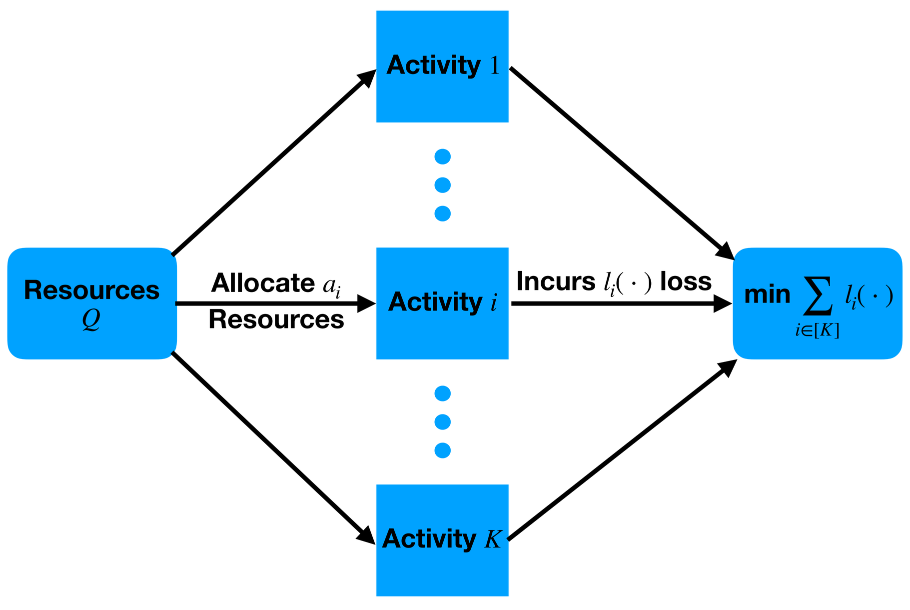
Manjesh K. Hanawal
IIT Bombay, India
mhanawal@iitb.ac.in

Arun Rajkumar
IIT Madras, India
arunr@cse.iitm.ac.in

Raman Sankaran
LinkedIn India
rsankara@linkedin.com

Resource Allocation Problem

- A fix amount of resources need to be allocated among different activities such that the total expected loss is minimized.



Resource Allocation Problem.

- Examples: advertisement budget allocation, police patrolling, supplier selection, etc.

Problem Setup

- Amount of resources: Q
- Number of arms (activities): K
- $\boldsymbol{a} := \{a_i : i \in \{1, 2, \ldots, K\}\}$, where $a_i \in [0, 1]$, denotes the resource allocated to arm i.
- Allocation $\boldsymbol{a}$ is feasible if $\sum_{i \in [K]} a_i \le Q$. The set of all feasible allocations is denoted by $\mathcal{A}_Q$.
- Expected loss observed from arm i is:

$$\mathbb{E}\left[l_i(\mu_i, \theta_i, a_i)\right] = \begin{cases} \mu_i & \text{if } a_i < \theta_i \\ 0 & \text{otherwise} \end{cases}$$

  where μ_i is the mean loss of arm i and θ_i is the associated threshold with arm i.
- **Note that both μ_i and θ_i are unknown.**
- Environment-Learner interaction in round t:

1. **Environment** generates a loss vector $\boldsymbol{X_t} = (X_{t,1}, X_{t,2}, \ldots, X_{t,K}) \in \{0,1\}^K$, where $\mathbb{E}\left[X_{t,i}\right] = \mu_i$ and sequence $(X_{t,i})_{t \ge 1}$ is i.i.d. for all $i \in [K]$ where $[K] := \{1, 2, \ldots, K\}$.
2. **Learner** picks an allocation vector $\boldsymbol{a}_t \in \mathcal{A}_Q$.
3. **Feedback and Loss:** The learner observes a random **censored** feedback $\boldsymbol{Y_t} = \{Y_{t,i} : i \in [K]\}$, where $Y_{t,i} = X_{t,i}\mathbb{1}_{\{a_{t,i} < \theta_i\}}$ and incurs loss $\sum_{i \in [K]} Y_{t,i}$.

- **Goal: Find an allocation that minimizes the total expected loss.**

Performance Measure: Regret

- Expected (pseudo) regret over a period of T:

$$\mathbb{E}\left[\mathcal{R}_T\right] = \sum_{t=1}^{T}\sum_{i=1}^{K} \mu_i \left(\mathbb{1}_{\{a_{t,i} < \theta_i\}} - \mathbb{1}_{\{a_i^\star < \theta_i\}}\right)$$

  where $\boldsymbol{a}^\star \in \arg\min_{\boldsymbol{a} \in \mathcal{A}_Q} \sum_{i=1}^{K} \mu_i \mathbb{1}_{\{a_i < \theta_i\}}$.
- A good policy should have sub-linear expected regret, i.e., $\mathbb{E}\left[\mathcal{R}_T\right]/T \to 0$ as $T \to \infty$.

Allocation Equivalent

- $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ are `allocation equivalent` iff:

$$\min_{\boldsymbol{a} \in \mathcal{A}_Q} \sum_{i=1}^{K} \mu_i \mathbb{1}_{\{a_i < \theta_i\}} = \min_{\boldsymbol{a} \in \mathcal{A}_Q} \sum_{i=1}^{K} \mu_i \mathbb{1}_{\{a_i < \hat{\theta}_i\}}.$$

CSB-ST: Same Threshold Case

- $\forall i \in [K] : \theta_i = \theta_c$ where $\theta_c \in \mathbb{R}^+$ and $Q \ge \theta_c$.
- Let $M = \min\{\lfloor Q/\theta_c \rfloor, K\}$. Then the optimal allocation allocates the θ_c fraction of resources to top M arms with highest mean loss.
- Let $\Theta = \{Q/K, Q/(K-1), \cdots, Q\}$. Then the allocation equivalent of θ_c is $\hat{\theta}_c$ where $\hat{\theta}_c \in \Theta$.
- Though $\theta_c \in \mathbb{R}^+$ but its allocation equivalent can be found in finite set Θ with high probability δ using binary search on Θ.
- Let for all $i \in [K]$, $\mu_i \ge \epsilon > 0$. Then with probability at least $1 - \delta$, the number of rounds needed to find an allocation equivalent threshold of θ_c is bounded as

$$T_{\theta_s}(\delta) \le \frac{\log(\log_2(|\Theta|)/\delta)\log_2(|\Theta|)}{\max\{1, \lfloor Q \rfloor\}\log\left(1/(1-\epsilon)\right)}.$$

- Once $\hat{\theta}_c$ is known, $\boldsymbol{\mu}$ needs to be estimated.
- Using Thomson Sampling (TS) based algorithm [1], bottom $K - M$ arms with the least mean loss are selected and no resources are allocated to them. Observe their losses and update empirical estimate of mean loss.

CSB-DT: Different Threshold Case

- Threshold may not be the same for all arms.
- Optimal allocation is the solution of 0-1 knapsack having capacity Q and K items where item i has weight θ_i and value μ_i.
- Define $\gamma := \left(Q - \sum_{a_i^\star \ge \theta_i} \theta_i\right)/K > 0$ and $\forall i \in [K] : \hat{\theta}_i \in [\theta_i, \theta_i + \gamma]$ where $\theta_i \in [0, 1]$. Then $\hat{\boldsymbol{\theta}}$ is allocation equivalent of $\boldsymbol{\theta}$.
- Each $\hat{\boldsymbol{\theta}}_i$ is estimated by using binary search in $[0, 1]$ interval.
- Let $\gamma > 0$ and for all $i \in [K]$, $\mu_i \ge \epsilon > 0$. Then with probability at least $1 - \delta$, the number of rounds needed to find an allocation equivalent of $\boldsymbol{\theta}$ is bounded as

$$T_{\theta_d}(\delta) \le \frac{K \log\left(\frac{K\log_2\left(\left\lceil 1 + \frac{1}{\gamma}\right\rceil\right)}{\delta}\right)\log_2\left(\left\lceil 1 + \frac{1}{\gamma}\right\rceil\right)}{\max\{1, \lfloor Q \rfloor\}\log\left(\frac{1}{1-\epsilon}\right)}.$$

- Once $\hat{\theta}$ is known, a subset of arms is selected using TS based algorithm [2] and no resources are allocated to them. Observe their losses and update empirical estimate of mean loss.

Regret Bounds

- Let $\mu_1 \le \mu_2 \le \ldots \le \mu_{K-M} < \mu_{K-M+1} \le \ldots \le \mu_K$, $\Delta_a = \sum_{i=1}^{K} \mu_i\left(\mathbb{1}_{\{a_i < \theta_i\}} - \mathbb{1}_{\{a_i^\star < \theta_i\}}\right)$, $\Delta_m = \max_{\boldsymbol{a} \in \mathcal{A}_Q} \Delta_a$, and $\delta = 1/T$. Then the expected regret of CSB-ST over a period of T is given by

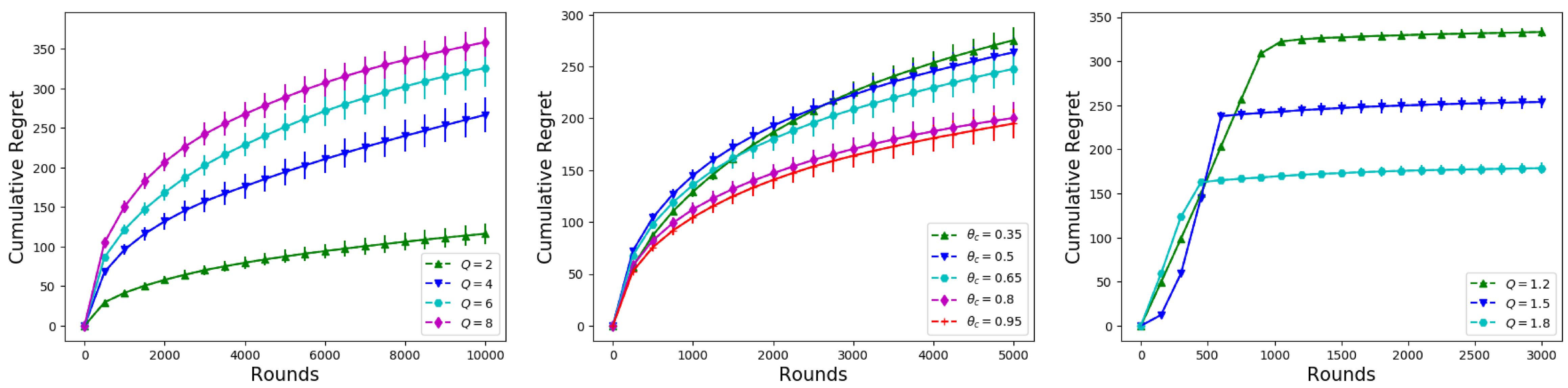
$$\mathbb{E}\left[\mathcal{R}_T\right] \le O\left(\sum_{i \in [K]\setminus[K-M]} \frac{(\mu_i - \mu_{K-M})\log T}{d(\mu_{K-M}, \mu_i)}\right).$$

- Let $\mu_1 \le \mu_2 \le \ldots \le \mu_K$, $\gamma > 0$, $S_a = \{i : a_i < \theta_i\}$ for any feasible allocation a, and $k^\star = |S_{a^\star}|$. Then for any η such that $\forall \boldsymbol{a} \in \mathcal{A}_Q, \Delta_a > 2(k^{\star 2} + 2)\eta$, the expected regret of CSB-DT over a period of T is given by

$$\mathbb{E}\left[\mathcal{R}_T\right] \le O\left(\sum_{i \in [K]} \max_{S_a : i \in S_a} \frac{8|S_a|\log T}{\Delta_a - 2(k^{\star 2} + 2)\eta}\right)$$

Experiment Results

- **Same Threshold Problem Instance:** It has $K = 50, C = 20$, $\theta_c = 0.7, \delta = 0.1$ and $\epsilon = 0.1$. The mean loss of arm $i \in [K]$ is $0.25 + (i-1)/100$.
- **Different Threshold Problem Instance:** It has $K = 5, \delta = 0.1$, $\epsilon = 0.1$, $\gamma = 10^{-3}$, $\boldsymbol{\mu} = [0.9, 0.89, 0.87, 0.6, 0.3]$, and $\boldsymbol{\theta} = [0.7, 0.7, 0.7, 0.58, 0.35]$.



Cumulative Regret of CSB-ST v/s Amount of Resource (Leftmost Figure) and Different Values of Same Threshold (Middle Figure). Cumulative regret of CSB-DT v/s Amount of Resource (Rightmost Figure).

Future Directions

- We decoupled the problem of threshold and mean loss estimation. It can be done jointly, leading to better performance guarantees.
- Another extension of our work is to relax the assumptions that mean losses are strictly positive, and time horizon T is known.

References

[1] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pages 1152–1161, 2015.

[2] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5101–5109, 2018.

[3] Arun Verma, Manjesh K Hanawal, Arun Rajkumar, and Raman Sankaran. Censored semi-bandits: A framework for resource allocation with censored feedback. *Appearing in Neural Information Processing Systems (NeurIPS)*, 2019.