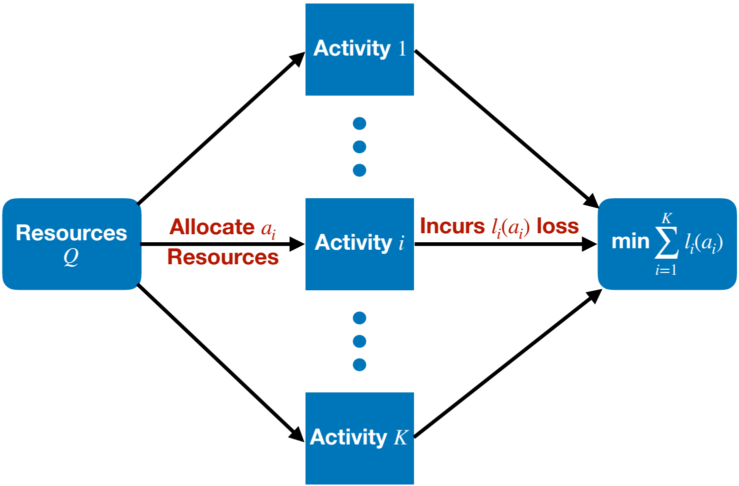
Censored Semi-Bandits: A Framework for Resource Allocation with Censored Feedback

Arun Verma, IIT Bombay
Manjesh K. Hanawal, IIT Bombay
Arun Rajkumar, IIT Madras
Raman Sankaran, LinkedIn

Resource Allocation Problem

Resource Allocation Problem



- **How to do resource allocation with stochastic loss function?**

Motivation

Many real-world problems

- Stochastic Network Utility Maximization (Yi and Chiang, 2008)
- Police patrolling (Curtin et al., 2010)
- Advertisement budget allocation (Lattimore et al., 2014)
- Traffic regulations and enforcement (Adler et al., 2014; Rosenfeld and Kraus, 2017)
- Supplier selection (Abernethy et al., 2016)
- Poaching control (Nguyen et al., 2016; Gholami et al., 2018)

Outline

Censored Semi-Bandits

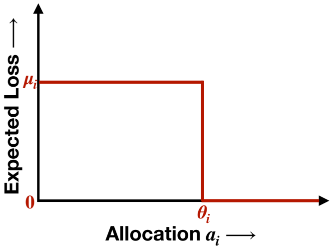
Censored Semi-Bandits

- Amount of resources: Q
- Number of arms (activities): K
- Resource allocation: $\boldsymbol{a} \doteq \{a_i\}_{i=1}^K$, where a_i denotes the resource allocated to arm i.
- All feasible allocations: $\mathcal{A}_Q \doteq \{\boldsymbol{a} : \sum_{i=1}^K a_i \leq Q\}$

- Expected loss observed from arm i is:

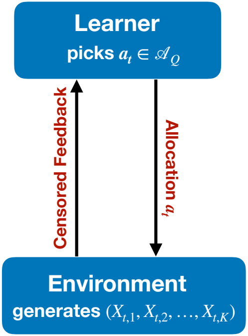
$$\mathbb{E}\left[l(a_i)\right] = \begin{cases} \mu_i & \text{if } a_i < \theta_i \\ 0 & \text{otherwise} \end{cases}$$

where μ_i is the mean loss and θ_i is the associated threshold of arm i.



- **Both $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^K$ and $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^K$ are unknown vectors.**

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 $\mathbb{E}[X_{t,i}] = \mu_i$ and sequence $(X_{t,i})_{t\geq 1}$ is i.i.d. for all $i \in [K]$.

2. **Learner** picks an allocation vector $\boldsymbol{a}_t \in \mathcal{A}_Q$.

3. **Feedback:** The learner observes a random **censored** feedback $\boldsymbol{Y_t} = \{Y_{t,i} : i \in [K]\}$, where $Y_{t,i} = X_{t,i}\mathbb{1}_{\{a_{t,i} < \theta_i\}}$.

4. **Incurs Loss:** $\sum_{i\in[K]} Y_{t,i}$.

Performance Measure

- **Optimal allocation**

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

- **Expected (pseudo) regret over a period of T for policy π:**

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

where $a_{t,i}(\pi)$ is the resources allocated to arm i by policy π in the round t.

- **A good policy should have sub-linear expected regret, i.e.,**

$$\lim_{T\to\infty} \frac{\mathbb{E}\left[\mathcal{R}_T\right]}{T} \to 0.$$

Threshold Equivalence

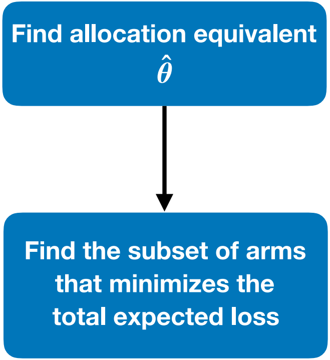
Allocation Equivalent

The two threshold vectors $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ are `allocation equivalent` if:

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

where $\boldsymbol{\mu}$ and Q are fixed.

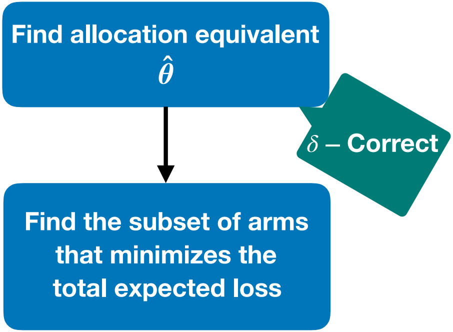
Algorithm Idea



Algorithm has two phases:

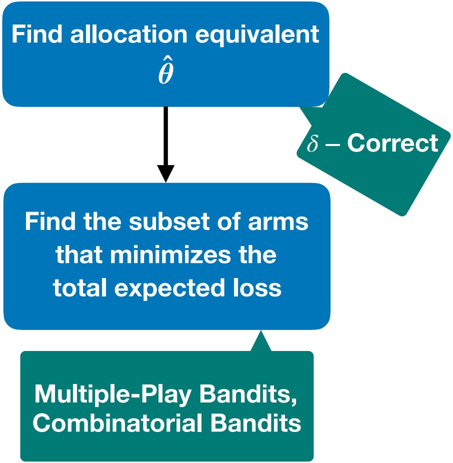
1. **Threshold Estimation Phase:** Find an allocation equivalent to $\boldsymbol{\theta}$
2. **Regret Minimization Phase:** Select the subset of arms that minimizes the total expected loss

How does Algorithm work?



- **$\boldsymbol{\delta}$—correct:** $\hat{\boldsymbol{\theta}}$ is an allocation equivalent to $\boldsymbol{\theta}$ with probability at least $1 - \delta$

How does Algorithm work?



- Selecting the best subset of arms using bandit Algorithms (MP-TS (Komiyama et al., 2015), CTS (Wang and Chen, 2018))

Same Threshold Case

Same Threshold Case

Setting:

- $\forall i \in [K] : \theta_i = \theta_c$ where $\theta_c \in \mathbb{R}^+$ and $Q \geq \theta_c$.

Optimal Allocation

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.

Allocation Equivalent (Verma et al., 2019, Lemma 1)

Let $\hat{\theta}_c = Q/M$. Then the allocation equivalent of θ_c is $\hat{\theta}_c$.
Further $\hat{\theta}_c \in \Theta = \{Q/K, Q/(K-1), \cdots, Q\}$.

Allocation Equivalent:

- Example: $K = 5, Q = 1, \theta_c = 0.3$, and $\Theta = \{0.2, 0.25, 0.33, 0.5, 1\}$.
  Given problem, $\hat{\theta}_c = 0.33$ is allocation equivalent to θ_c.

Algorithm for Same Threshold Case: CSB-ST

Threshold Estimation Phase

Let $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9\}$ and $\theta_c = \theta_6$

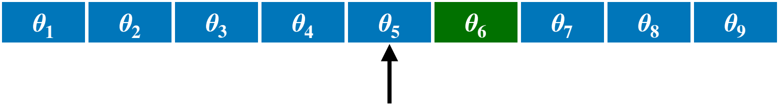
θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9

- Start a binary search to find allocation equivalent in Θ.

Algorithm for Same Threshold Case: CSB-ST

Threshold Estimation Phase

Let $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9\}$ and $\theta_c = \theta_6$

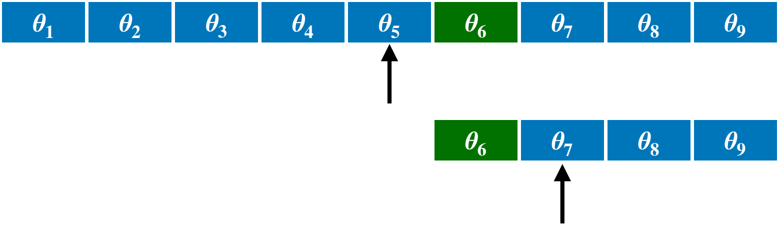


- Select $\theta_i \in \Theta$ and allocate θ_i resources to randomly selected $\frac{Q}{\theta_i}$ arms
- If loss is observed, θ_i is underestimate of θ_c.

Algorithm for Same Threshold Case: CSB-ST

Threshold Estimation Phase

Let $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9\}$ and $\theta_c = \theta_6$

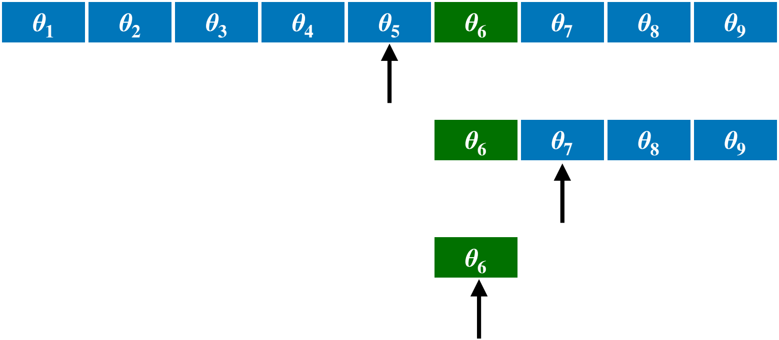


- If loss is not observed for consecutive $N(\delta)$ rounds, θ_i is overestimate of θ_c.

Algorithm for Same Threshold Case: CSB-ST

Threshold Estimation Phase

Let $\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9\}$ and $\theta_c = \theta_6$



- $\delta-$correct allocation is found.

Algorithm for Same Threshold Case: CSB-ST

Number of Rounds (Verma et al., 2019, Lemma 2)

Let for all $i \in [K]$, $\mu_i \geq \epsilon > 0$. Then with probability at least $1-\delta$, the number of rounds needed by CSB-ST to find an allocation equivalent to θ_c is bounded as

$$T_{\theta_s}(\delta) \leq \frac{N(\delta)}{\max\{1, \lfloor Q \rfloor\}} \log_2 K$$

where $N(\delta) = \frac{\log\left(\frac{\log_2 K}{\delta}\right)}{\log(1/(1-\epsilon))}$.

Regret Minimization Phase

- Once $\hat{\theta}_c$ is known, top $Q/\hat{\theta}_c$ arms are selected using Multiple-Play Thomson Sampling (MP-TS) algorithm (Komiyama et al., 2015) in subsequent rounds.

Regret Bounds

Lower Bound (Anantharam et al., 1987, Theorem 3.1)

$$\lim_{T\to\infty} \mathbb{P}\left\{ \frac{\mathbb{E}[\mathcal{R}_T]}{\log T} \geq \sum_{i\in[K]\setminus[K-M]} \frac{(1-o(1))(\mu_i - \mu_{K-M})}{d(\mu_{K-M}, \mu_i)} \right\} = 1,$$

where $d(p, q)$ is the Kullback-Leibler (KL) divergence between two Bernoulli distributions with parameter p and q.

Upper Bound (Verma et al., 2019, Theorem 1)

Let $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_{K-M} < \mu_{K-M+1} \leq \ldots \leq \mu_K$ and $\delta = 1/T$. Then the expected regret of CSB-ST over a period of T is given by

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

$\implies$ **The regret of CSB-ST is asymptotically optimal.**

Different Threshold Case

Different Threshold Case

Setting:

- Threshold may not be the same for all arms.

Optimal Allocation (Verma et al., 2019, Proposition 2)

The optimal allocation is the solution given by 0-1 knapsack having capacity Q and K items where item i has weight θ_i and value μ_i.

Allocation Equivalent:

- Let $r := \left(Q - \sum_{a_i^\star \geq \theta_i} \theta_i\right)$. If $r = 0 \implies$ 'hopeless problem'
- An allocation equivalent can be found if $\boldsymbol{r > 0}$

Allocation Equivalent (Verma et al., 2019, Lemma 3)

Let $\gamma := r/K > 0$ and $\forall i \in [K] : \hat{\theta}_i \in [\theta_i, \theta_i + \gamma]$. Then $\hat{\boldsymbol{\theta}}$ is an allocation equivalent to $\boldsymbol{\theta}$.

Algorithm for Different Threshold Case: CSB-DT

Threshold Estimation Phase

- Each $\hat{\theta}_i$ is estimated by using binary search in $[0, Q]$ interval and keep track of lower bound $\theta_{l,i}$ and upper bound $\theta_{u,i}$.
- Stop search when $\boldsymbol{\theta_{u,i} - \theta_{l,i} \leq \gamma}$

Number of Rounds (Verma et al., 2019, Lemma 4)

Let $\gamma > 0$ and for all $i \in [K]$, $\mu_i \geq \epsilon > 0$. Then with probability at least $1 - \delta$, the number of rounds needed by CSB-DT to find an allocation equivalent to $\boldsymbol{\theta}$ is bounded as

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

Regret Minimization Phase

- Once $\hat{\theta}$ is known, a subset of arms is selected using Combinatorial Thomson Sampling (CTS) algorithm (Wang and Chen, 2018).

Regret Bounds

Upper Bound Verma et al. (2019, Theorem 2)

Let $\gamma > 0$, $S_a = \{i : a_i < \theta_i\}$ for any feasible allocation a, and $\Delta_a = \sum_{i=1}^{K} \mu_i \big(\mathbb{1}_{\{a_i < \theta_i\}} - \mathbb{1}_{\{a_i^\star < \theta_i\}}\big)$. 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

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

- **Instance I (Same Threshold Problem Instance):** It has $K = 20, Q = 7, \theta_c = 0.7, \delta = 0.1$ and $\epsilon = 0.1$. The mean loss of arm $i \in [K]$ is $\mu_i = 0.25 + (i-1)/50$.
- **Instance II (Different Threshold Problem Instance):** It has $K = 5, Q = 2, \delta = 0.1, \epsilon = 0.1, \gamma = 10^{-3}$. The mean loss vector is $\boldsymbol{\mu} = [0.9, 0.89, 0.87, 0.58, 0.3]$ and corresponding threshold vector is $\boldsymbol{\theta} = [0.7, 0.7, 0.7, 0.6, 0.35]$.

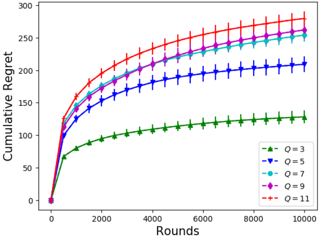


Figure 1: Cumulative Regret of CSB-ST for different amount of resources in Instance I.

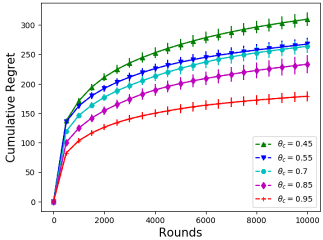


Figure 2: Cumulative Regret of CSB-ST for different thresholds in Instance I.

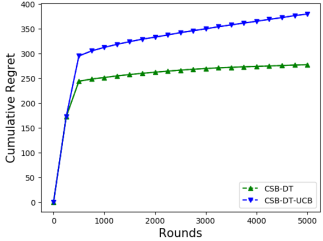


Figure 3: Cumulative Regret of CSB-DT and UCB based Algorithms for Instance II.

References

Y. Yi and M. Chiang. Stochastic network utility maximisation—a tribute to kelly's paper published in this journal a decade ago. *European Transactions on Telecommunications*, 19(4):421–442, 2008.

Kevin M Curtin, Karen Hayslett-McCall, and Fang Qiu. Determining optimal police patrol areas with maximal covering and backup covering location models. *Networks and Spatial Economics*, 10(1):125–145, 2010.

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Optimal resource allocation with semi-bandit feedback. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 477–486. AUAI Press, 2014.

Nicole Adler, Alfred Shalom Hakkert, Jonathan Kornbluth, Tal Raviv, and Mali Sher. Location-allocation models for traffic police patrol vehicles on an interurban network. *Annals of Operations Research*, 221(1):9–31, 2014.

Ariel Rosenfeld and Sarit Kraus. When security games hit traffic: Optimal traffic enforcement under one sided uncertainty. In *IJCAI*, pages 3814–3822, 2017.

Jacob D Abernethy, Kareem Amin, and Ruihao Zhu. Threshold bandits, with and without censored feedback. In *Advances In Neural Information Processing Systems*, pages 4889–4897, 2016.

References ii

Thanh H Nguyen, Arunesh Sinha, Shahrzad Gholami, et al. Capture: A new predictive anti-poaching tool for wildlife protection. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 767–775, 2016.

Shahrzad Gholami, Sara Mc Carthy, Bistra Dilkina, , et al. Adversary models account for imperfect crime data: Forecasting and planning against real-world poachers. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 823–831, 2018.

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.

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

Arun Verma, Manjesh K Hanawal, Arun Rajkumar, and Raman Sankaran. Censored semi-bandits: A framework for resource allocation with censored feedback. *To appear in Advances In Neural Information Processing Systems*, 2019.

V. Anantharam, P. Varaiya, and J. Walrand. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays- part I. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, pages 964–972, 2015.

Yuval Dagan and Crammer Koby. A better resource allocation algorithm with semi-bandit feedback. In *Proceedings of Algorithmic Learning Theory*, pages 268–320, 2018.

Xavier Fontaine, Shie Mannor, and Vianney Perchet. A problem-adaptive algorithm for resource allocation. *arXiv preprint arXiv:1902.04376*, 2019.