Near-Optimal Algorithms for Piecewise-Stationary Cascading Bandits

Lingda Wang

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign

2021 IEEE International Conference on Acoustics, Speech and Signal Processing

June 6th - 11th, 2021

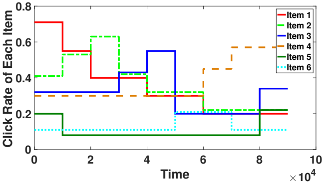
Joint work with Huozhi Zhou (UIUC), Bingcong Li (UMN), Lav R. Varshney (UIUC), and Zhizhen Zhao (UIUC)

Motivation

- Cascading bandit (CB) is a variant of multi-armed bandit (MAB) tailored for cascade model (CM) that depicts a user's online behavior. CB can be characterized by L different attraction distributions $\{f_i\}_{i=1}^{L}$ associated with items (e.g., web pages and ads).
- **Goal of CB**: Identifying the K most attractive items to the user and maximizing the number of clicks during the Learning process.

Motivation

- Cascading bandit (CB) is a variant of multi-armed bandit (MAB) tailored for cascade model (CM) that depicts a user's online behavior. CB can be characterized by L different attraction distributions $\{f_i\}_{i=1}^{L}$ associated with items (e.g., web pages and ads).
- **Goal of CB**: Identifying the K most attractive items to the user and maximizing the number of clicks during the Learning process.
- What if the attraction distributions are **non-stationary** (e.g. User's preference might change)?



Learning Protocol of Cascading Bandits

- CB $= (\mathcal{L},\ T,\ \{f_{\ell,t}\}_{\ell\in\mathcal{L},t\leq T}, K)$:
  - $\mathcal{L}$: Ground set containing L items (e.g., web pages or ads);
  - $\{f_{\ell,t}\}_{\ell\in\mathcal{L},t\leq T}$: Pmfs of attraction distributions of items in $\mathcal{L}$;
  - T is the time horizon, and K is the number of items recommended by the learner to the user.

Learning Protocol of Cascading Bandits

- CB $= (\mathcal{L},\ T,\ \{f_{\ell,t}\}_{\ell\in\mathcal{L},t\leq T}, K)$:
  - $\mathcal{L}$: Ground set containing L items (e.g., web pages or ads);
  - $\{f_{\ell,t}\}_{\ell\in\mathcal{L},t\leq T}$: Pmfs of attraction distributions of items in $\mathcal{L}$;
  - T is the time horizon, and K is the number of items recommended by the learner to the user.
- In each time step $t = 1, 2, \ldots, T$:
  - Given historical data, the learner selects a K-item ranked list $\mathcal{A}_t := (a_{1,t}, \ldots, a_{K,t}) \in \Pi_K(\mathcal{L})$ to recommend to the user;
  - The learner observes the feedback from the user at time t:

$$F_t = \begin{cases} \emptyset, & \text{if no click,} \\ \arg\min_k\{1 \leq k \leq K : Z_{a_{k,t},t} = 1\}, & \text{otherwise,} \end{cases}$$

which indicates the first item clicked by the user $(Z_{a_{k,t},t})$ or no click $(\emptyset)$.

Piecewise-Stationary Cascading Bandits: Problem Formulation

- Consider Piecewise-Stationary CB (PS-CB) with N segments, where $N = 1 + \sum_{t=1}^{T-1} \mathbb{I}\{\exists \ell \in \mathcal{L} \text{ s.t. } f_{\ell,t} \neq f_{\ell,t+1}\}$.
- For the ith piecewise-stationary segment $t \in [\nu_{i-1}+1, \nu_i]$, the attraction distribution of item ℓ, denoted as f_ℓ^i, remains unchanged.

Piecewise-Stationary Cascading Bandits: Problem Formulation

- Consider Piecewise-Stationary CB (PS-CB) with N segments, where $N = 1 + \sum_{t=1}^{T-1} \mathbb{I}\{\exists \ell \in \mathcal{L} \text{ s.t. } f_{\ell,t} \neq f_{\ell,t+1}\}$.
- For the ith piecewise-stationary segment $t \in [\nu_{i-1}+1, \nu_i]$, the attraction distribution of item ℓ, denoted as f_ℓ^i, remains unchanged.
- **Goal**: Minimize the expected cumulative regret:

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T} R\left(\mathcal{A}_t, \mathrm{w}_t, \mathrm{Z}_t\right)\right],$$

where w_t is attraction probability vector. Here, $R(\mathcal{A}_t, \mathrm{w}_t, \mathrm{Z}_t) = r(\mathcal{A}_t^*, \mathrm{w}_t) - r(\mathcal{A}_t, \mathrm{Z}_t)$ with $\mathcal{A}_t^* = \arg\max_{\mathcal{A}_t \in \Pi_K(\mathcal{L})} r\left(\mathcal{A}_t, \mathrm{w}_t\right)$ being the optimal list that maximizes the expected reward, where $r\left(\mathcal{A}_t, \mathrm{w}_t\right) = 1 - \prod_{k=1}^{K}\left(1 - w_{a_{k,t},t}\right)$.

Contributions

- **Tighter regret bounds.** The proposed two algorithms are shown to have regret bounds $\mathcal{O}(\sqrt{NLT \log T})$, which tightens those in *Li et al.* [1] by a factor of $\sqrt{L}$ and $\sqrt{L} \log T$, respectively.

[1] Chang Li and Maarten de Rijke, "Cascading non-stationary bandits: Online learning to rank in the non-stationary cascade model," in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI 2019), 2019, pp. 2859–2865.

Contributions

- **Tighter regret bounds.** The proposed two algorithms are shown to have regret bounds $\mathcal{O}(\sqrt{NLT \log T})$, which tightens those in *Li et al.* [1] by a factor of $\sqrt{L}$ and $\sqrt{L}\log T$, respectively.

- **Matching lower bound.** We establish that the minimax regret lower bound for PS-CB is $\Omega(\sqrt{NLT})$. Such a lower bound: i) implies the proposed algorithms are optimal up to a logarithm factor; ii) is the first to characterize dependence on N, L, and T for PS-CB.

[1] Chang Li and Maarten de Rijke, "Cascading non-stationary bandits: Online learning to rank in the non-stationary cascade model," in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI 2019), 2019, pp. 2859–2865.

Contributions

- **Tighter regret bounds.** The proposed two algorithms are shown to have regret bounds $\mathcal{O}(\sqrt{NLT \log T})$, which tightens those in *Li et al.* [1] by a factor of $\sqrt{L}$ and $\sqrt{L}\log T$, respectively.

- **Matching lower bound.** We establish that the minimax regret lower bound for PS-CB is $\Omega(\sqrt{NLT})$. Such a lower bound: i) implies the proposed algorithms are optimal up to a logarithm factor; ii) is the first to characterize dependence on N, L, and T for PS-CB.

- **Better numerical peformance.** Numerical experiments on a real-world benchmark dataset reveal the merits of proposed algorithms over state-of-the-art approaches.

[1] Chang Li and Maarten de Rijke, "Cascading non-stationary bandits: Online learning to rank in the non-stationary cascade model," in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI 2019), 2019, pp. 2859–2865.

The Proposed Algorithms

- `GLRT-CascadeUCB` and `GLRT-CascadeKL-UCB` algorithms run in three phases:
  - For p fraction of the time, the algorithms select K items by uniform sampling. For the rest of the time, the algorithms select K items with highest UCB/KL-UCB indices;
  - Update the statistics of K selected items:

$$\mathrm{UCB}(\ell) = \hat{\mathrm{w}}(\ell) + \sqrt{\frac{3\log(t-\tau)}{2n_\ell}},$$
$$\mathrm{UCB}_{\mathrm{KL}}(\ell) = \max\{q \in [\hat{\mathrm{w}}(\ell), 1] : n_\ell \times \mathrm{KL}(\hat{\mathrm{w}}(\ell), q) \le g(t-\tau)\};$$

  - At the end of each round, run GLRT change-point detector [2] on selected items at this round. If at least one item's click probability has changed, restart the UCB indices/KL-UCB indices of all items.

[2] Lilian Besson and Emilie Kaufmann, "The generalized likelihood ratio test meets klucb: an improved algorithm for piecewise non-stationary bandits," arXiv preprintarXiv:1902.01575, 2019.

Theoretical Analysis: Regret Upper Bounds

Theorem (Wang et al. 2019, `GLRT-CascadeUCB`)

Under mild assumptions, `GLRT-CascadeUCB` *guarantees*

$$\mathcal{R}(T) \leq \sum_{i=1}^{N} \widetilde{C}_i + Tp + \sum_{i=1}^{N-1} d_i + 3NTL\delta,$$

where $\widetilde{C}_i = \sum_{\ell=K+1}^{L} \frac{12}{\Delta^{i}_{s_i(\ell),s_i(K)}} \log T + \frac{\pi^2}{3} L$.

Corollary (Wang et al. 2019, `GLRT-CascadeUCB`)

The regret of `GLRT-CascadeUCB` *is established by choosing* $\delta = 1/T$ *and* $p = \sqrt{NL \log T / T}$*:*

$$\mathcal{R}(T) = \mathcal{O}\left(\frac{N(L-K)\log T}{\Delta^{\min}_{\text{opt}}} + \frac{\sqrt{NLT \log T}}{\left(\Delta^{\min}_{\text{change}}\right)^2} \right).$$

Theoretical Analysis: Regret Upper Bounds

Theorem (Wang et al. 2019, `GLRT-CascadeKL-UCB`)

Under mild assumptions, `GLRT-CascadeKL-UCB` *guarantees*

$$\mathcal{R}(T) \leq T(N-1)(L+1)\delta + Tp + \sum_{i=1}^{N-1} d_i + NK \log\log T + \sum_{i=0}^{N-1} \widetilde{D}_i,$$

where $\widetilde{D}_i$ *is a term depending on* $\log T$ *and the suboptimal gaps.*

Corollary (Wang et al. 2019, `GLRT-CascadeKL-UCB`)

Choosing the same δ *and* p *as* `GLRT-CascadeUCB`, `GLRT-CascadeKL-UCB` *has the same regret as* `GLRT-CascadeUCB`*.*

- As T becomes larger, the regret is dominated by the cost of the change-point detection component, implying the regret is $\mathcal{O}(\sqrt{NLT \log T})$.

Theoretical Analysis: Regret Lower Bound

Theorem (Wang et al. 2019, Lower Bound)

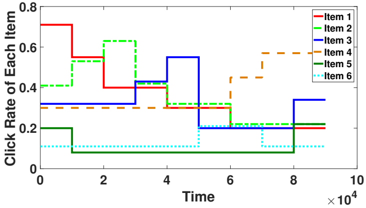
If $L \geq 3$ and $T \geq MN\frac{(L-1)^2}{L}$, then for any policy, the worst-case regret is at least $\Omega(\sqrt{NLT})$, where $M = 1/\log\frac{4}{3}$, and $\Omega(\cdot)$ notation hides a constant factor that is independent of N, L, and T.

- This lower bound is the first characterization involving N, L, and T. And it indicates our proposed algorithms are nearly order-optimal within a logarithm factor $\sqrt{\log T}$.

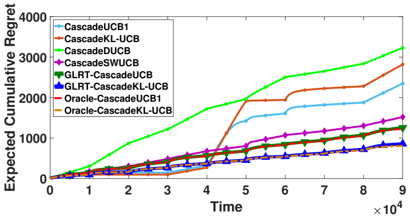
Summary

	Regret	Diff
Lower Bound	$\mathcal{O}(\sqrt{NLT})$	0
CascadeDUCB[3]	$\mathcal{O}(L\sqrt{NT}\log T)$	$\mathcal{O}(\sqrt{L}\log T)$
CascadeSWUCB[3]	$\mathcal{O}(L\sqrt{NT\log T})$	$\mathcal{O}(\sqrt{L\log T})$
GLRT-CascadeUCB	$\mathcal{O}\left(\sqrt{NLT\log T}\right)$	$\mathcal{O}(\sqrt{\log T})$
GLRT-CascadeKL-UCB	$\mathcal{O}\left(\sqrt{NLT\log T}\right)$	$\mathcal{O}(\sqrt{\log T})$

[3] Chang Li and Maarten de Rijke, "Cascading non-stationary bandits: Online learning to rank in the non-stationary cascade model," in Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI 2019), 2019, pp. 2859–2865.



(a) Click rate.



(b) Expected cumulative regret.

- Experiment settings:
  - Use the Yahoo! benchmark dataset[4];
  - Pre-process the dataset by adopting the same method as *Cao et al.* [5], where $L = 6$, $K = 2$, $N = 9$, and $T = 90000$.

[4] https://webscope.sandbox.yahoo.com

[5] Yang Cao, Zheng Wen, Branislav Kveton, and Yao Xie, "Nearly optimal adaptive procedure with change detection for piecewise-stationary bandit," in Proc. 22nd Int. Conf. Artif. Intell. Stat. (AISTATS 2019), 2019, pp. 418–427.

Thank You!

- Full version of *Nearly Optimal Algorithms for Piecewise-Stationary Cascading Bandits* is available online at: `https://arxiv.org/abs/1909.05886`.