Online Learning to Rank under Corruption: A Robust Cascading Bandits Approach

Fatemeh Ghaffari
UMass Amherst
Amherst, Massachusetts, USA

Siddarth Sitaraman
Brown University
Providence, Rhode Island, USA

Xutong Liu
University of Washington - Tacoma
Tacoma, Washington, USA

Xuchuang Wang
Chinese University of Hong Kong
Hong Kong, China

Mohammad Hajiesmaili
UMass Amherst
Amherst, Massachusetts, USA

ABSTRACT

Online learning to rank (OLTR) studies how to recommend a short ranked list of items from a large pool and improves future rankings based on user clicks. This setting is commonly modeled as cascading bandits, where the objective is to maximize the likelihood that the user clicks on at least one of the presented items across as many timesteps as possible. However, such systems are vulnerable to click fraud and other manipulations (i.e., corruption), where bots or paid click farms inject corrupted feedback that misleads the learning process and degrades user experience. In this paper, we propose M^{2}UCB-V, a robust algorithm that incorporates a novel mean-of-medians estimator, which to our knowledge is applied to bandits with corruption setting for the first time. This estimator behaves like a standard mean in the absence of corruption, so no cost is paid for robustness. Under corruption, the median step filters out outliers and corrupted samples, keeping the estimate close to its true value. Updating this estimate at every round further accelerates empirical convergence in experiments. Hence, M^{2}UCB-V achieves optimal logarithmic regret in the absence of corruption and degrades gracefully under corruptions, with regret increasing only by an additive term tied to the total corruption. Comprehensive and extensive experiments on real-world datasets further demonstrate that our approach consistently outperforms prior methods while maintaining strong robustness. In particular, it achieves a 97.35% and a 91.60% regret improvement over two state-of-the-art methods.

1. INTRODUCTION

Learning to rank lies at the core of modern recommendation and information retrieval systems, where the goal is to present users with an ordered list of items tailored to their preferences. Online learning to rank extends this paradigm by updating the recommendations as new feedback arrives in sequence, enabling systems to adapt quickly to user behavior [7, 8, 17]. For example on Yelp, the system ranks local businesses so that users find good services quickly [12]. On music and movie platforms, it surfaces songs and films that match a user's taste while exploring new options [17, 10, 8].

A widely studied model of user interactions in these applications is the *cascade model* [7, 10, 8], where users examine results sequentially from top to bottom and click the first attractive item. The cascade model captures the position-dependent nature of user feedback and leads naturally to the cascading bandits framework, a special case of combinatorial bandits [1, 2] with non-linear rewards. Each item k has an unknown click probability μ_k, and the learner aims to sequentially construct ranked lists to maximize the probability of obtaining a click. The learning agent explores by estimating item attractiveness from past observations, while balancing the need to exploit high-probability items to maximize user satisfaction.

Practical OLTR systems face substantial challenges from corrupted environments [4, 12, 17]. For example, click fraud in online advertising generates misleading clicks through bots, degrading both system revenue and user experience. Similarly, e-commerce platforms suffer from fake reviews. These corruptions are pervasive in deployed systems. Thus, robust algorithms must achieve near-optimal performance in benign settings while degrading gracefully under corruption, even when the amount of corruption is unknown.

Designing a robust algorithm for the cascading bandit setting poses unique challenges. The non-linear reward function breaks the applicability of standard linear confidence bounds widely used in existing works, making them face significant limitations: some achieve optimality in uncorrupted settings but fail under corruption [7, 10]; others offer robustness but sacrifice optimal guarantees [4, 12]; yet others rely on epoch-based designs that slow convergence and underperform in practice [13]. As a result, there remains a critical gap in the literature: we lack algorithms that are simultaneously (i) optimal in trustworthy environments, (ii) provably robust under adversarial corruption, and (iii) reliable in real-world deployments.

Contributions. In [3], We address the central question: *How can we efficiently and robustly learn in corrupted cascading bandit environments in a way that also performs reliably on real data?*

Algorithm design We develop Algorithm 1, Model selection calibrated Mean-of-medians Variance-aware UCB (M^{2}UCB-V), an algorithm that remains robust under adversarial corruption, yet performs optimally when no corruption is present, without requiring prior knowledge of the corruption level. The key design ideas behind M^{2}UCB-V include: (i) incorporating a calibrated mean-of-medians estimator that ensures reliable performance both with and without corruption, (ii) introducing a variance-aware refinement of UCB to further tighten the regret bound, and (iii) employing a model selection mechanism to automatically adapt to unknown corruption.

First, we incorporate a calibrated mean-of-medians mechanism that leverages a robust median-based estimator. We show that in the absence of corruption, the mean-of-medians behaves like a UCB mean estimator and achieves optimal performance. Under corruption, the median step selects uncorrupted samples with high probability, ensuring robustness. While this estimator has been previously used in heavy-tailed bandits [16, 14], applying it to adversarial corruption is novel and requires addressing new challenges in careful adaptation to guarantee sufficient uncorrupted samples per item. Because click feedback is asymmetric and typically Bernoulli, unlike

Learning-Augmented Algorithms: Theory and Applications (LATA) 2026
Ann Arbor, MI, USA

Table 1: Comparison of cascading bandit algorithms in stochastic and corrupted settings.

Algorithm	Regret w/o corruption	Stoch. LB†	Regret w/ corruption	Robust	Corr. factor
CascadeUCB-V [10]	$O\left(\sum_{k\notin S^*}\frac{\log T}{\Delta_k}\right)$	✓	–	✗	–
CascadeRAC [12]	$O\left(\sum_{k=d+1}^{K}\frac{d\log T\log(KT)}{\Delta_k}\right)$	✗	$O\left(\sum_{k=d+1}^{K}\frac{d(CK\log(KT)+\log T)\log(KT)}{\Delta_k}\right)$	✓	Multiplicative
FORC [4]	$O\left(\sum_{k=1}^{K}\sum_{j=k+1}^{K}\frac{\log T\log(KT)}{\Delta_{kj}}\right)$	✗	$O\left(\sum_{k=d+1}^{K}\frac{d(CK\log(KT)+\log T)\log(KT)}{\Delta_k}\right)$	✓	Multiplicative
CascadeCBARBAR [13]	$O\left(\frac{d^2K}{\Delta}\log^2 T\right)$	✗	$O\left(dC+\frac{d^2K}{\Delta}\log^2 T\right)$	✓	Additive
M^{2}UCB-V (Ours)	$O\left(\frac{K\log T}{\Delta}\right)$	✓	$O\left(KC+\frac{K\log T}{\Delta}\right)$	✓	Additive

Note (†): ✓indicates the algorithm matches the known stochastic lower bound in the uncorrupted stochastic setting, ✗indicates it does not.

the symmetric assumptions common in heavy-tailed bandits, we introduce a calibration step that maps the mean-of-medians back to the underlying Bernoulli mean, producing estimates centered around the desired true value. Then, by combining this estimator with the state-of-the-art variance-aware UCB radius, we develop an algorithm that is robust against corruption, albeit with the requirement of prior knowledge of the corruption level. Last, and to remove the dependence on prior knowledge of corruption level, we incorporate the model selection framework inspired by [11] and develop the corruption-agnostic Model Selection MUCB-V (M^{2}UCB-V). Because the algorithm updates its estimates every round, it adapts quickly to changes and rapidly converges to the true optimal items once corruption is removed.

Algorithm 1 MUCB-V (calibrated Mean-of-medians variance-aware UCB)

Input: Horizon T, list size d, corr. budget C, Constants $\alpha, A, B > 0$

for each item $k \in [K]$ **do**
  $\boldsymbol{X}_k \leftarrow [\emptyset]$ ▷ Observed clicks of item K
  for $r = 1$ to $10C$ **do**
    Recommend $S_k = \{k,\ d-1$ items from $[K] \setminus \{k\}\}$.
    Observe click feedback and append to X_k.
  end for
end for
for rounds $t = 10KC, \ldots, T$ **do**
  for each item $k \in [K]$ **do**
    $T_k(t) \leftarrow |\boldsymbol{X}_k|$, $b \leftarrow \lceil \alpha \log \max\{T_k(t), 2\} \rceil$
    $\hat{\mu}_k \leftarrow \texttt{CalibratedMeanOfMedians}(\boldsymbol{X}_k, b)$
    $\hat{v}_k \leftarrow \hat{\mu}_k(1-\hat{\mu}_k)$ ▷ empirical variance
    $s \leftarrow \max\{1, T_k(t)\}$ ▷ avoid divide-by-zero
    $\rho_k \leftarrow A\sqrt{\frac{\hat{v}_k \log t}{s}} + B\frac{\log t}{s}$
    $\bar{\mu}_k \leftarrow \min\{\hat{\mu}_k + \rho_k, 1\}$
  end for
  $S_t \leftarrow$ Top-d items by $\bar{\mu}_k$
  Play S_t,Observe click feedback and append to X_k for all observed k
end for

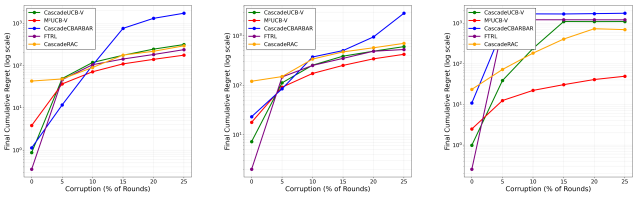
Theoretical guarantees. We provide a regret analysis establishing near-optimal bounds both in the corruption-free regime and in terms of dependence on the total corruption. In particular, we establish that the regret of M^{2}UCB-V is bounded by $O\left(KC + \frac{K\log T}{\Delta}\right)$, where T is the time horizon, C is the total corruption level, K is the number of items, S^* is the optimal recommended list, and Δ is the minimum gap between any suboptimal item k and the worst optimal item. *This means that in the absence of corruption, M^{2}UCB-V achieves optimal regret, and under corruption, the additional term is optimal up to a factor of K.* Furthermore, we prove an $\Omega(C)$ lower bound for cascading bandits under adversarial corruption, showing that any learner can incur regret at least linear under the corruption budget C.

Algorithm 2 CalibratedMeanOfMedians

Input: Reward vector $\boldsymbol{X}$, number of groups $b = \lceil \alpha \log T \rceil$

if $|\boldsymbol{X}| < b$ **then**
  return $\texttt{Mean}(\boldsymbol{X})$ ▷ not enough samples
end if
Set block size $b \leftarrow 2\left\lfloor \frac{|\boldsymbol{X}|}{2b} \right\rfloor + 1$ ▷ nearest odd
Uniformly partition $\boldsymbol{X}$ into b blocks $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(G)}$ with sizes $\approx b$
for each $j = 1, \ldots, b$ **do**
  $M_j \leftarrow \texttt{Median}\left(\boldsymbol{X}^{(j)}\right) \in \{0, 1\}$
end for
$\overline{M} \leftarrow \texttt{Mean}(M_1, \ldots, M_b)$
$\hat{\mu} \leftarrow \texttt{Calibrate}(b, \overline{M})$ ▷ invert majority map
return $\hat{\mu}$

Empirical evaluation. We conduct experiments on three large-scale real-world datasets: Yelp [15], MovieLens [5], and LastFM [9], which together represent some of the most common applications of OLTR. We also implement an OLTR-specific attack from [17], which they show is both highly effective and efficient in terms of corruption budget. Our empirical results support the theoretical guarantees across several scenarios. We validate our algorithms on both synthetic and real-world datasets, comparing against strong baselines including CascadeUCB-V [10], FTRL [6], CascadeRAC [12], and CascadeCBARBAR[13]. In a set of representative experiments, our method's cumulative regret after $40k$ rounds improves FTRL by 99.60%, CascadeUCB-V by 97.35%, CascadeRAC by 91.60, and CascadeCBARBAR by 98.41%.



(a) Yelp Dataset (b) MovieLens Dataset (c) LastFM Dataset

Figure 1: Comparing final cumulative regret of the algorithms after 40K rounds with list size $d = 10$.

2. REFERENCES

[1] W. Chen, Y. Wang, and Y. Yuan. Combinatorial multi-armed bandit: General framework, results and applications. *Journal of Machine Learning Research*, 17(1):1–46, 2016.

[2] R. Combes, M. S. Talebi, A. Proutiere, and M. Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, volume 28, pages 2116–2124, 2015.

[3] F. Ghaffari, S. Sitaraman, X. Liu, X. Wang, and M. Hajiesmaili. Online learning to rank under corruption: A robust cascading bandits approach. *arXiv preprint arXiv:2511.03074*, 2025.

[4] N. Golrezaei, V. Manshadi, J. Schneider, and S. Sekar. Learning product rankings robust to fake users. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, EC '21, page 560–561, New York, NY, USA, 2021. Association for Computing Machinery.

[5] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), Dec. 2015.

[6] S. Ito. Hybrid regret bounds for combinatorial semi-bandits and adversarial linear bandits. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2654–2667. Curran Associates, Inc., 2021.

[7] B. Kveton, C. Szepesvari, Z. Wen, and A. Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International conference on machine learning*, pages 767–776. PMLR, 2015.

[8] C. Li, H. Feng, and M. d. Rijke. Cascading hybrid bandits: Online learning to rank for relevance and diversity. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 33–42, 2020.

[9] M. Schedl. The lfm-1b dataset for music retrieval and recommendation. ICMR '16, page 103–110, New York, NY, USA, 2016. Association for Computing Machinery.

[10] D. Vial, S. Sanghavi, S. Shakkottai, and R. Srikant. Minimax regret for cascading bandits. *Advances in Neural Information Processing Systems*, 35:29126–29138, 2022.

[11] C.-Y. Wei, C. Dann, and J. Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

[12] J. Xie, C. Chen, Z. Wang, and S. Li. Cascading bandits robust to adversarial corruptions. *arXiv preprint arXiv:2502.08077*, 2025.

[13] H. Xu and J. Li. Simple combinatorial algorithms for combinatorial bandits: Corruptions and approximations. In *Uncertainty in Artificial Intelligence*, pages 1444–1454. PMLR, 2021.

[14] B. Xue, Y. Wang, Y. Wan, J. Yi, and L. Zhang. Efficient algorithms for generalized linear bandits with heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36:70880–70891, 2023.

[15] Yelp Inc. Yelp open dataset. https://business.yelp.com/data/resources/open-dataset/, 2024. Accessed: 2025-10-07.

[16] H. Zhong, J. Huang, L. Yang, and L. Wang. Breaking the moments condition barrier: No-regret algorithm for bandits with super heavy-tailed payoffs. *Advances in Neural Information Processing Systems*, 34:15710–15720, 2021.

[17] J. Zuo, Z. Zhang, Z. Wang, S. Li, M. Hajiesmaili, and A. Wierman. Adversarial attacks on online learning to rank with click feedback. *Advances in Neural Information Processing Systems*, 36:41675–41692, 2023.