Combinatorial Logistic Bandits

Xutong Liu
The Chinese University of Hong Kong
Hong Kong, China
liuxt@cse.cuhk.edu.hk

Xiangxiang Dai
The Chinese University of Hong Kong
Hong Kong, China
xxdai23@cse.cuhk.edu.hk

Xuchuang Wang
University of Massachusetts Amherst
Amherst, MA, USA
xuchuangw@gmail.com

Mohammad Hajiesmaili
University of Massachusetts Amherst
Amherst, MA, USA
hajiesmaili@cs.umass.edu

John C.S. Lui
The Chinese University of Hong Kong
Hong Kong, China
cslui@cse.cuhk.edu.hk

Abstract

Combinatorial multi-armed bandit (CMAB) is a fundamental online learning framework that can optimize cumulative rewards in networked systems under uncertainty. Real-world applications like content delivery and channel allocation often feature binary base arm rewards and nonlinear total reward functions. This paper introduces combinatorial logistic bandits (CLogB), a contextual CMAB framework with the base arm reward modeled as a nonlinear logistic function of the context, and the feedback is governed by a general arm-triggering process. We study CLogB with smooth reward functions, covering applications such as online content delivery, online multi-LLM selection, and dynamic channel allocation. Our first algorithm, CLogUCB, uses a variance-agnostic exploration bonus and achieves a regret bound of $\tilde{O}(d\sqrt{\kappa KT})$, where d is the feature dimension, κ reflects logistic model nonlinearity, K is the maximum number of triggered arms, and $\tilde{O}$ ignores logarithmic factors. This improves on prior results by $\tilde{O}(\sqrt{\kappa})$. We further propose VA-CLogUCB, a variance-adaptive enhancement achieving regret bounds of $\tilde{O}(d\sqrt{KT})$ under standard smoothness conditions and $\tilde{O}(d\sqrt{T})$ under stronger variance conditions, removing dependence on K. For time-invariant feature maps, we enhance computational efficiency by avoiding nonconvex optimization while maintaining $\tilde{O}(d\sqrt{T})$ regret. Experiments on synthetic and real-world datasets validate the superior performance of our algorithms, demonstrating their effectiveness and scalability for real-world networked systems.

CCS Concepts

• **Theory of computation** → **Online learning theory**; • **Computing methodologies** → *Planning under uncertainty*; **Learning from implicit feedback**; • **Networks** → **Network performance analysis**.

Keywords

Multi-armed bandits, combinatorial multi-armed bandits, logistic model, variance-adaptive, regret

SIGMETRICS Abstracts '25, Stony Brook, NY, USA.

ACM ISBN 979-8-4007-1593-8/25/06
https://doi.org/10.1145/3726854.3727279

ACM Reference Format:
Xutong Liu, Xiangxiang Dai, Xuchuang Wang, Mohammad Hajiesmaili, and John C.S. Lui. 2025. Combinatorial Logistic Bandits. In *Abstracts of the 2025 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS Abstracts '25), June 9–13, 2025, Stony Brook, NY, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3726854.3727279

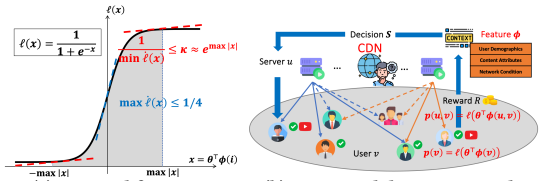
1 Combinatorial Logistic Bandit Model

We introduce our model for the combinatorial logistic bandit (CLogB) problem. A CLogB instance is a tuple $([m], \Phi, \Theta, \mathcal{S}, D_{\text{trig}}, R)$, where $[m]$ are base arms, $\mathcal{S}$ are combinatorial actions, Φ are feature maps, Θ is the parameter space, D_{trig} is the probabilistic triggering function, and R is the reward function, which are introduced later. Our model is mainly based on [3, 4], but introduces a nonlinear parameterization for the binary outcome of each base arm based on the logistic bandit model [1]. For more comparisons with existing models, see Section 2.2 in our full paper [2].

Base arms and combinatorial actions. The environment has a set of $[m] = \{1, 2, ..., m\}$ base arms and chooses an unknown parameter $\boldsymbol{\theta}^* \in \Theta$, where Θ is the set of feasible parameters. At each round $t \in [T]$, the environment first reveals a feature map $\phi_t \in \Phi$ to the learner, where ϕ_t is a function $[m] \to \mathbb{R}^d$. The environment then draws Bernoulli outcomes $\boldsymbol{X}_t = (X_{t,1}, ...X_{t,m}) \in \{0, 1\}^m$ with mean $\mu_{t,i} := \mathbb{E}[X_{t,i}|\mathcal{H}_t] = \ell(\boldsymbol{\theta}^{*\top}\boldsymbol{\phi}_t(i))$. Here, $\boldsymbol{\theta}^{*\top}\boldsymbol{\phi}_t(i)$ is the linear predictor, $\ell : \mathbb{R} \to \mathbb{R}_+$ is the *sigmoid* function $\ell(x) := (1 + e^{-x})^{-1}$ as shown in Fig. 1a that links the linear predictor and the mean $\mu_{t,i}$ in a nonlinear manner, and $\mathcal{H}_t$ denotes the history information. Then the learner selects a combinatorial action $S_t \in \mathcal{S}$, where $\mathcal{S}$ is the set of feasible actions.

Probablistically triggering arm feedback. We consider a feedback process that involves scenarios where each base arm in a super arm S_t does not always reveal its outcome, even probabilistically. To handle such probabilistic feedback, we assume that after the action S_t is selected, the base arms in a random set $\tau_t \sim D_{\text{trig}}(S_t, \boldsymbol{X}_t)$ are triggered depending on the outcome $\boldsymbol{X}_t$, where $D_{\text{trig}}(S, \boldsymbol{X})$ is the probabilistic triggering function on the subsets $2^{[m]}$. This means that the outcomes of the arms in τ_t, i.e., $(X_{t,i})_{i\in\tau_t}$ are revealed as feedback to the learner. We let $p_i^{\boldsymbol{\mu},S}$ denote the probability that base arm i is triggered when the action is S, the mean vector is $\boldsymbol{\mu}$.

Reward function. At the end of round t, the learner receives a nonnegative reward $R(S_t, \boldsymbol{X}_t, \tau_t)$, determined by action S_t, outcome $\boldsymbol{X}_t$, and triggered arm set τ_t. Similarly to [4], we assume the expected reward to be $r(S_t; \boldsymbol{\mu}_t) := \mathbb{E}[R(S_t, \boldsymbol{X}_t, \tau_t)]$, a function of



(a) Sigmoid function (b) Content delivery network

Figure 1: Left: illustration of a sigmoid function. Right: CLogB for content delivery networks as an application.

the unknown mean vector $\boldsymbol{\mu}_t$, where the expectation is taken over the randomness of $\boldsymbol{X}_t$ and $\tau_t \sim D_{\text{trig}}(S_t, \boldsymbol{X}_t)$.

Offline approximation oracle and approximate regret. The goal of CLogB is to accumulate as much reward as possible over T rounds by learning the underlying parameter $\boldsymbol{\theta}^*$. For many reward functions, it is NP hard to compute the exact S_t^* even when $\boldsymbol{\mu}_t$ is known, so similar to [3, 4], we assume that algorithm A has access to an offline α-approximation oracle ORACLE, which takes any mean vector $\boldsymbol{\mu} \in [0,1]^m$ as input, and outputs an α-approximate solution $S \in \mathcal{S}$, i.e.,

$$S = \text{ORACLE}(\boldsymbol{\mu}) \text{ s.t. } r(S;\boldsymbol{\mu}) \ge \alpha \cdot \max_{S' \in \mathcal{S}} r(S';\boldsymbol{\mu}) \tag{1}$$

The T-round α-approximate regret is then defined as

$$\text{Reg}(T) = \mathbb{E}\left[\sum_{t=1}^{T} \left(\alpha \cdot r(S_t^*;\boldsymbol{\mu}_t) - r(S_t;\boldsymbol{\mu}_t)\right)\right], \tag{2}$$

where the expectation is taken over the randomness of outcomes $\boldsymbol{X}_1, ..., \boldsymbol{X}_T$, the triggered sets $\tau_1, ..., \tau_T$, as well as the randomness of the algorithm itself.

1.1 Key Assumptions and Conditions

We consider the following assumptions/conditions at the arm and reward level, see Seciton 2.3 of the full paper [2] for their intuitions and more explanations.

Arm-level assumptions. Assumption 1 bounds the range of plausible feature vectors for each base arm and unknown parameters. Assumption 2 bounds the nonlinearity level of the base arm mean regarding all plausible linear predictor $\boldsymbol{\theta}^\top \boldsymbol{\phi}(i)$.

Assumption 1 (Bounded parameter and arm feature). *There exists a known constant $L > 0$ such that for any $\boldsymbol{\theta} \in \Theta$, $\|\boldsymbol{\theta}\|_2 \le L$. For any $\phi \in \Phi$ and $i \in [m]$, it holds that $\|\boldsymbol{\phi}(i)\|_2 \le 1$.*

Assumption 2 (Arm-level nonlinearity). *There exists a known $\kappa > 0$ such that $\left(\min_{i\in[m],\phi\in\Phi,\boldsymbol{\theta}^*\in\Theta} \dot{\ell}(\boldsymbol{\theta}^\top \boldsymbol{\phi}(i))\right)^{-1} \le \kappa$.*

Reward-level conditions. Condition 1 indicates the reward monotonically increases when the parameter $\boldsymbol{\mu}$ increases. Condition 2 and 3 both bound the reward smoothness/sensitivity, i.e., the amount of the reward change caused by the parameter change from $\boldsymbol{\mu}$ to $\boldsymbol{\mu}'$.

Condition 1 (Monotonicity). *A CLogB problem satisfies monotonicity condition if for any action $S \in \mathcal{S}$, any mean vectors $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [0,1]^m$ s.t. $\mu_i \le \mu_i'$ for all $i \in [m]$, we have $r(S;\boldsymbol{\mu}) \le r(S;\boldsymbol{\mu}')$.*

Condition 2 (1-norm TPM bounded smoothness, [4]). *We say that a CLogB problem satisfies the 1-norm triggering probability modulated (TPM) B_1-bounded smoothness condition, if there exists $B_1 > 0$, such that for any action $S \in \mathcal{S}$, any mean vectors $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [0,1]^m$, we have $|r(S;\boldsymbol{\mu}') - r(S;\boldsymbol{\mu})| \le B_1 \sum_{i\in[m]} p_i^{\boldsymbol{\mu},S} |\mu_i - \mu_i'|$.*

Condition 3 (TPVM bounded smoothness, [3]). *We say that a CLogB problem satisfies the triggering probability and variance modulated (TPVM) (B_v, B_1, λ)-bounded smoothness condition, if there exists $B_v, B_1, \lambda > 0$ such that for any action $S \in \mathcal{S}$, any mean vector $\boldsymbol{\mu}, \boldsymbol{\mu}' \in (0,1)^m$, for any $\boldsymbol{\zeta}, \boldsymbol{\eta} \in [-1,1]^m$ s.t. $\boldsymbol{\mu}' = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\eta}$, we have $|r(S;\boldsymbol{\mu}') - r(S;\boldsymbol{\mu})| \le B_v \sqrt{\sum_{i\in[m]} (p_i^{\boldsymbol{\mu},S}) \lambda \frac{\zeta_i^2}{(1-\mu_i)\mu_i}} + B_1 \sum_{i\in[m]} p_i^{\boldsymbol{\mu},S} |\eta_i|$.*

2 Variance-Agnostic CLogUCB Algorithm, Regret, Applications, and Experiments

In this abstract, we will introduce a variance-agnostic CLogUCB algorithm with $\tilde{O}(d\sqrt{\kappa KT})$ regret under the 1-norm TPM condition. Then in the full paper [2], we devise the variance-adaptive VA-CLogUCB algorithm with improved $\tilde{O}(d\sqrt{T})$ regret under the stronger TPVM condition. Finally, we improve the computational efficiency of VA-CLogUCB while maintaining the regret results. We summarize their regret bounds and per-round time complexity in Table 1.

Maximum likelihood estimation. We first introduce the parameter learning process, which utilizes the maximum likelihood estimation (MLE) and lays the foundations of our combinatorial UCB-based algorithms throughout the paper. Based on historical data $\mathcal{H}_t = (\boldsymbol{\phi}_s, S_s, \tau_s, (X_{s,i})_{i\in\tau_s})_{s<t} \bigcup \boldsymbol{\phi}_t$, we consider the following regularized log-likelihood (or cross-entropy loss) for $t \in [T]$:

$$\begin{aligned}\mathcal{L}_t(\boldsymbol{\theta}) \coloneqq -\sum_{s=1}^{t-1} \sum_{i\in\tau_s} \big[X_{s,i} \log \ell\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_s(i)\right) + \\ \left(1 - X_{s,i}\right) \cdot \log\left(1 - \ell\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_s(i)\right)\right)\big] + \tfrac{\lambda_t}{2} \|\boldsymbol{\theta}\|_2^2.\end{aligned} \tag{3}$$

where $\lambda_t = O(d \log(Kt))$ is a time-varying regularizer that will be specified later on in Algorithms. Our MLE estimator is defined as

$$\hat{\boldsymbol{\theta}}_t \coloneqq \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}_t(\boldsymbol{\theta}). \tag{4}$$

For this loss function $\mathcal{L}_t(\boldsymbol{\theta})$, it is convenient to define a mapping $\boldsymbol{g}_t : \mathbb{R}^d \to \mathbb{R}^d$ $\boldsymbol{g}_t(\boldsymbol{\theta}) \coloneqq \sum_{s=1}^{t-1} \sum_{i\in\tau_s} \ell\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_s(i)\right) \boldsymbol{\phi}_s(i) + \lambda_t \boldsymbol{\theta}$. Then we can express the gradient $\nabla_{\boldsymbol{\theta}} \mathcal{L}_t(\boldsymbol{\theta})$ at $\boldsymbol{\theta}$ as: $\nabla_{\boldsymbol{\theta}} \mathcal{L}_t(\boldsymbol{\theta}) = \boldsymbol{g}_t(\boldsymbol{\theta}) - \sum_{s=1}^{t-1} \sum_{i\in\tau_s} X_{s,i} \boldsymbol{\phi}_s(i)$. Lastly, we define Hessian $\boldsymbol{H}_t(\boldsymbol{\theta})$ of the log-loss and the covariance matrix $\boldsymbol{V}_t$ as two important quantities used in our algorithm design and analysis.

$$\boldsymbol{H}_t(\boldsymbol{\theta}) \coloneqq \sum_{s=1}^{t-1} \sum_{i\in\tau_s} \dot{\ell}\left(\boldsymbol{\theta}^\top \boldsymbol{\phi}_s(i)\right) \boldsymbol{\phi}_s^\top(i) \boldsymbol{\phi}_s(i) + \lambda_t \boldsymbol{I}_d, \tag{5}$$

$$\boldsymbol{V}_t \coloneqq \sum_{s=1}^{t-1} \sum_{i\in\tau_s} \boldsymbol{\phi}_s^\top(i) \boldsymbol{\phi}_s(i) + \kappa \lambda_t \boldsymbol{I}_d.. \tag{6}$$

We show that $\hat{\boldsymbol{\theta}}_t$ is a good estimator by bounding the distance between $\hat{\boldsymbol{\theta}}_t$ and $\boldsymbol{\theta}^*$ via mapping g_t.

Lemma 1 (Concentration inequality for MLE). *Let $\hat{\boldsymbol{\theta}}_t$ be the MLE as defined in Eq. (4), it holds with probability at least $1-\delta$ that:*

$$\left\|\boldsymbol{g}_t\left(\hat{\boldsymbol{\theta}}_t\right) - \boldsymbol{g}_t\left(\boldsymbol{\theta}^*\right)\right\|_{\boldsymbol{H}_t^{-1}(\boldsymbol{\theta}^*)} \le \gamma_t(\delta), \quad \forall t \ge 1,$$

where $\gamma_t(\delta) \coloneqq (L + 3/2)\sqrt{d \log\left(4(1+tK)/\delta\right)}$ is the confidence radius. Therefore, it holds with probability at least $1-\delta$ that $\{\forall t \ge 1, \boldsymbol{\theta}^* \in \mathcal{A}_t(\delta)\}$, where $\mathcal{A}_t(\delta) \coloneqq \left\{\boldsymbol{\theta} \in \Theta : \left\|\boldsymbol{g}_t\left(\hat{\boldsymbol{\theta}}_t\right) - \boldsymbol{g}_t\left(\boldsymbol{\theta}\right)\right\|_{\boldsymbol{H}_t^{-1}(\boldsymbol{\theta})} \le \gamma_t(\delta)\right\}$.

Variance-agnostic confidence region. We first construct the following variance-agnostic confidence region around the MLE $\hat{\boldsymbol{\theta}}_t$.

Table 1: Summary of the main results and additional results for CLogB with time-invariant features (CLogB-TI).

CLogB	Algorithm	Condition	Coefficient	Regret Bound	Per-round Cost
(Main Result 1)	CLogUCB (Algorithm 1)	1-norm TPM	B_1	$\tilde{O}\left(B_1 d\sqrt{\kappa KT}\right)$	$\tilde{O}\left(dK^2T^2 + T_\alpha\right)$
(Main Result 2)	VA-CLogUCB (Full paper [2])	1-norm TPM	B_1	$\tilde{O}\left(B_1 d\sqrt{KT} + B_1\kappa d^2\right)$	$\tilde{O}\left(dK^2T^2 + T_{\text{nc}} + T_\alpha\right)$
(Main Result 3)	VA-CLogUCB (Full paper [2])	TPVM	B_v †, $\lambda \geq 1$‡	$\tilde{O}\left(B_v d\sqrt{T} + B_1\kappa d^2\right)$	$O\left(dK^2T^2 + T_{\text{nc}} + T_\alpha\right)$
CLogB-TI	**Algorithm**	**Condition**	**Coefficient**	**Regret Bound**	
(Additional Result 1)	EVA-CLogUCB (Full paper [2])	1-norm TPM	B_1	$\tilde{O}\left(B_1 d\sqrt{KT} + B_1\kappa Kd^2\right)$	$O\left(dK^2T^2 + T_\alpha\right)$
(Additional Result 2)	EVA-CLogUCB (Full paper [2])	TPVM	$B_v, \lambda \geq 1$	$\tilde{O}\left(B_v d\sqrt{T} + B_1\kappa Kd^2\right)$	$\tilde{O}\left(dK^2T^2 + T_\alpha\right)$

This table assumes $T \gg m \geq K \gg d$. ** T_{nc} and T_α are the time to solve a nonconvex projection problem and an α-approximation for the combinatorial optimization problem, respectively. † Generally, coefficient $B_v = O(B_1\sqrt{K})$ and the existing regret bound is improved when $B_v = o(B_1\sqrt{K})$ ‡ λ is a coefficient in TPVM condition: when λ is larger, the condition is stronger with smaller regret but can include fewer applications.

Algorithm 1 CLogUCB: **C**ombinatorial **Log**isitc **U**pper **C**onfidence **B**ound Algorithm for CLogB

1: **Input:** Base arms $[m]$, dimension d, parameter space Θ, nonlinearity coefficient κ, probability $\delta = 1/T$, offline ORACLE.
2: **for** $t = 1, ..., T$ **do**
3: Compute MLE $\hat{\boldsymbol{\theta}}_t = \text{argmax}_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}_t(\boldsymbol{\theta})$ according to Eq. (3) with $\lambda_t = d\log\left(4(1+tK)/\delta\right)$.
4: Compute the covariance matrix $\boldsymbol{V}_t$ according to Eq. (6).
5: **for** $i \in [m]$ **do**
6: $\bar{\mu}_{t,i} = \ell\left(\hat{\boldsymbol{\theta}}_t^\top \boldsymbol{\phi}_t(i)\right) + \beta_t(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{V}_t^{-1}}$ with $\beta_t(\delta) = \left(L^2 + 4L + 19/4\right)\sqrt{\kappa d\log\left(4(1+tK)/\delta\right)}$.
7: **end for**
8: $S_t = \text{ORACLE}(\bar{\mu}_{t,1}, ..., \bar{\mu}_{t,m})$ as in Eq. (1).
9: Play S_t and observe triggering arm set τ_t with their outcomes $(X_{t,i})_{i\in\tau_t}$.
10: **end for**

Lemma 2. *Let $\delta \in (0,1]$ and set the confidence radius $\beta_t(\delta) := \left(L^2 + 4L + 19/4\right)\sqrt{\kappa d\log\left(4(1+tK)/\delta\right)}$. The following region*

$$\mathcal{B}_t(\delta) := \left\{\boldsymbol{\theta}\in\Theta : \left\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_t\right\|_{\mathbf{V}_t} \leq \beta_t(\delta)\right\}, \tag{7}$$

is an anytime confidence region for $\boldsymbol{\theta}^*$ with probability at least $1-\delta$, i.e., $\Pr\left(\forall t \geq 1, \boldsymbol{\theta}^* \in \mathcal{B}_t(\delta)\right) \geq 1-\delta$.

Based on the above confidence region, we can now construct our variance-agnostic exploration bonus.

Lemma 3. *Let $B_t(\delta)$ be the confidence region with the confidence radius $\beta_t(\delta)$ as defined in Lemma 2. Let the exploration bonus be $\rho_{t,V}(i) := \frac{1}{4}\beta_t(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{V}_t^{-1}}$. Under the event $\{\forall t \geq 1, \boldsymbol{\theta}^* \in \mathcal{B}_t(\delta)\}$, it holds that, for any $i \in [m], t \geq 1$, $\ell\left(\boldsymbol{\theta}^{*\top}\boldsymbol{\phi}_t(i)\right) \leq \ell\left(\hat{\boldsymbol{\theta}}_t^\top\boldsymbol{\phi}_t(i)\right) + \rho_{t,V}(i) \leq \ell\left(\boldsymbol{\theta}^{*\top}\boldsymbol{\phi}_t(i)\right) + 2\rho_{t,V}(i)$.*

Finally we can use the variance-agnostic upper confidence bound $\bar{\mu}_{t,i} := \ell\left(\hat{\boldsymbol{\theta}}_t^\top\boldsymbol{\phi}_t(i)\right) + \rho_{t,V}(i)$ as our optimistic estimation of the true mean $\mu_{t,i}$ to balance the exploration-exploitation tradeoff.

Variance-agnostic CLogUCB algorithm. Based on this exploration bonus, we introduce a simple yet efficient CLogUCB algorithm and prove the first regret bound for applications under the 1-norm TPM smoothness condition. In Line 4, we compute the covariance matrix $\boldsymbol{V}_t$ in order to compute the exploration bonus $\rho_{t,V}(i)$ defined as in Lemma 3. In Line 6, we construct an upper confidence bound $\bar{\mu}_{t,i}$ for each arm i based on Lemma 2, where $\ell\left(\hat{\boldsymbol{\theta}}_t^\top\boldsymbol{\phi}_t(i)\right)$ is the MLE estimation of $\mu_{t,i}$ and $\beta_t(\delta)\left\|\boldsymbol{\phi}_t(i)\right\|_{\boldsymbol{V}_t^{-1}}$ is the exploration bonus in the direction $\boldsymbol{\phi}_t(i)$. After computing the UCB values $\bar{\boldsymbol{\mu}}_t$, the learner selects action S_t through the offline oracle with $\bar{\boldsymbol{\mu}}_t$ as input. Then, the base arms in τ_t are triggered, and the learner receives the observation set $(X_{t,i})_{i\in\tau_t}$ as feedback to improve future decisions. Now we provide the regret upper bound for applications under the 1-norm TPM condition.

Theorem 1. *For a CLogB instance that satisfies monotonicity (Condition 1) and 1-norm TPM smoothness (Condition 2) with coefficient B_1, CLogUCB (Algorithm 1) with an α-approximation oracle achieves an α-approximate regret bounded by $O\left(B_1 d\sqrt{\kappa KT}\log(KT)\right)$.*

Representative applications and experiments. We show that our framework can cover a diverse range of application scenarios, including online content delivery, dynamic channel allocation, online packet routing, and online multi-LLM selection, which are detailed in Section 3 in the full paper [2]. We also validate our theoretical results through experiments on both synthetic and real-world datasets, demonstrating at least 53% regret improvement compared to baseline algorithms. See the full paper [2] for details.

Acknowledgement

The work of Xutong Liu was supported in part by a fellowship award from the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK PDFS2324-4S04). The work of John C.S. Lui was supported in part by the RGC GRF-14202923. The work is supported by the National Science Foundation under awards CNS-2102963, CAREER-2045641, CNS-2106299, CPS-2136199, and CNS-2325956. (*Corresponding author: Xuchuang Wang.*)

References

[1] Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. 2020. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*. PMLR, 3052–3060.
[2] Xutong Liu, Xiangxiang Dai, Xuchuang Wang, Mohammad Hajiesmaili, and John Lui. 2024. Combinatorial Logistic Bandits. *arXiv preprint arXiv:2410.17075* (2024).
[3] Xutong Liu, Jinhang Zuo, Siwei Wang, John CS Lui, Mohammad Hajiesmaili, Adam Wierman, and Wei Chen. 2023. Contextual combinatorial bandits with probabilistically triggered arms. In *International Conference on Machine Learning*. PMLR, 22559–22593.
[4] Qinshi Wang and Wei Chen. 2017. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*. 1161–1171.