Topic-Aware Contextual Cascading Bandits

Hyun-jun Choi[*]
Seoul National University
nschj1@snu.ac.kr

Taehyun Hwang[*]
Seoul National University
th.hwang@snu.ac.kr

Min-hwan Oh
Seoul National University
minoh@snu.ac.kr

Abstract

We study a contextual cascading bandit problem in which at most one item from each topic can appear in a cascade. Unlike standard cascading bandits, we allow an item's click probability to depend on both its own features and the ordered history of previously displayed topics, capturing order-sensitive user responses even when all displayed items belong to distinct topics. This order dependence makes cascade planning nontrivial, since item selection and ordering must be optimized jointly. To address the resulting planning challenge, we reformulate cascade construction as a Markov decision process whose state summarizes the previously selected topics. Based on this formulation, we develop an optimism-based algorithm and prove a regret upper bound of $\widetilde{\mathcal{O}}(\bar{p}^{\frac{K-1}{2}} d\sqrt{T})$, where d is the number of unknown parameters, T is the horizon, K is the cascade length, and $\bar{p} < 1$ upper bounds the no-click probability of an examined item. To the best of our knowledge, this is the first regret guarantee showing a decreasing dependence on K in the more general order-sensitive setting, without relying on permutation invariance or regularization that scales with the cascade length. We further provide a local regret lower bound showing that this decreasing dependence on K is intrinsic, and validate our theory through experiments with topic-order-dependent click probabilities.

1 Introduction

In news briefings or content recommendation lists with limited display slots, it is often desirable to present one representative item from each of several distinct topics rather than multiple items from the same topic. Users typically examine such ranked lists sequentially and may stop once they click an item of interest, a behavior naturally captured by the cascading bandit model [11, 17, 18]. Motivated by this setting, we study a contextual cascading bandit problem in which at most one item from each topic can appear in a cascade. However, simply imposing this no-topic-repetition constraint on standard contextual cascading bandits is insufficient: their click probabilities depend only on the current item and an unknown preference parameter, not on the topics shown earlier or their order [22, 35, 21, 23, 34, 31, 8, 9]. Consequently, they cannot capture order-dependent user responses even when all displayed items belong to distinct topics, motivating a click model that accounts for the ordered history of previously presented topics.

Related to such history-dependent user responses, prior work has modeled how previously selected or exposed items affect the value of subsequent recommendations or click behavior in various ways. For example, submodular topic-coverage approaches use the additional topic coverage contributed by previously selected items as features [33, 14], while relevance–diversity hybrid models and fatigue-aware click models incorporate diversity and repeated-exposure effects through dedicated structures [20, 5]. These approaches model the effect of history through specific notions such as coverage, diversity, or fatigue; in particular, submodular coverage-based methods assume that the corresponding features satisfy a diminishing-returns structure [14].

[*]Equal contribution

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

However, actual user responses need not always conform to such structural assumptions. The influence of previously exposed topics may be better represented through more general features, and not only which topics have appeared but also the order in which they were presented may affect the click probability of subsequent items. We therefore consider modeling the previous topic history through a feature representation and combining it with the current item's features to learn click probabilities. Motivated by this perspective, we propose *Topic-aware contextual Cascading Bandits* (TopicCB), in which an item's click probability depends jointly on its own features and the ordered history of previously displayed topics, allowing the learner to capture how preceding topics influence subsequent click probabilities.

This order dependence creates a new planning challenge even when the click-model parameters are known. In history-independent models, the no-topic-repetition constraint can be handled using item-level scores within each topic. In TopicCB, however, each choice changes the click probabilities at later positions, requiring item selection and ordering to be optimized jointly. Direct enumeration is therefore costly, as it repeatedly evaluates item combinations across possible topic sequences. The key observation is that, under no topic repetition, different item-level prefixes inducing the same ordered topic history lead to the same subsequent planning problem: the history determines both future click probabilities and the set of remaining topics. We exploit this structure by reformulating cascade construction as a finite-horizon Markov decision process (MDP), with the topic history as the state and items from unselected topics as actions. Bellman recursion then reuses the computation associated with each topic-history state, avoiding redundant item-level enumeration.

Another important question is how the cascade length K affects regret. Longer cascades increase the planning space, but they also reduce the probability of no click, potentially mitigating the effect of click-probability estimation errors on the expected reward. Recent work [9] established a leading regret term that is non-increasing in K for standard contextual cascading bandits, but its analysis relies on permutation invariance and exposure swapping, which do not apply to TopicCB because click probabilities are order-sensitive. Beyond this issue, several existing cascading-bandit analyses require regularization that scales with K to control uncertainty across multiple positions within a cascade [8, 9, 26], potentially inflating the confidence radius when the cascade length becomes large relative to the feature dimension.[2] This raises a natural question: *how does the cascade length K affect regret when click probabilities are order-sensitive?*

In this work, we address these challenges with `TopicCB-UCB`, an optimism-based algorithm for TopicCB that combines confidence-based learning with the MDP reformulation for efficient cascade planning. We prove a regret upper bound whose leading horizon-dependent term decreases with the cascade length K, even though click probabilities are order-sensitive. Our analysis preserves the selected order rather than relying on permutation invariance, and it avoids regularization that grows with K, thereby extending the guarantee to regimes where the cascade length is large relative to the feature dimension. Our main contributions are summarized as follows:

- We introduce *Topic-aware contextual Cascading Bandits* (TopicCB), a novel contextual cascading bandit model where an item's click probability depends on both its own features and the ordered history of previously displayed topics. By incorporating topic-history dependence, TopicCB captures richer and more realistic user preferences than standard cascading bandit models.
- We reformulate cascade construction as a finite-horizon MDP, where states represent ordered topic histories and actions correspond to item placements. Building on this formulation, we propose `TopicCB-UCB`, an optimism-based algorithm that computes optimistic values through Bellman backups and constructs a fixed cascade along the hypothetical no-click trajectory. This approach exploits the topic-history structure to avoid exhaustive enumeration of ordered item lists.
- We establish a regret upper bound of order $\widetilde{\mathcal{O}}(\bar{p}^{(K-1)/2} d\sqrt{T} + Kd^2 + d^2\kappa_\star)$, where d is the dimension of the unknown parameter, T is the horizon, $\bar{p} < 1$ upper bounds the no-click probability, and $\kappa_\star$ is a problem-dependent constant characterizing the inverse derivative of the logistic link function. More importantly, since $\bar{p} < 1$, the leading horizon-dependent

[2] Some existing analyses require $\lambda \geq K$ to control uncertainty across up to K positions in a cascade using the same Gram matrix constructed from observations collected before the current round. In logistic models [8, 26], the confidence radius typically scales as $\widetilde{O}(\sqrt{\lambda} + d\log T/\sqrt{\lambda})$, with the two terms balanced at $\lambda \asymp d\log T$. Thus, when $K \gtrsim d\log T$, the constraint $\lambda \geq K$ can introduce a $\sqrt{K}$ dependence. A similar issue arises in linear models [22] through the $\sqrt{\lambda}$ term.

term in our regret bound strictly decreases with the cascade length K through the factor $\bar{p}^{(K-1)/2}$. To the best of our knowledge, this is the first regret guarantee establishing such a strictly decreasing dependence on K in order-sensitive cascading bandits. The key technical ingredients are an order-preserving regret decomposition that retains both prefix and suffix no-click factors and auxiliary within-cascade information matrices that avoid K-dependent regularization, which may be of independent interest.

- We establish a local minimax regret lower bound for a subclass of TopicCB, matching the leading upper bound in its dependence on the parameter dimension d and horizon T, up to logarithmic and local problem-dependent factors. The lower bound also exhibits a decreasing no-click factor in K, supporting the intrinsic benefit of longer cascades in order-sensitive environments.
- We provide numerical experiments in environments where click probabilities depend on the order of previously displayed topics, demonstrating the effectiveness of `TopicCB-UCB` and the computational advantages of MDP-based planning over brute-force enumeration.

1.1 Related Work

Cascading bandits. The cascade model is a standard click model for online learning to rank, capturing position bias through sequential examination and first-click stopping [11]. Cascading bandits formulate this interaction as a bandit problem with partial feedback, where the learner recommends a ranked list and observes feedback only up to the first attractive item [17, 10]. Early works studied non-contextual cascading bandits and established regret guarantees for learning item attraction probabilities under cascade feedback [17, 10]. The framework has also been extended to combinatorial action spaces and more general semi-bandit-style feedback models [18, 32, 7]. Feature-based variants use item features to generalize across large item collections [35], and contextual combinatorial cascading bandits incorporate contextual information and combinatorial constraints [22]. Other works study richer click models and behavioral effects, including dependent-click models with multiple clicks [16], user abandonment [4], and Thompson-sampling methods under cascade feedback [34]. Minimax regret guarantees have been investigated to characterize the worst-case statistical difficulty of cascading bandits [31]. More recent works study contextual cascading bandits with richer assortment structures [8] and refined analyses of the cascade-length dependence [9], including exposure-swapping techniques that control the leading dependence on the cascade length.

Diversity-aware and topic-aware recommendation. Another line of work incorporates topic-level or diversity structure into recommendation and learning to rank. Diversified retrieval objectives promote non-redundant rankings through criteria such as maximal marginal relevance or ranked bandit formulations [6, 28, 29]. Contextual combinatorial bandits have been applied to diversified online recommendation [27], and linear submodular bandits model diversity through marginal gains in topic coverage [33]. Cascading linear submodular bandits further combine cascade feedback with diversified retrieval by defining item attractiveness through topic-coverage gains [14]. Related works also study relevance–diversity hybrid cascading models [20], diversity-aware choice models [3], and fatigue effects under click feedback [5].

Logistic and combinatorial bandits. Our click probabilities are parameterized through a link function, the analysis analysis builds on techniques from generalized linear and logistic bandits. For generalized linear bandits, confidence-set based optimistic algorithms extend linear bandit methods to nonlinear reward models [13]. For logistic bandits, refined analyses use the local curvature of the logistic link and concentration of the regularized maximum likelihood estimator [12, 2]. Our confidence-set construction uses likelihood-based confidence sequences for the constrained, unregularized maximum likelihood estimator [19], adapted to cascading feedback, where only the examined positions of each cascade are observed.

Our setting is also related to combinatorial semi-bandits and probabilistically triggered arms. The general combinatorial multi-armed bandit framework studies super-arm decisions that trigger multiple base-arm observations [7], while contextual combinatorial bandits use features for combinatorial recommendation [27]. Recent batch-size-independent analyses reduce the regret dependence on the number of selected or triggered arms [24], and combinatorial logistic bandits extend such questions to logistic reward models [26]. Matrix-shrinkage techniques have also been used to obtain sharper confidence or uncertainty control in related bandit and dynamic decision problems [15].

2 Preliminaries

Notation. For $N \in \mathbb{N}$, let $[N] := \{1, \dots, N\}$. For $n \geq r$, define $P(n, r) := n!/(n-r)!$ and $C(n, r) := \binom{n}{r}$, the numbers of ordered r-permutations and unordered r-combinations from n elements, respectively. Let $\mathcal{B}_d(S) := \{x \in \mathbb{R}^d : \|x\|_2 \leq S\}$. We use $\otimes$ for the Kronecker product, $\mathrm{vec}(A)$ for matrix vectorization, and $\|A\|_{\mathrm{F}}$ for the Frobenius norm.

Problem setting. We introduce **Topic**-aware contextual **C**ascading **B**andits (TopicCB), a novel contextual cascading bandit model where the click probability depends on the topic history of earlier positions *within* the cascade. Let $\mathcal{I} = [N]$ be the ground set of items. Each item $i \in \mathcal{I}$ has a topic $z_i \in \mathcal{Z}$, where $\mathcal{Z}$ is a finite set of topics; we write $Z := (z_i)_{i \in \mathcal{I}}$. At each round $t \in [T]$, the learner observes round-varying item features $X_t := \{x_{t,i}\}_{i \in \mathcal{I}} \subset \mathbb{R}^{d_a}$. Then, the learner selects a list of K items, called a cascade, with distinct items and distinct topics:

$$C_t = (i_{t,1}, \dots, i_{t,K}) \in \Pi, \quad \Pi := \{(i_1, \dots, i_K) : i_k \in \mathcal{I},\ i_k \neq i_\ell,\ z_{i_k} \neq z_{i_\ell} \text{ for } k \neq \ell\}.$$

Remark 1. *The no-topic-repetition constraint can be interpreted as a diversity requirement that prevents a cascade from allocating multiple positions to the same topic. Its restrictiveness depends on the granularity of the topic partition. If the constraint is too strong for a given application, the topic space can be refined so that closely related but distinct items are assigned to different topics and may therefore appear together in the same cascade. Thus, the constraint can be adjusted through the choice of topic granularity to reflect the desired level of diversity.*

Once the cascade $C_t = (i_{t,1}, \dots, i_{t,K})$ is presented to the user, the user examines its items sequentially from the first position and stops upon the first click or after reaching the end of the cascade. For each position $k \in [K]$, let $Y_{t,k} \in \{0, 1\}$ denote the click outcome of the item at position k, with $Y_{t,k} = 1$ indicating that the item would be clicked if examined. Define the stopping position as

$$\tau_t := \begin{cases} \min\{k \in [K] : Y_{t,k} = 1\}, & \text{if such a position exists,} \\ K + 1, & \text{otherwise.} \end{cases}$$

The learner observes $\mathbf{Y}_t := (Y_{t,1}, \dots, Y_{t,\min\{\tau_t, K\}})$, the partial binary feedback up to the stopping position. Feedback for positions $k > \min\{\tau_t, K\}$ remains *unobserved*.

Unlike standard cascading bandits [17, 22, 9], TopiCB allows an item's click probability to depend not only on its own feature but also on the topic history induced by items displayed earlier in the cascade. For a cascade $C = (i_1, \dots, i_K)$, define the *topic history* before position k by

$$z_{<1}(C) := \emptyset, \qquad z_{<k}(C) := (z_{i_1}, \dots, z_{i_{k-1}}) \quad \text{for } k \in \{2, \dots, K+1\}.$$

Let $\phi(z_{<k}(C)) \in \mathbb{R}^{d_h}$ be a known feature representation of this topic history. Let $\mathcal{H}_t$ denote the information available to the learner before choosing C_t, including the current item features X_t and the item topic labels. We assume that the conditional click probability at position k is given by

$$p_{t,k}(C; \Theta_\star) := \mathbb{P}\left(Y_{t,k} = 1 \mid \mathcal{H}_t, C_t = C\right) = \mu\left(\phi(z_{<k}(C))^\top \Theta_\star x_{t,i_k}\right), \tag{1}$$

where $\mu(u) = 1/(1 + \exp(-u))$ is the logistic function and $\Theta_\star \in \mathbb{R}^{d_h \times d_a}$ is an *unknown* time-invariant parameter. Conditional on $\mathcal{H}_t$ and C, the latent click outcomes $Y_{t,1}, \dots, Y_{t,K}$ are assumed to be independent Bernoulli random variables with means $p_{t,1}(C; \Theta_\star), \dots, p_{t,K}(C; \Theta_\star)$, respectively.

We define the expected reward of cascade C as the probability that at least one item in C is clicked, namely $f_t(C; \Theta_\star) := 1 - \prod_{k=1}^{K}(1 - p_{t,k}(C; \Theta_\star))$. Unlike in standard cascading bandits, $f_t(C; \Theta_\star)$ is generally *not permutation-invariant*. In particular, two cascades containing the same set of items may have different expected rewards if their topic orders differ.

The learner aims to maximize the cumulative expected reward, or equivalently to minimize the cumulative expected regret

$$\mathrm{Regret}_T^{\mathrm{TopicCB}} := \mathbb{E}\left[\sum_{t=1}^{T} \{f_t(C_t^\star; \Theta_\star) - f_t(C_t; \Theta_\star)\}\right], \tag{2}$$

where $C_t^\star \in \mathrm{argmax}_{C \in \Pi} f_t(C; \Theta_\star)$ is an optimal cascade at round t, and the expectation is taken over the randomness of the learner and the user feedback.

Remark 2. *TopicCB subsumes standard history-independent contextual cascading bandit models as a special case. In particular, suppose the topic-history representation is constant, i.e., $\phi(h) = c$ for every history h, for some fixed vector c. Then $p_{t,k}(C; \Theta_\star) = \mu\big(c^\top \Theta_\star x_{t,i_k}\big) = \mu\big(\theta_\star^\top x_{t,i_k}\big)$, where $\theta_\star := \Theta_\star^\top c$, so the click probability depends only on the current item feature and is independent of the preceding topic history. Consequently, the expected cascade reward becomes permutation-invariant, recovering the standard contextual cascading bandit setting. Thus, TopicCB can be viewed as a generalization that allows the effective preference parameter to vary with the ordered topic history.*

3 Main Results

3.1 MDP Reformulation for Order-Sensitive Planning

Order-sensitive planning challenge. We first discuss the planning challenge induced by topic-history-conditioned click probabilities. In standard contextual cascading bandits, each item's click probability depends only on its own feature. Hence, the expected reward is determined by the selected item set, and the ordering of those items does not change the expected reward. This structure allows optimistic planning to be reduced to selecting the top-K items according to item-level scores.

In TopicCB, this reduction is no longer valid. Since the click probability at position k depends on the topic history $z_{<k}(\cdot)$, the same item may have different click probabilities under different preceding topic orders. Thus, two cascades with the same item set can induce different expected rewards, and the learner must optimize over ordered item sequences rather than merely selecting K items. Even at the topic-order level, the search space has size at least $P(|\mathcal{Z}|, K)$, provided $K \le |\mathcal{Z}|$ and each topic contains at least one item.[3] Moreover, exposure-swapping arguments [8, 9] are not directly applicable, since swapping items changes the topic histories and the click probabilities of later positions.

Topic-history MDP. The key observation is that, under the no-topic-repetition constraint, the ordered topic history summarizes both the effect of previous exposures on future click probabilities and the set of topics that remain available for selection. Consequently, different item-level prefixes that induce the same topic history share the same remaining planning problem and need not be treated separately. This motivates a finite-horizon MDP with horizon K, where the state is the ordered topic history and the actions are items from topics not yet selected.

Fix a round t and condition on the item features X_t and topic labels Z. We define the round-t topic-history MDP as $\mathcal{M}_t^\star := (\mathcal{S}, \{\mathcal{A}(s)\}_{s\in\mathcal{S}}, P_t^\star, r, K, s_{\text{init}})$. The base state space is

$$\mathcal{S}_{\text{base}} := \{\emptyset\} \cup \bigcup_{m=1}^{K} \{(z_{i_1}, \dots, z_{i_m}) : (i_1, \dots, i_m) \in \Pi_m\},$$

where Π_m denotes the set of feasible length-m item history. Let $s_{\text{init}} := \emptyset$, introduce an absorbing click state s_{click}, and set $\mathcal{S} := \mathcal{S}_{\text{base}} \cup \{s_{\text{click}}\}$. The admissible action sets implement the no-topic-repetition constraint. For $s \in \mathcal{S}_{\text{base}}$ with $|s| < K$, define $\mathcal{A}(s) := \{i \in \mathcal{I} : z_i \notin s\}$. For terminal topic-history states $s \in \mathcal{S}_{\text{base}}$ with $|s| = K$, and for $s = s_{\text{click}}$, define $\mathcal{A}(s) := \emptyset$. The transition kernel and reward are defined using the topic-history-conditioned click probability in Eq. (1). For $s \in \mathcal{S}_{\text{base}}$ with $|s| < K$ and $a \in \mathcal{A}(s)$, let $p_t^\star(s, a) := \mu(\phi(s)^\top \Theta_\star x_{t,a})$, the probability that item a is clicked after topic history s. The transition kernel satisfies $P_t^\star(s_{\text{click}} \mid s, a) = p_t^\star(s, a)$ and $P_t^\star((s, z_a) \mid s, a) = 1 - p_t^\star(s, a)$. The realized reward is $r(s, a, s') := \mathbf{1}\{s' = s_{\text{click}}\}$. Thus, a click yields reward 1 and moves the process to s_{click}, while a non-click yields reward 0 and moves the process to the extended topic history (s, z_a).

A deterministic policy for $\mathcal{M}_t^\star$ is a sequence $\pi = (\pi_1, \dots, \pi_K)$, where $\pi_k(s) \in \mathcal{A}(s)$ for each feasible state s with $|s| = k - 1$. For such a policy, set $V_{t,k}^\pi(s_{\text{click}}) = 0$ for all k and $V_{t,K+1}^\pi(s) = 0$. For $k = K, \dots, 1$, define

$$V_{t,k}^\pi(s) = p_t^\star(s, \pi_k(s)) + \big(1 - p_t^\star(s, \pi_k(s))\big) V_{t,k+1}^\pi((s, z_{\pi_k(s)})),$$

for $s \in \mathcal{S}_{\text{base}}$ with $|s| = k - 1$. The optimal initial value is $V_{t,1}^\star(s_{\text{init}}) := \max_\pi V_{t,1}^\pi(s_{\text{init}})$, where the maximum is over deterministic policies.

[3] A feasible cascade is obtained by choosing an ordered K-tuple of distinct topics and then choosing one item from each chosen topic. For a topic order $(z_1, \dots, z_K)$, the number of compatible item cascades is $\prod_{k=1}^K n_{z_k}$, where $n_z := |\{i \in \mathcal{I} : z_i = z\}|$. Thus $|\Pi| = \sum_{(z_1,\dots,z_K)\in \text{Perm}(\mathcal{Z},K)} \prod_{k=1}^K n_{z_k}$, and $|\Pi| \ge P(|\mathcal{Z}|, K)$ whenever every topic is nonempty.

Algorithm 1 `TopicCB-UCB`

Inputs: confidence level $\delta \in (0,1)$, norm bound $S > 0$
Initialize $\hat{\theta}_1 = \mathbf{0}_d$ and set $\lambda = 1/(4S^2)$
for round $t = 1, \ldots,$ **do**
  Observe item features $X_t = \{x_{t,i}\}_{i \in \mathcal{I}}$
  Compute $\mathbf{H}_t(\hat{\theta}_t)$ from all previously observed feedback
  Compute $\{\widetilde{Q}_{t,k}\}_{k \in [K]}$ as described in Eq. (3) ▷ *Optimistic Bellman backup*
  Set $s_{t,1} = s_{\text{init}}$
  for $k = 1, \ldots, K$ **do** ▷ *No-click trajectory extraction*
    Set $a_{t,k} \in \operatorname{argmax}_{a \in \mathcal{A}(s_{t,k})} \widetilde{Q}_{t,k}(s_{t,k}, a)$ and $s_{t,k+1} = (s_{t,k}, z_{a_{t,k}})$
  end for
  Recommend $C_t = (a_{t,1}, \ldots, a_{t,K})$ and observe feedback $\{Y_{t,k}\}_{k \in [\min\{\tau_t, K\}]}$
  Compute $\hat{\theta}_{t+1} \in \operatorname{argmin}_{\theta \in \mathcal{B}_d(S)} \mathcal{L}_{t+1}(\theta)$
end for

We now connect this MDP policy to the original cascade decision. Although π is written as a policy, the learner in TopicCB must recommend a fixed cascade before observing any within-round feedback. This fixed cascade is obtained from the *hypothetical no-click trajectory of* π. Starting from $s_1^\pi = s_{\text{init}}$, define $a_k^\pi := \pi_k(s_k^\pi)$ and $s_{k+1}^\pi := (s_k^\pi, z_{a_k^\pi})$ for $k \in [K]$. Then π induces the cascade $C^\pi := (a_1^\pi, \ldots, a_K^\pi)$. Conversely, any feasible cascade $C = (i_1, \ldots, i_K) \in \Pi$ can be embedded into a deterministic policy by setting $\pi_k(s_k^C) = i_k$ along its no-click trajectory $s_1^C = s_{\text{init}}$, $s_{k+1}^C = (s_k^C, z_{i_k})$, and defining π_k arbitrarily on all other non-terminal feasible states. These off-trajectory choices do not affect the value from s_{init}.

Hence, the MDP is a planning device: *it exposes the dynamic-programming structure of ordered planning while preserving the original non-adaptive cascade decision*. The following proposition makes this equivalence precise. In particular, the regret measured in the topic-history MDP is exactly the original TopicCB regret in Eq. (2).

Proposition 1 (Equivalence of MDP and TopicCB regret)**.** *Consider any learner that, at each round t, selects a deterministic policy π_t in $\mathcal{M}_t^\star$ and recommends the induced cascade $C_t = C^{\pi_t}$. Define the auxiliary MDP regret by*

$$\text{Regret}_T^{\text{MDP}} := \mathbb{E}\left[\sum_{t=1}^{T} \left\{V_{t,1}^\star(s_{\text{init}}) - V_{t,1}^{\pi_t}(s_{\text{init}})\right\}\right].$$

Then $\text{Regret}_T^{\text{MDP}} = \text{Regret}_T^{\text{TopicCB}}$.

Remark 3 (Computational benefit of MDP planning)**.** *The MDP reformulation also gives a computational benefit for exact planning. To illustrate this, suppose $|\mathcal{Z}| = g$ and each topic contains $\bar{N}$ items. A brute-force search over cascades enumerates $P(g,K)\bar{N}^K$ ordered item lists and costs $\mathcal{O}(K)$ to evaluate the reward of each list, giving complexity $\mathcal{O}\big(KP(g,K)\bar{N}^K\big)$. By contrast, Bellman planning in the topic-history MDP performs $\bar{N}\sum_{\ell=1}^{K} P(g,\ell)$ state-action backups. This quantity lies between $\bar{N}P(g,K)$ and $K\bar{N}P(g,K)$. Therefore, compared with brute-force enumeration, the MDP reformulation saves a factor between $\bar{N}^{K-1}$ and $K\bar{N}^{K-1}$ in the item-choice component of planning. A detailed derivation is provided in Appendix B.*

3.2 Algorithm

We now present `TopicCB-UCB`, an optimism-based algorithm for TopicCB model based on the topic-history MDP. At each round, the algorithm constructs an ellipsoidal confidence set around the constrained logistic maximum likelihood estimate (MLE), computes optimistic click probabilities, performs an optimistic Bellman backup over topic-history states, and extracts a fixed cascade along the hypothetical no-click trajectory. Algorithm 1 presents the full procedure.

Linearized topic-history click model. The topic-history-conditioned click probability in Eq. (1) can be written in the MDP notation as $p_t^\star(s,a) = \mu(\phi(s)^\top \Theta_\star x_{t,a})$, where s is the topic history

before the current position and a is the selected item. For estimation, we vectorize the unknown matrix parameter by setting $\theta_\star := \mathrm{vec}(\Theta_\star) \in \mathbb{R}^d$, where $d = d_h d_a$, and define $\varphi_t(s,a) := x_{t,a} \otimes \phi(s) \in \mathbb{R}^d$. Then $p_t^\star(s,a) = \mu(\langle \varphi_t(s,a), \theta_\star \rangle)$. For any candidate parameter θ, write $p_t(s,a;\theta) := \mu(\langle \varphi_t(s,a), \theta \rangle)$.

Estimation and confidence set. Let $\mathcal{L}_t$ be the negative log-likelihood built from all feedback observed before round t. Equivalently, if $\mathcal{D}_t$ denotes the observed tuples $(\varphi_{t'}(s_{t',k}, a_{t',k}), Y_{t',k})$ for $t' < t$ and $k \le \min\{\tau_{t'}, K\}$, then

$$\mathcal{L}_t(\theta) := \sum_{(u,y)\in\mathcal{D}_t} \ell(y,u;\theta), \qquad \ell(y,u;\theta) := -y \log \mu(u^\top\theta) - (1-y)\log(1-\mu(u^\top\theta)).$$

The constrained MLE is $\hat{\theta}_t \in \mathrm{argmin}_{\theta\in\mathcal{B}_d(S)} \mathcal{L}_t(\theta)$. If there are multiple minimizers, select the one with minimum Euclidean norm. For $\lambda = 1/(4S^2)$, define

$$\mathbf{H}_t(\theta) := \sum_{(u,y)\in\mathcal{D}_t} \dot{\mu}(u^\top\theta) uu^\top + \lambda \mathbf{I}_d.$$

Using the ellipsoidal confidence construction of Lee et al. [19, Theorem 3.2], we define

$$\mathcal{E}_t := \left\{ \theta \in \mathbb{R}^d : \|\theta - \hat{\theta}_t\|^2_{\mathbf{H}_t(\hat{\theta}_t)} \le \gamma_t(\delta)^2 \right\},$$

where $\gamma_t(\delta) = \mathcal{O}(\sqrt{d\log(TK)})$ for $t \le T$, taking $\delta = 1/(T+1)$. Its exact definition is given in Appendix C.1. By Corollary C.2, this set contains $\theta_\star$ for all rounds t with probability at least $1-\delta$.

Optimistic Bellman backup. Given the confidence set $\mathcal{E}_t$, the algorithm computes optimistic values in the topic-history MDP. For each $s \in \mathcal{S}_{\mathrm{base}}$, let $|s|$ denote the length of the topic history. Set $\widetilde{V}_{t,K+1}(s) = 0$ for every terminal topic-history state with $|s| = K$. For $k = K, \dots, 1$, for every $s \in \mathcal{S}_{\mathrm{base}}$ with $|s| = k-1$ and $a \in \mathcal{A}(s)$, define the optimistic logit and the corresponding optimistic click probability by

$$U_{t,k}(s,a) := \langle \varphi_t(s,a), \hat{\theta}_t \rangle + \gamma_t(\delta) \|\varphi_t(s,a)\|_{\mathbf{H}_t(\hat{\theta}_t)^{-1}}, \qquad \widetilde{p}_{t,k}(s,a) := \mu(U_{t,k}(s,a)).$$

The optimistic action-value and value are then defined by

$$\widetilde{Q}_{t,k}(s,a) := \widetilde{p}_{t,k}(s,a) + \big(1 - \widetilde{p}_{t,k}(s,a)\big)\widetilde{V}_{t,k+1}((s,z_a)), \qquad \widetilde{V}_{t,k}(s) := \max_{a\in\mathcal{A}(s)} \widetilde{Q}_{t,k}(s,a). \quad (3)$$

Eq. (3) is the Bellman backup of the topic-history MDP with the true click probability replaced by its most optimistic plausible value. Since μ is monotone and $\widetilde{V}_{t,k+1}((s,z_a)) \in [0,1]$, this optimistic backup uses the maximum linear score over $\mathcal{E}_t$. The displayed bonus formula equals $\max_{\theta\in\mathcal{E}_t} \langle \varphi_t(s,a), \theta \rangle$, so the optimistic logit is available in closed form.

Cascade selection. After the optimistic Bellman backup, the algorithm extracts a fixed cascade along the hypothetical no-click trajectory. Starting from $s_{t,1} = s_{\mathrm{init}}$, for $k = 1, \dots, K$, it chooses

$$a_{t,k} \in \operatorname*{argmax}_{a\in\mathcal{A}(s_{t,k})} \widetilde{Q}_{t,k}(s_{t,k}, a), \qquad s_{t,k+1} := (s_{t,k}, z_{a_{t,k}}).$$

The recommended cascade is $C_t = (a_{t,1}, \dots, a_{t,K})$. After observing the partial feedback $\{Y_{t,k}\}_{k\le\min\{\tau_t,K\}}$, the algorithm updates the likelihood $\mathcal{L}_{t+1}$ and computes the next estimate $\hat{\theta}_{t+1}$.

3.3 Regret Bound

We now state the cumulative regret guarantee for Algorithm 1. We first introduce the boundedness assumption and the problem-dependent quantities that appear in the bound.

Assumption 1 (Boundedness)**.** *For all rounds $t \in [T]$ and items $i \in \mathcal{I}$, the item feature satisfies $\|x_{t,i}\|_2 \le 1$. For all topic-history states $s \in \mathcal{S}_{\mathrm{base}}$, the topic-history feature satisfies $\|\phi(s)\|_2 \le 1$. The unknown parameter satisfies $\|\Theta_\star\|_{\mathrm{F}} \le S$ for a known constant $S > 0$.*

Under Assumption 1, the linearized state-action feature $\varphi_t(s,a) := x_{t,a} \otimes \phi(s) \in \mathbb{R}^d$, with $d = d_h d_a$, satisfies $\|\varphi_t(s,a)\|_2 \le 1$ for every feasible state-action pair (s,a). Moreover, since $\|\mathrm{vec}(\Theta_\star)\|_2 = \|\Theta_\star\|_{\mathrm{F}}$, the vectorized parameter $\theta_\star := \mathrm{vec}(\Theta_\star)$ satisfies $\|\theta_\star\|_2 \le S$.

Definition 1 (Minimum click probability)**.** *For a horizon T, define the minimum click probability as $\underline{q} := \inf\{p_t^\star(s,a) : t \in [T],\ s \in \mathcal{S}_{\mathrm{base}},\ |s| < K,\ a \in \mathcal{A}(s)\}$. We also write $\bar{p} := 1 - \underline{q}$, which is an upper bound on the no-click probability of any feasible state-action pair.*

Definition 2 (Inverse link derivative)**.** *For a horizon T, define the inverse link-derivative constant over all feasible state-action pairs by $\kappa_\star := \sup_{t\in[T],\, s\in\mathcal{S}_{\mathrm{base}}:\, |s|<K,\, a\in\mathcal{A}(s)} 1/\dot{\mu}(\langle \varphi_t(s,a), \theta_\star\rangle)$.*

The constant $\bar{p}$ controls the probability that the user continues scanning the cascade, and hence appears through products of no-click probabilities. The constant $\kappa_\star$ controls the smallest local slope of the logistic link at the true logits; larger $\kappa_\star$ corresponds to flatter likelihood and harder estimation. Under Assumption 1, both constants are finite. Indeed, $|\langle \varphi_t(s,a), \theta_\star\rangle| \le S$ for every feasible (t,s,a), so $\underline{q} \ge \mu(-S) > 0$, $\bar{p} < 1$, and $\kappa_\star \le (\min_{|u|\le S} \dot{\mu}(u))^{-1} < \infty$.

We now state our main regret guarantee for Algorithm 1.

Theorem 1 (Regret upper bound for Algorithm 1)**.** *Suppose Assumption 1 holds. Run Algorithm 1 with $\lambda_t \equiv 1/(4S^2)$, $\delta = 1/(T+1)$, and the confidence radius $\gamma_t(\delta)$ defined in Appendix C.1. For fixed S, this radius satisfies $\gamma_t(\delta) = \mathcal{O}(\sqrt{d_h d_a \log(1+TK)})$ uniformly over $t \in [T]$, and the expected cumulative regret satisfies*

$$\mathrm{Regret}_T^{\mathrm{TopicCB}} = \widetilde{\mathcal{O}}\left(\bar{p}^{(K-1)/2} d_h d_a \sqrt{T} + K d_h^2 d_a^2 + d_h^2 d_a^2 \kappa_\star\right),$$

where logarithmic factors are suppressed and the implicit constants may depend on S.

Discussion of Theorem 1. The leading horizon-dependent term contains the factor $\bar{p}^{(K-1)/2}$, which strictly decreases with the cascade length K since $\bar{p} < 1$ under Assumption 1. This captures the cascading nature of the reward: *longer cascades reduce the probability that no click occurs throughout the cascade*. Recent work on standard contextual cascading bandits shows that the leading regret term can be non-increasing in the cascade length K [9]. However, that result relies on permutation invariance: because item click probabilities do not depend on preceding items, reordering a selected set does not change the expected reward, enabling an optimistic exposure swapping argument [8]. In TopicCB, this argument is unavailable because reordering items changes the topic history and hence the expected reward. Instead, we develop a refined regret analysis that yields the K-decreasing factor in the leading term. As noted in Remark 2, permutation-invariant cascading bandits are recovered as a special case when the click probability is independent of the topic history. Thus, our result establishes a decreasing-in-K leading regret term in the more general order-sensitive setting.

We next establish a local minimax lower bound that complements Theorem 1. The lower bound is established on a constructed subclass of TopicCB with fixed item features and a specially designed, known topic-history representation. Since this subclass is contained in the general contextual model, a lower bound for this subclass also applies to the general problem.

Theorem 2 (Local minimax lower bound)**.** *Fix $d_h \ge 1$, $d_a \ge 2$, and $K \ge \max\{1, d_h - 1\}$. For any fixed reference logit $b \in \mathbb{R}$, there exist a finite TopicCB instance satisfying Assumption 1, with item features fixed across rounds, and a reference parameter Θ_0 independent of T. For some positive sequence $\varepsilon_T \to 0$, define the Frobenius neighborhood $\mathcal{N}_T(\Theta_0) := \{\Theta : \|\Theta - \Theta_0\|_{\mathrm{F}} \le \varepsilon_T\}$. Then, for all sufficiently large T,*

$$\inf_{\mathsf{Alg}} \sup_{\Theta\in\mathcal{N}_T(\Theta_0)} \mathrm{Regret}_T^{\mathrm{TopicCB}}(\Theta) \ge c\, p_{\mathrm{loc}}^{K-1} d_h d_a \sqrt{\frac{T}{\kappa_\star(\Theta_0)}},$$

where $c > 0$ is a universal constant, $p_{\mathrm{loc}} := 1 - \mu(b + \log 2) \in (0,1)$, and $\kappa_\star(\Theta_0)$ is the inverse link-derivative constant at the reference parameter. The infimum is over all possibly randomized learning algorithms Alg.

Discussion of Theorem 2. The hard-instance construction exploits topic-history features to activate different rows of the unknown parameter matrix, embedding independent hidden signs across multiple rows. The proof builds on information-theoretic lower-bound techniques for logistic bandits [2], while requiring a topic-history-aware construction and a cascade-specific regret comparison.

Comparing Theorem 2 with Theorem 1, the leading upper-bound term matches the local minimax lower bound in its dependence on the dimension $d_h d_a$ and the horizon T, up to logarithmic and problem-dependent factors. Notably, both bounds involve no-click factors that decay exponentially with the cascade length K. A related decreasing-in-K lower-bound factor was established for standard contextual cascading bandits by Choi et al. [9]. Our result shows that a similar dependence persists despite the order sensitivity introduced by topic histories, consistent with the favorable K-dependence of our upper bound.

3.4 Regret Analysis

We now summarize the main technical ingredients behind the K-decreasing leading regret term for TopicCB. First, the optimistic Bellman backup makes order-sensitive optimistic planning tractable. Although the learner must optimize over ordered cascades, the topic-history MDP allows the algorithm to compute optimistic value functions recursively over feasible state-action pairs and then extract a fixed cascade along the hypothetical no-click trajectory. Thus, the algorithm obtains an optimistic cascade without brute-force enumeration of all ordered item lists, while still accounting for the effect of each chosen item on future topic-history-dependent click probabilities.

Second, we refine the regret decomposition to extract the no-click factor without exposure swapping [8, 9]. In standard cascading bandits, exposure swapping can move uncertain items to earlier positions because the expected reward is permutation-invariant. This improves the exposure of uncertain items without changing the cascade reward, and therefore avoids simply summing position-wise uncertainty over all K positions. In TopicCB, such an argument is invalid. Our analysis instead keeps the selected order fixed and decomposes how the local optimistic error at each position affects the cascade reward. For the cascade selected at round t, let $p^\star_{t,k}$ be the true click probability at position k, and let $\widetilde{p}_{t,k}$ be the optimistic click probability used in the Bellman backup. Lemma D.2 shows that the local error $\widetilde{p}_{t,k} - p^\star_{t,k}$ is weighted as

$$\left(\prod_{\ell<k}(1-p^\star_{t,\ell})\right)(\widetilde{p}_{t,k}-p^\star_{t,k})\left(\prod_{m>k}(1-p^\star_{t,m})\right).$$

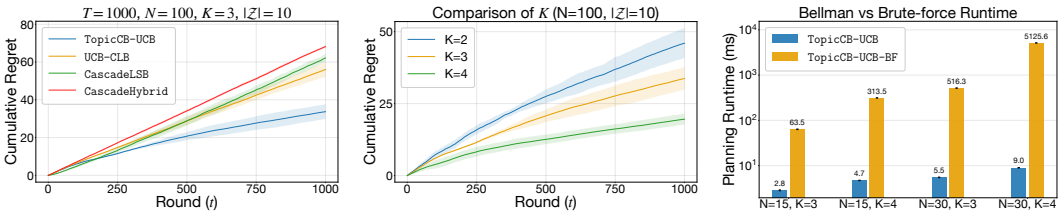
The suffix factor is crucial: if we only kept the prefix factor, the analysis would essentially add uncertainty over positions and lead to an undesirable K-dependence. Prior works [24, 26, 9] reduced the dependence on K by inserting local variance factor. However, previous analyses handle the suffix factor rather crudely. Since the suffix factor is less than 1, they upper bound the suffix factor by its square root. Our key difference is to split the suffix factor into two square-root factor, so both the prefix and suffix factor remain after Cauchy-Schwarz:

$$\prod_{\ell<k}(1-p^\star_{t,\ell})\prod_{m>k}(1-p^\star_{t,m}) = \prod_{j\neq k}(1-p^\star_{t,j}) \le \bar{p}^{K-1}.$$

Thus, instead of adding K position-wise errors at full scale, the analysis multiplies each local error by the probability that the other positions do not click. After combining this decomposition with the local click-probability error bound, the weighted cascade-potential argument applies Cauchy–Schwarz, which turns this product into the factor $\bar{p}^{(K-1)/2}$ in the leading regret term.

Third, we avoid the cascade-length-dependent regularization that often appears in analyses of delayed combinatorial feedback. In existing combinatorial and cascading bandit analyses [27, 22, 8, 9, 26], all K positions in a cascade are often controlled using the same pre-round information matrix. To apply an elliptical-potential argument directly to this length-K block, these analyses typically require a regularization level satisfying $\lambda \ge K$. Since λ enters the confidence radius, this leads to a radius scaling as $\sqrt{\max\{K, d\log T\}}$. Thus, the confidence radius has no additional polynomial dependence on K only when $K \lesssim d\log T$; otherwise, it scales as $\sqrt{K}$.

In contrast, our proof introduces auxiliary within-cascade information matrices only for the analysis. These matrices compare the pre-round uncertainty with the uncertainty that would have been accumulated from earlier examined positions in the same planned cascade, allowing a sequential elliptical-potential argument without requiring $\lambda \ge K$. This permits the fixed choice $\lambda = 1/(4S^2)$. For fixed S and $\delta = 1/(T+1)$, the resulting confidence radius satisfies $\gamma_t(\delta) = \mathcal{O}(\sqrt{d\log(1+TK)})$ uniformly over $t \in [T]$, so its dependence on K is only logarithmic.



(a) Regret comparison (b) Effect of cascade length K (c) Planning runtime

Figure 1: Experimental results in a TopicCB environment with $N = 100$ items and $|\mathcal{Z}| = 10$ topics. (1a) Cumulative regret of `TopicCB-UCB` and baselines over $T = 1000$ rounds. (1b) Cumulative regret of `TopicCB-UCB` under different cascade lengths $K \in \{2, 3, 4\}$. (1c) Per-round runtime comparison between Bellman-backup and brute-force planning in `TopicCB-UCB`. The bar heights are plotted on a logarithmic scale.

4 Numerical Experiments

We evaluate `TopicCB-UCB` in synthetic environments with $\Theta_\star \in \mathbb{R}^{d_h \times d_a}$, where $d_s = d_h = 3$ and $\|\Theta_\star\|_{\mathrm{F}} \leq 1$. The environments are order-sensitive to previously observed topics; details are given in Appendix G. Since TopicCB has topic-history-dependent logistic click probabilities, no existing baseline directly matches our setting. We compare `TopicCB-UCB` with three representative baselines: `UCB-CLB` [9] for item-feature logistic learning, `CascadeLSB` [14] for topic-coverage learning, and `CascadeHybrid` [20] for hybrid item-plus-topic learning.

Regret comparison & K-ablation test. Figure 1a compares `TopicCB-UCB` with the three baselines. `TopicCB-UCB` achieves much smaller cumulative regret throughout the horizon T, while the baselines increase almost linearly. This shows the benefit of modeling topic-history-dependent click probabilities. Figure 1b reports the cumulative regret of `TopicCB-UCB` for $K \in \{2, 3, 4\}$. The regret clearly decreases as cascade length K increases, supporting our theoretical claim.

Computational cost of planning. Figure 1c compares `TopicCB-UCB` with `TopicCB-UCB-BF`, which replaces only the action-plan selection step with brute-force enumeration. We vary $N \in \{15, 30\}$ and $K \in \{3, 4\}$ with $|\mathcal{Z}| = 4$. Brute-force planning grows rapidly with N and K, whereas `TopicCB-UCB` remains efficient. This agrees with Remark 3, which predicts a saving of order $\bar{N}^{K-1}$. When N doubles from 15 to 30, the observed speedup increases by $4.2\times$ for $K = 3$ and $8.6\times$ for $K = 4$, close to the predicted factors 4 and 8, respectively.

5 Conclusion

In this work, we introduced TopicCB, a contextual cascading bandit model with topic-history-dependent click probabilities. This model captures order-sensitive user preferences beyond standard contextual cascading bandits, but also creates a new planning challenge because the optimal cascade is no longer determined solely by the selected item set. We addressed this challenge by reformulating cascade construction as a finite-horizon topic-history MDP and proposing `TopicCB-UCB`, which computes optimistic value functions by Bellman backups and extracts a fixed cascade along the hypothetical no-click trajectory.

We proved a regret upper bound whose leading term contains the strictly decreasing no-click factor $\bar{p}^{(K-1)/2}$, showing that the beneficial effect of longer cascades persists even in the order-sensitive setting where exposure swapping is unavailable. Our analysis avoids regularization that scales with the cascade length and isolates the delayed-feedback cost into lower-order additive terms. We also provided a local minimax lower bound showing that the leading dimension and horizon dependence is essentially tight. Finally, numerical experiments support the benefit of topic-history modeling, the decrease of regret with K, and the computational advantage of MDP-based planning.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant both funded by the Korea government (MSIT) (No. RS-2022-NR071853, RS-2023- 00222663, RS-2025-25463302, RS-2026-25507282, RS-2026-25613008).

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

[2] Marc Abeille, Louis Faury, and Clément Calauzènes. Instance-wise minimax-optimal algorithms for logistic bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 3691–3699. PMLR, 2021.

[3] Heesang Ann, Taehyun Hwang, and Min hwan Oh. Diversified multinomial logit contextual bandits. In *The Fourteenth International Conference on Learning Representations*, 2026.

[4] Junyu Cao, Wei Sun, and Zuo-Jun Max Shen. Doubly adaptive cascading bandits with user abandonment. *Available at SSRN 3355211*, 2019.

[5] Junyu Cao, Wei Sun, Zuo-Jun Max Shen, and Markus Ettl. Fatigue-aware bandits for dependent click models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3341–3348, 2020.

[6] Jaime Carbonell and Jade Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336, 1998.

[7] Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17 (50):1–33, 2016.

[8] Hyunjun Choi, Rajan Udwani, and Min-hwan Oh. Cascading contextual assortment bandits. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[9] Hyunjun Choi, Joongkyu Lee, and Min hwan Oh. True impact of cascade length in contextual cascading bandits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[10] Richard Combes, Stefan Magureanu, Alexandre Proutiere, and Cyrille Laroche. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS international conference on measurement and modeling of computer systems*, pages 231–244, 2015.

[11] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*, pages 87–94, 2008.

[12] Louis Faury, Marc Abeille, Clément Calauzènes, and Olivier Fercoq. Improved optimistic algorithms for logistic bandits. In *International Conference on Machine Learning*, pages 3052–3060. PMLR, 2020.

[13] Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, page 586–594, Red Hook, NY, USA, 2010. Curran Associates Inc.

[14] Gaurush Hiranandani, Harvineet Singh, Prakhar Gupta, Iftikhar Ahamath Burhanuddin, Zheng Wen, and Branislav Kveton. Cascading linear submodular bandits: Accounting for position bias and diversity in online learning to rank. In *Uncertainty in Artificial Intelligence*, pages 722–732. PMLR, 2020.

[15] Rong Jin, David Simchi-Levi, Li Wang, Xinshang Wang, and Sen Yang. Shrinking the upper confidence bound: A dynamic product selection problem for urban warehouses. *Management Science*, 67(8):4756–4771, 2021.

[16] Sumeet Katariya, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. Dcm bandits: Learning to rank with multiple clicks. In *International Conference on Machine Learning*, pages 1215–1224. PMLR, 2016.

[17] Branislav Kveton, Csaba Szepesvari, Zheng Wen, and Azin Ashkan. Cascading bandits: Learning to rank in the cascade model. In *International Conference on Machine Learning*, pages 767–776, 2015.

[18] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Combinatorial cascading bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

[19] Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[20] Chang Li, Haoyun Feng, and Maarten de 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.

[21] Shuai Li and Shengyu Zhang. Online clustering of contextual cascading bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[22] Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. Contextual combinatorial cascading bandits. In *International conference on machine learning*, pages 1245–1253. PMLR, 2016.

[23] Weiwen Liu, Shuai Li, and Shengyu Zhang. Contextual dependent click bandit algorithm for web recommendation. In *International computing and combinatorics conference*, pages 39–50. Springer, 2018.

[24] Xutong Liu, Jinhang Zuo, Siwei Wang, Carlee Joe-Wong, John Lui, and Wei Chen. Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms. *Advances in Neural Information Processing Systems*, 35:14904–14916, 2022.

[25] Xutong Liu, Jinhang Zuo, Siwei Wang, John CS Lui, Mohammad Hajiesmaili, Adam Wierman, and Wei Chen. Contextual combinatorial bandits with probabilistically triggered arms. In *International Conference on Machine Learning*, pages 22559–22593. PMLR, 2023.

[26] Xutong Liu, Xiangxiang Dai, Xuchuang Wang, Mohammad Hajiesmaili, and John CS Lui. Combinatorial logistic bandits. In *Abstracts of the 2025 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 112–114, 2025.

[27] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

[28] Filip Radlinski, Robert Kleinberg, and Thorsten Joachims. Learning diverse rankings with multi-armed bandits. In *Proceedings of the 25th international conference on Machine learning*, pages 784–791, 2008.

[29] Karthik Raman, Pannaga Shivaswamy, and Thorsten Joachims. Online learning to diversify from implicit feedback. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 705–713, 2012.

[30] Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*, volume 11. Springer, 2009.

[31] Daniel Vial, Sujay Sanghavi, Sanjay Shakkottai, and Rayadurgam Srikant. Minimax regret for cascading bandits. *Advances in Neural Information Processing Systems*, 35:29126–29138, 2022.

[32] Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 1113–1122, 2015.

[33] Yisong Yue and Carlos Guestrin. Linear submodular bandits and their application to diversified retrieval. *Advances in neural information processing systems*, 24, 2011.

[34] Zixin Zhong, Wang Chi Chueng, and Vincent YF Tan. Thompson sampling algorithms for cascading bandits. *Journal of Machine Learning Research*, 22(218):1–66, 2021.

[35] Shi Zong, Hao Ni, Kenny Sung, Nan Rosemary Ke, Zheng Wen, and Branislav Kveton. Cascading bandits for large-scale recommendation problems. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 835–844, 2016.

Contents of Appendix

A Equivalence of the MDP and Cascade Formulations **14**

A.1 Proof of Proposition 1 15

B Computation Benefit of MDP Planning (Remark 3) **15**

C Concentration Bounds **17**

C.1 Notations for Concentration Bound 17

C.2 Likelihood confidence and estimation error bounds 18

C.3 Proofs of Corollaries 18

C.3.1 Proof of Corollary C.1 18

C.3.2 Proof of Corollary C.2 18

C.3.3 Proof of Corollary C.3 19

D Regret Upper Bound **19**

D.1 Lemmas used in the regret analysis 19

D.2 Proof of Theorem 1 21

D.3 Proofs of Lemmas 24

D.3.1 Proof of Lemma D.1 24

D.3.2 Proof of Lemma D.2 24

D.3.3 Proof of Lemma D.3 25

D.3.4 Proof of Lemma D.4 25

D.3.5 Proof of Lemma D.5 25

D.3.6 Proof of Corollary D.1 26

E Local Minimax Lower Bound **26**

E.1 Hard-Instance Construction 26

E.2 Main Proof of Theorem 27

E.3 Proofs of Lemmas for Theorem 2 29

E.3.1 Proof of Lemma E.1 29

E.3.2 Proof of Lemma E.2 30

E.3.3 Proof of Corollary E.1 30

E.3.4 Proof of Lemma E.3 31

F Auxiliary Lemmas **32**

G Details of Experiments **33**

A Equivalence of the MDP and Cascade Formulations

In this section, we show that our MDP reformulation is only a planning device and does not change the learning objective through Proposition 1. The proof follows by identifying each no-click action plan with its induced cascade and showing that their values are equal.

Proposition 1 (Equivalence of MDP and TopicCB regret)**.** *Consider any learner that, at each round t, selects a deterministic policy π_t in $\mathcal{M}_t^\star$ and recommends the induced cascade $C_t = C^{\pi_t}$. Define the auxiliary MDP regret by*

$$\mathrm{Regret}_T^{\mathrm{MDP}} := \mathbb{E}\left[\sum_{t=1}^{T}\left\{V_{t,1}^\star(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}})\right\}\right].$$

Then $\mathrm{Regret}_T^{\mathrm{MDP}} = \mathrm{Regret}_T^{\mathrm{TopicCB}}$.

A.1 Proof of Proposition 1

Proof of Proposition 1. Fix a round t. We condition on X_t, the topic labels Z, and the learner's selected plan π_t.

Let $\pi = (a_1, \ldots, a_K)$ be any non adaptive action plan. Define its induced hypothetical no-click trajectory by

$$s_1^\pi = s_{\mathrm{init}}, \qquad s_{k+1}^\pi = (s_k^\pi, z_{a_k}), \quad k \in [K].$$

The induced cascade is $C^\pi = (a_1, \ldots, a_K)$. Since each action is feasible at the corresponding topic history, $C^\pi \in \Pi$. Conversely, every feasible cascade $C = (i_1, \ldots, i_K) \in \Pi$ defines a no-click action plan by taking $a_k = i_k$. Hence the non-adaptive action plans and feasible cascades are in one-to-one correspondence.

For a fixed plan π, the MDP reaches s_k^π if and only if no click occurs in the first $k-1$ positions. Therefore,

$$V_{t,1}^\pi(s_{\mathrm{init}}) = 1 - \prod_{k=1}^{K}\left(1 - p_t(s_k^\pi, a_k; \Theta_\star)\right).$$

Since $s_k^\pi = z_{<k}(C^\pi)$, we have $p_t(s_k^\pi, a_k; \Theta_\star) = p_{t,k}(C^\pi; \Theta_\star)$. Thus,

$$V_{t,1}^\pi(s_{\mathrm{init}}) = 1 - \prod_{k=1}^{K}\left(1 - p_{t,k}(C^\pi; \Theta_\star)\right) = f_t(C^\pi; \Theta_\star).$$

Taking the maximum over no-click action plans gives

$$V_{t,1}^\star(s_{\mathrm{init}}) = \max_\pi V_{t,1}^\pi(s_{\mathrm{init}}) = \max_{C \in \Pi} f_t(C; \Theta_\star) = f_t(C_t^\star; \Theta_\star).$$

For the learner's selected plan π_t, with induced cascade $C_t = C^{\pi_t}$, the same identity gives

$$V_{t,1}^{\pi_t}(s_{\mathrm{init}}) = f_t(C_t; \Theta_\star).$$

Therefore, for every t,

$$V_{t,1}^\star(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}}) = f_t(C_t^\star; \Theta_\star) - f_t(C_t; \Theta_\star).$$

Summing over $t \in [T]$ and taking expectation yields our claim. □

B Computation Benefit of MDP Planning (Remark 3)

We provide a simple counting argument that explains the computational benefit of the MDP reformulation for exact planning. This comparison isolates the combinatorial planning cost and treats the optimistic score computation for each state-action pair as one oracle call. The statistical estimation cost is not included in this comparison.

Suppose there are $g := |\mathcal{Z}|$ topics and each topic contains exactly $\bar{N}$ items. The no-topic-repetition constraint implies that a feasible cascade is obtained by first choosing an ordered K-tuple of distinct topics and then choosing one item from each selected topic.

We first analyze the computational cost of brute-force search. The number of possible topic orders is $P(g, K) = g!/(g-K)!$. For each topic order, there are $\bar{N}^K$ item choices. Hence the number of feasible ordered cascades is $P(g, K)\bar{N}^K$. Evaluating the reward of one cascade requires computing

$$f_t(C; \Theta_\star) = 1 - \prod_{k=1}^{K}(1 - p_{t,k}(C; \Theta_\star)),$$

which costs $\mathcal{O}(K)$. Therefore, brute-force search costs

$$A := \mathcal{O}\big(KP(g,K)\bar{N}^K\big)\,.$$

We next analyze the computational cost of Bellman backups in the topic-history MDP. In the MDP formulation, a state at depth k is an ordered topic history of length k. Thus the number of non-terminal states at depth k is $P(g,k)$, for $k = 0,\dots,K-1$. At such a state, $g-k$ topics remain available, and each remaining topic contains $\bar{N}$ items. Hence the number of admissible actions at depth k is $(g-k)\bar{N}$. The Bellman backup for a state-action pair (s,a) only needs the value of the unique no-click successor (s,z_a); the click successor is the absorbing state and contributes no continuation value. Therefore, the number of state-action backups is

$$B := \bar{N}\sum_{k=0}^{K-1} P(g,k)(g-k).$$

Using $P(g,k)(g-k) = P(g,k+1)$, this becomes

$$B = \bar{N}\sum_{\ell=1}^{K} P(g,\ell).$$

To compare this with brute-force search, factor out $P(g,K)$:

$$\begin{aligned} B &= \bar{N}P(g,K)\sum_{\ell=1}^{K}\frac{P(g,\ell)}{P(g,K)} \\ &= \bar{N}P(g,K)\sum_{j=0}^{K-1}\frac{(g-K)!}{(g-K+j)!}. \end{aligned}$$

Define

$$S_{g,K} := \sum_{j=0}^{K-1}\frac{(g-K)!}{(g-K+j)!}.$$

Since the first term of $S_{g,K}$ is 1 and all K terms are at most 1, we have

$$1 \le S_{g,K} \le K.$$

Therefore,

$$\bar{N}P(g,K) \le B \le K\bar{N}P(g,K).$$

We now compare the two planning costs. Since

$$A = K\bar{N}^K P(g,K), \qquad B = \bar{N}P(g,K)S_{g,K},$$

we obtain

$$\frac{A}{B} = \frac{K\bar{N}^{K-1}}{S_{g,K}}.$$

Using $1 \le S_{g,K} \le K$, this yields

$$\bar{N}^{K-1} \le \frac{A}{B} \le K\bar{N}^{K-1}.$$

Thus, compared with brute-force enumeration, Bellman planning saves at least a factor of order $\bar{N}^{K-1}$ in the item-choice component of planning.

This saving comes from dynamic programming. A brute-force search repeatedly evaluates all item completions of each topic history. In contrast, the Bellman backup stores the value of each topic-history suffix once, so all downstream choices after a state are summarized by a single continuation value. Thus the MDP reformulation removes the repeated enumeration over item choices in the suffix, while still accounting for the order dependence of topic histories.

C Concentration Bounds

C.1 Notations for Concentration Bound

We first establish the concentration guarantee for the constrained MLE used in Algorithm 1. Recall that the click probability is modeled as

$$p_t(s,a;\theta_\star) = \mu\big(\varphi_t(s,a)^\top \theta_\star\big)\,, \qquad \varphi_t(s,a) := x_{t,a} \otimes \phi(s) \in \mathbb{R}^d,$$

where $x_{t,a} \in \mathbb{R}^{d_a}$, $\phi(s) \in \mathbb{R}^{d_h}$, and $d = d_h d_a$. For any candidate parameter $\theta \in \mathbb{R}^d$, write

$$p_t(s,a;\theta) := \mu\big(\varphi_t(s,a)^\top \theta\big)\,, \qquad p_t^\star(s,a) := p_t(s,a;\theta_\star).$$

Throughout the analysis, $\min\{\tau_t, K\}$ denotes the number of observed positions in episode t. Thus the observed tuples before episode t is

$$\mathcal{D}_t := \big\{\big(\varphi_{t'}(s_{t',k}, a_{t',k}), r_{t',k}\big) : 1 \le t' < t,\ 1 \le k \le \min\{\tau_{t'}, K\}\big\}\,.$$

For $u \in \mathbb{R}^d$, $y \in \{0,1\}$, define

$$\ell(y,u;\theta) := -y \log \mu(u^\top\theta) - (1-y)\log\big(1-\mu(u^\top\theta)\big).$$

The negative log-likelihood and constrained MLE are

$$\mathcal{L}_t(\theta) := \sum_{(u,y)\in\mathcal{D}_t} \ell(y,u;\theta), \qquad \hat\theta_t \in \arg\min_{\theta\in\mathcal{B}_d(S)} \mathcal{L}_t(\theta).$$

Since $\mathcal{L}_t$ is continuous and $\mathcal{B}_d(S)$ is compact, a minimizer exists. If there are multiple minimizers, let $\hat\theta_t$ be the one with minimum Euclidean norm. This choice is unique because the minimizer set is compact and convex and the squared Euclidean norm is strictly convex. The resulting estimator is measurable with respect to the observed data.

For $\theta \in \mathcal{B}_d(S)$, define the hessian matrix

$$\mathbf{H}_t(\theta) := \nabla^2\mathcal{L}_t(\theta) + \lambda\mathbf{I}_d = \sum_{(u,y)\in\mathcal{H}_t} \dot\mu(u^\top\theta) uu^\top + \lambda\mathbf{I}_d, \quad \lambda = \frac{1}{4S^2}$$

For $\delta \in (0,1]$, let

$$\beta_t(\delta)^2 = \log\frac{1}{\delta} + d\log\left(e \vee \frac{2eS(t-1)K}{d}\right),$$

$$\gamma_t(\delta)^2 = 2(1+S)\{4S^2\lambda + \beta_t(\delta)^2\} = 2(1+S)\{1+\beta_t(\delta)^2\}, \tag{4}$$

$$\Gamma_t(\delta)^2 = 2(1+2S)^2\beta_t(\delta)^2 + 4S^2\lambda = 2(1+2S)^2\beta_t(\delta)^2 + 1. \tag{5}$$

The radius $\gamma_t(\delta)$ is used in the optimistic score; $\Gamma_t(\delta)$ is used only in the regret analysis.

The confidence set is

$$\mathcal{E}_t = \left\{\theta \in \mathbb{R}^d : \|\theta - \hat\theta_t\|^2_{H_t(\hat\theta_t)} \le \gamma_t(\delta)^2\right\}, \tag{6}$$

and the optimistic score is

$$U_{t,k}(s,a) = \varphi_t(s,a)^\top\hat\theta_t + \gamma_t(\delta)\|\varphi_t(s,a)\|_{\mathbf{H}_t(\hat\theta_t)^{-1}}, \qquad \widetilde{p}_{t,k}(s,a) = \mu(U_{t,k}(s,a)). \tag{7}$$

Here $\mathcal{E}_t$ is the full ellipsoid in $\mathbb{R}^d$, without an intersection with $\mathcal{B}_d(S)$. Consequently $U_{t,k}(s,a) = \max_{\theta\in\mathcal{E}_t}\varphi_t(s,a)^\top\theta$, exactly as in the closed-form construction of [19, Eq. (4)]. The optimistic Bellman recursion in Eq. 3 uses these click probabilities, and the selected cascade follows its maximizing hypothetical no-click trajectory.

For the selected cascade π_t, let $s_{t,k}^{\pi_t}$ denote the state at position k along its hypothetical no-click trajectory. For every $k \in [K]$, write

$$\varphi_{t,k}^{\pi_t} := \varphi_t(s_{t,k}^{\pi_t}, a_{t,k}), \qquad p_{t,k}^\star := p_t^\star(s_{t,k}^{\pi_t}, a_{t,k}), \qquad \widetilde{p}_{t,k} := \widetilde{p}_{t,k}(s_{t,k}^{\pi_t}, a_{t,k}).$$

C.2 Likelihood confidence and estimation error bounds

Corollary C.1 (Likelihood confidence under cascading feedback)**.** *Under the boundedness assumptions* $\|\theta_\star\|_2 \le S$ *and* $\|\varphi_t(s,a)\|_2 \le 1$*, define*

$$\mathcal{G}_\delta = \left\{\forall t \in [T] : \mathcal{L}_t(\theta_\star) - \mathcal{L}_t(\hat{\theta}_t) \le \beta_t(\delta)^2\right\}. \tag{8}$$

Then $\mathbb{P}(\mathcal{G}_\delta) \ge 1-\delta$.

This follows by applying [19, Theorem 3.1] to the observed cascade feedback. We use an ordering argument analogous to [26, Appendix D.1].

Corollary C.2 (Estimation error in the Hessian norm at the MLE)**.** *On* $\mathcal{G}_\delta$*, simultaneously for every* $t \in [T]$*,*

$$\|\hat{\theta}_t - \theta_\star\|^2_{\mathbf{H}_t(\hat{\theta}_t)} \le \gamma_t(\delta)^2. \tag{9}$$

This is [19, Theorem 3.2] with logistic self-concordance constant $R_s = 1$.

Corollary C.3 (Estimation error in the Hessian norm at the true parameter)**.** *On* $\mathcal{G}_\delta$*, simultaneously for every* $t \in [T]$*,*

$$\|\hat{\theta}_t - \theta_\star\|^2_{\mathbf{H}_t(\theta_\star)} \le \Gamma_t(\delta)^2. \tag{10}$$

The proof uses the constrained MLE condition and the logistic derivative comparison in Auxiliary Lemma G.2.

C.3 Proofs of Corollaries

C.3.1 Proof of Corollary C.1

Proof. Following the lexicographic ordering argument of [26, Appendix D.1], list the actually observed tuples in lexicographic order of the round and the position within that round. At a reached position k, the feature $\varphi^{\pi_t}_{t,k}$ is known before $Y_{t,k}$, and

$$\mathbb{P}(Y_{t,k} = 1 \mid \mathcal{H}_t, \pi_t, Y_{t,1}, \dots, Y_{t,k-1}) = \mu((\varphi^{\pi_t}_{t,k})^\top \theta_\star).$$

Whether a position is observed depends only on earlier outcomes, not on $Y_{t,k}$. The ordered observations therefore meet the predictable-feature logistic assumptions of [19, Theorem 3.1]. Their likelihood is exactly $\mathcal{L}_t$ at the end of round $t-1$, and their constrained MLE has the same minimum likelihood value as $\hat{\theta}_t$.

For a prefix of n observed tuples, the Lipschitz constant of its loss on $\mathcal{B}_d(S)$ is at most n, since

$$\left\| \sum_{(u,y) \text{ in the prefix}} (\mu(u^\top\theta) - y)u \right\|_2 \le \sum_{(u,y) \text{ in the prefix}} |\mu(u^\top\theta) - y| \|u\|_2 \le n.$$

The source theorem is simultaneous over all observation counts. Thus it also holds at the random count $\sum_{s<t} \min\{\tau_s, K\}$, without a union bound over rounds. Since $\sum_{s<t} \min\{\tau_s, K\} \le (t-1)K$, with probability at least $1-\delta$, simultaneously for every $t \in [T]$,

$$\begin{aligned}
\mathcal{L}_t(\theta_\star) - \mathcal{L}_t(\hat{\theta}_t) &\le \log\frac{1}{\delta} + d\log\left(e \vee \frac{2eS\sum_{s<t}\min\{\tau_s, K\}}{d}\right) \\
&\le \log\frac{1}{\delta} + d\log\left(e \vee \frac{2eS(t-1)K}{d}\right) = \beta_t(\delta)^2.
\end{aligned}$$

Thus $\mathbb{P}(\mathcal{G}_\delta) \ge 1-\delta$. □

C.3.2 Proof of Corollary C.2

Proof. The parameter set $\mathcal{B}_d(S)$ is compact and convex, $\hat{\theta}_t$ is its unregularized likelihood minimizer, and $|\ddot{\mu}| \le \dot{\mu}$ gives $R_s = 1$. The likelihood-to-ellipsoid inclusion in [19, Theorem 3.2 and Appendix D], applied to the observations in Corollary C.1, yields

$$\|\hat{\theta}_t - \theta_\star\|^2_{\nabla^2\mathcal{L}_t(\hat{\theta}_t)+\lambda I_d} \le 2(1+S)\{4S^2\lambda + \beta_t(\delta)^2\} = \gamma_t(\delta)^2.$$

This inclusion uses constrained first-order optimality, and does not require a unique MLE or a zero likelihood gradient at the boundary. □

C.3.3 Proof of Corollary C.3

Proof. Fix $t \in [T]$ and work on $\mathcal{G}_\delta$. Since $\theta_\star \in \mathcal{B}_d(S)$ and $\hat{\theta}_t$ minimizes $\mathcal{L}_t$ over $\mathcal{B}_d(S)$, the first-order optimality condition gives

$$\nabla \mathcal{L}_t(\hat{\theta}_t)^\top (\theta_\star - \hat{\theta}_t) \geq 0.$$

Taylor's theorem and $\mathcal{G}_\delta$ therefore imply

$$\begin{aligned} \beta_t(\delta)^2 &\geq \mathcal{L}_t(\theta_\star) - \mathcal{L}_t(\hat{\theta}_t) \\ &\geq \sum_{(u,y)\in\mathcal{D}_t} \{u^\top(\theta_\star - \hat{\theta}_t)\}^2 \int_0^1 (1-v)\dot{\mu}(u^\top \hat{\theta}_t + v u^\top(\theta_\star - \hat{\theta}_t))\, dv. \end{aligned} \tag{11}$$

For each observed u, $|u^\top(\theta_\star - \hat{\theta}_t)| \leq 2S$. The endpoint curvature comparison of Auxiliary Lemma G.2 gives

$$\begin{aligned} &\dot{\mu}(u^\top \hat{\theta}_t + v u^\top(\theta_\star - \hat{\theta}_t)) \\ &\qquad \geq \dot{\mu}(u^\top \theta_\star) e^{-(1-v)|u^\top(\theta_\star - \hat{\theta}_t)|} \geq \dot{\mu}(u^\top \theta_\star) e^{-2S(1-v)}. \end{aligned}$$

Since $e^{2S} \geq 1 + 2S + 2S^2$ and $S > 0$,

$$\begin{aligned} \int_0^1 (1-v) e^{-2S(1-v)}\, dv &= \int_0^1 r e^{-2Sr}\, dr = \frac{1 - (1+2S)e^{-2S}}{4S^2} \\ &\geq \frac{1 - (1+2S)/(1+2S+2S^2)}{4S^2} = \frac{1}{2+4S+4S^2} \geq \frac{1}{2(1+2S)^2}. \end{aligned}$$

Substituting into (11) yields

$$\|\hat{\theta}_t - \theta_\star\|^2_{\nabla^2 \mathcal{L}_t(\theta_\star)} \leq 2(1+2S)^2 \beta_t(\delta)^2.$$

Adding $\lambda \|\hat{\theta}_t - \theta_\star\|_2^2 \leq 4S^2\lambda$ and using (5) gives

$$\|\hat{\theta}_t - \theta_\star\|^2_{\mathbf{H}_t(\theta_\star)} \leq 2(1+2S)^2 \beta_t(\delta)^2 + 4S^2\lambda = \Gamma_t(\delta)^2.$$

□

D Regret Upper Bound

D.1 Lemmas used in the regret analysis

Lemmas D.1 and D.2 bound the value regret in each round in terms of optimistic prediction errors. Lemma D.3 converts sums weighted by reach probabilities into sums over observed positions. Lemma D.4 compares information matrices to bound the exploration bonus. Lemma D.5 bounds sums of clipped squared norms, and Corollary D.1 applies this bound to count observed positions outside $\mathcal{I}_{\mathrm{good}}$ or in $\mathcal{J}$. Proofs are collected in Appendix D.3.

Lemma D.1 (Optimism and prediction error)**.** *On $\mathcal{G}_\delta$, for every feasible (t, k, s, a),*

$$\widetilde{p}_{t,k}(s,a) \geq p_t^\star(s,a), \qquad \widetilde{V}_{t,k}(s) \geq V_{t,k}^\star(s).$$

Along the selected no-click trajectory,

$$0 \leq \widetilde{p}_{t,k} - p_{t,k}^\star \leq 2\gamma_t(\delta)\dot{\mu}((\varphi_{t,k}^{\pi_t})^\top \theta_\star) \|\varphi_{t,k}^{\pi_t}\|_{\mathbf{H}_t(\hat{\theta}_t)^{-1}} + \frac{\gamma_t(\delta)^2}{2} \|\varphi_{t,k}^{\pi_t}\|^2_{\mathbf{H}_t(\hat{\theta}_t)^{-1}}.$$

In Lemma D.1, the first assertion is the usual monotone Bellman argument; the second is a Taylor bound applied to the closed-form score.

Lemma D.2 (Weighted value decomposition)**.** *On $\mathcal{G}_\delta$, for every $t \in [T]$,*

$$V_{t,1}^\star(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}}) \leq \sum_{k=1}^{K} \left\{ \prod_{\ell<k} (1 - p_{t,\ell}^\star) \right\} (\widetilde{p}_{t,k} - p_{t,k}^\star) \left\{ \prod_{m>k} (1 - p_{t,m}^\star) \right\}.$$

Lemma D.2 bounds the value regret of the selected cascade by a weighted sum of optimistic prediction errors. Each weight contains a prefix no-click factor, giving the probability of reaching the position, and a suffix no-click factor, giving the probability that none of the subsequent items would be clicked.

Lemma D.3 (Expectation identity for observed positions)**.** *For any $t \in [T]$ and $k \in [K]$, let $Z_{t,k}$ be a nonnegative random variable determined by $\mathcal{H}_t$ and the selected cascade π_t, before the feedback of round t. Then*

$$\mathbb{E}\left[Z_{t,k}\prod_{\ell<k}(1-p^\star_{t,\ell})\right] = \mathbb{E}[\mathbf{1}\{k \le \min\{\tau_t, K\}\}Z_{t,k}].$$

The identity also holds after summing over rounds and positions. This is the cascading specialization of the triggering probability equivalence in [25, Lemma 4]; the same conditional-expectation argument is used in [26, Appendix E.1].

Next, for the hessian comparison, we introduce only the following analysis objects:

$$\mathbf{V}_t = \kappa_\star \lambda \mathbf{I}_d + \sum_{(u,y)\in\mathcal{D}_t} uu^\top, \tag{12}$$

$$\mathcal{I}_{\text{good}} = \left\{(t,k) \in [T]\times[K] : \kappa_\star \Gamma_T(\delta)^2 \|\varphi^{\pi_t}_{t,k}\|^2_{\mathbf{V}_t^{-1}} \le 1\right\}, \tag{13}$$

$$\mathbf{H}^{\text{good}}_t = \lambda \mathbf{I}_d + \sum_{s<t}\sum_{\substack{1\le j\le \min\{\tau_s,K\}\\(s,j)\in\mathcal{I}_{\text{good}}}} \dot\mu((\varphi^{\pi_s}_{s,j})^\top\theta_\star)\varphi^{\pi_s}_{s,j}(\varphi^{\pi_s}_{s,j})^\top. \tag{14}$$

These quantities are used only in the analysis. The algorithm computes $\mathbf{H}_t(\hat\theta_t)$ from all observations and does not test membership in $\mathcal{I}_{\text{good}}$. Membership is determined before the feedback of round t.

Lemma D.4 (Comparison with the filtered true hessian matrix)**.** *On $\mathcal{G}_\delta$, for every $t \in [T]$,*

$$\mathbf{H}_t(\hat\theta_t) \succeq e^{-1}\mathbf{H}^{\text{good}}_t.$$

Consequently, for every $k \in [K]$, $\gamma_t(\delta)\|\varphi^{\pi_t}_{t,k}\|_{\mathbf{H}_t(\hat\theta_t)^{-1}} \le \sqrt{e}\,\gamma_T(\delta)\|\varphi^{\pi_t}_{t,k}\|_{(\mathbf{H}^{\text{good}}_t)^{-1}}$.

The proof of Lemma D.4 bounds $\|\hat\theta_t - \theta_\star\|_{H_t(\theta_\star)}$ and applies [2, Lemma 9] to compare the logistic derivatives at $\hat\theta_t$ and $\theta_\star$ for past good observations.

Lemma D.5 (Clipped pre-round potential)**.** *Let $a > 0$, $\|y_{t,k}\|_2 \le L$, and define*

$$\mathbf{M}_t = a\mathbf{I}_d + \sum_{s<t}\sum_{j=1}^{\min\{\tau_s,K\}} y_{s,j}y_{s,j}^\top.$$

Then, for every realized sequence,

$$\sum_{t=1}^{T}\sum_{k=1}^{\min\{\tau_t,K\}} \min\{1, \|y_{t,k}\|^2_{\mathbf{M}_t^{-1}}\} \le 2d\log\left(1+\frac{TKL^2}{da}\right) + \frac{KL^2d}{a}.$$

This is a consequence of the standard clipped elliptical potential and the squared-norm trace inequality in Appendix F.

We also define the set of good positions at which the following weighted squared norm exceeds one:

$$\mathcal{J} = \left\{(t,k) \in \mathcal{I}_{\text{good}} : \dot\mu((\varphi^{\pi_t}_{t,k})^\top\theta_\star)\|\varphi^{\pi_t}_{t,k}\|^2_{(\mathbf{H}^{\text{good}}_t)^{-1}} > 1\right\}. \tag{15}$$

Its membership is also determined before the current feedback. All past good observations, including those in $\mathcal{J}$, remain in $\mathbf{H}^{\text{good}}_t$.

Corollary D.1 (Counts of exceptional observations)**.** *On every realized path,*

$$\sum_{t=1}^{T}\sum_{k=1}^{\min\{\tau_t,K\}} \mathbf{1}\{(t,k)\notin\mathcal{I}_{\text{good}}\} \le \frac{\Gamma_T(\delta)^2Kd}{\lambda} + 2\kappa_\star\Gamma_T(\delta)^2 d\log\left(1+\frac{TK}{d\kappa_\star\lambda}\right),$$

$$\sum_{t=1}^{T}\sum_{k=1}^{\min\{\tau_t,K\}} \mathbf{1}\{(t,k)\in\mathcal{J}\} \le \frac{Kd}{4\lambda} + 2d\log\left(1+\frac{TK}{4d\lambda}\right).$$

Both inequalities follow by substitution into Lemma D.5.

D.2 Proof of Theorem 1

By the MDP equivalence in Proposition 1, $\mathrm{Regret}_T^{\mathrm{TopicCB}}$ equals $\mathrm{Regret}_T^{\mathrm{MDP}}$ which is the expected sum of $V_{t,1}^{\star}(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}})$. Each difference lies in $[0,1]$. Hence

$$\begin{aligned}\mathrm{Regret}_T^{\mathrm{TopicCB}} =& \mathbb{E}\left[\mathbf{1}_{\mathcal{G}_\delta}\sum_{t=1}^{T}\{V_{t,1}^{\star}(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}})\}\right] + \mathbb{E}\left[\mathbf{1}_{\mathcal{G}_\delta^c}\sum_{t=1}^{T}\{V_{t,1}^{\star}(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}})\}\right] \\ \le& \mathbb{E}\left[\mathbf{1}_{\mathcal{G}_\delta}\sum_{t=1}^{T}\{V_{t,1}^{\star}(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}})\}\right] + \delta T. \end{aligned} \tag{16}$$

On $\mathcal{G}_\delta$, apply Lemmas D.1–D.3 and partition all positions into $\mathcal{I}_{\mathrm{good}}^c$, $\mathcal{J}$, and $\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}$. The resulting pointwise inequality is

$$\begin{aligned}&\sum_{t=1}^{T}\{V_{t,1}^{\star}(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}})\} \\ &\le \sum_{t=1}^{T}\sum_{k=1}^{K}\left\{\prod_{\ell<k}(1-p_{t,\ell}^{\star})\right\}\left\{\prod_{m>k}(1-p_{t,m}^{\star})\right\} \\ &\quad\times\Bigg[\mathbf{1}\{(t,k)\notin\mathcal{I}_{\mathrm{good}}\} + \mathbf{1}\{(t,k)\in\mathcal{J}\} \\ &\quad\quad + \mathbf{1}\{(t,k)\in\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}\}\begin{Bmatrix} 2\sqrt{e}\,\gamma_T(\delta)\dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} \\ + \dfrac{e\gamma_T(\delta)^2}{2}\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_t^{\mathrm{good}})^{-1}}\end{Bmatrix}\Bigg].\end{aligned} \tag{17}$$

Since $\mathbf{H}_t^{\mathrm{good}} \succeq \lambda\mathbf{I}_d$, the right-hand side is well-defined and nonnegative on every path. Its expectation therefore bounds the expected regret restricted to $\mathcal{G}_\delta$.

Bounding the contribution from exceptional positions. For positions $(t,k)\notin\mathcal{I}_{\mathrm{good}}$ or $(t,k)\in\mathcal{J}$, use $\prod_{m>k}(1-p_{t,m}^{\star})\le 1$ and apply Lemma D.3:

$$\begin{aligned}&\mathbb{E}\sum_{t=1}^{T}\sum_{k=1}^{K}\left\{\prod_{\ell<k}(1-p_{t,\ell}^{\star})\right\}\left\{\prod_{m>k}(1-p_{t,m}^{\star})\right\}\left[\mathbf{1}\{(t,k)\notin\mathcal{I}_{\mathrm{good}}\} + \mathbf{1}\{(t,k)\in\mathcal{J}\}\right] \\ &\le \mathbb{E}\sum_{t=1}^{T}\sum_{k=1}^{\min\{\tau_t,K\}}\left[\mathbf{1}\{(t,k)\notin\mathcal{I}_{\mathrm{good}}\} + \mathbf{1}\{(t,k)\in\mathcal{J}\}\right].\end{aligned} \tag{18}$$

The identity applies because both memberships are determined before the current feedback. Corollary D.1 bounds the right-hand side of (18). Next, it remains to bound the contribution from $\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}$, whose prediction-error bound contains a norm term and a squared-norm term.

Decomposing the two weighted sums. For the contribution of $\mathbf{1}\{(t,k)\in\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}\}$ in (17), we will establish the following bounds:

$$\begin{aligned}&\mathbb{E}\sum_{t=1}^{T}\sum_{\substack{1\le k\le K\\(t,k)\in\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}}}\left\{\prod_{\ell\neq k}(1-p_{t,\ell}^{\star})\right\}\dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} \\ &\quad\le \frac{Kd}{\sqrt{\lambda}} + \bar{p}^{(K-1)/2}\sqrt{2Td\log\left(1+\frac{TK}{4d\lambda}\right)},\end{aligned} \tag{19}$$

$$\begin{aligned}&\mathbb{E}\sum_{t=1}^{T}\sum_{\substack{1\le k\le K\\(t,k)\in\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}}}\left\{\prod_{\ell\neq k}(1-p_{t,\ell}^{\star})\right\}\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} \\ &\quad\le \frac{Kd}{\lambda} + 2\kappa_\star d\log\left(1+\frac{TK}{4d\lambda}\right).\end{aligned} \tag{20}$$

To prove (19) and (20), we separate both the norm and its square into a term using $\mathbf{H}_{t,k}^{\mathrm{good}}$ and the difference from the term using $\mathbf{H}_t^{\mathrm{good}}$. For each position, define

$$\mathbf{H}_{t,k}^{\mathrm{good}} = \mathbf{H}_t^{\mathrm{good}} + \sum_{\substack{j<k \\ (t,j)\in\mathcal{I}_{\mathrm{good}}}} \dot{\mu}((\varphi_{t,j}^{\pi_t})^\top \theta_\star)\varphi_{t,j}^{\pi_t}(\varphi_{t,j}^{\pi_t})^\top.$$

For an observed position, all earlier positions have also been observed, so

$$\mathbf{H}_t^{\mathrm{good}} \preceq \mathbf{H}_{t,k}^{\mathrm{good}} \preceq \mathbf{H}_{t+1}^{\mathrm{good}}.$$

Since $\mathbf{H}_{t,k}^{\mathrm{good}} \succeq \mathbf{H}_t^{\mathrm{good}}$, we have

$$\|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} = \|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}} + \left\{\|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} - \|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\right\}, \tag{21}$$

and

$$\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} = \|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}} + \left\{\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} - \|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\right\}. \tag{22}$$

The differences in braces are nonnegative. Substitute (21) into the left-hand side of (19) and (22) into the left-hand side of (20), leaving all other weights unchanged. We bound the resulting terms separately.

Applying the elliptical potential lemma. We first bound the sum of curvature-weighted squared norms using $\mathbf{H}_{t,k}^{\mathrm{good}}$ over the observed positions in $\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}$. We will use it to bound the terms involving $\mathbf{H}_{t,k}^{\mathrm{good}}$ in (20) through $\dot{\mu} \geq \kappa_\star^{-1}$, and in (19) through Cauchy–Schwarz. List all observed good positions in order and consider the vectors $\sqrt{\dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)}\varphi_{t,k}^{\pi_t}$. Since $\dot{\mu} \leq 1/4$ and $\|\varphi_{t,k}^{\pi_t}\|_2 \leq 1$, these vectors have norm at most $1/2$. Applying Auxiliary Lemma G.4 with $a = \lambda$, $L = 1/2$, and at most TK vectors gives

$$\sum_{t=1}^{T} \sum_{\substack{k\leq\min\{\tau_t,K\} \\ (t,k)\in\mathcal{I}_{\mathrm{good}}}} \min\left\{1, \dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\right\} \leq 2d\log\left(1+\frac{TK}{4d\lambda}\right).$$

For $(t,k) \in \mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}$, $\dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}} \leq \dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} \leq 1$. Therefore, on this subset $\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}$, clipping can be removed. In particular,

$$\sum_{t=1}^{T} \sum_{\substack{k\leq\min\{\tau_t,K\} \\ (t,k)\in\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}}} \dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}} \leq 2d\log\left(1+\frac{TK}{4d\lambda}\right). \tag{23}$$

Bounding the norm differences. We now bound the contributions from the terms in braces in (21) and (22). For the former, keep the weights in (19), discard the suffix product, and apply Lemma D.3. The contribution is at most

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T} \sum_{\substack{k\leq\min\{\tau_t,K\} \\ (t,k)\in\mathcal{I}_{\mathrm{good}}\backslash\mathcal{J}}} \dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\left\{\|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_t^{\mathrm{good}})^{-1}} - \|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\right\} \\
&\leq \frac{K}{2}\,\mathbb{E}\sum_{t=1}^{T}\{\mathrm{tr}((\mathbf{H}_t^{\mathrm{good}})^{-1/2}) - \mathrm{tr}((\mathbf{H}_{t+1}^{\mathrm{good}})^{-1/2})\} \leq \frac{Kd}{2\sqrt{\lambda}} \leq \frac{Kd}{\sqrt{\lambda}}.
\end{aligned} \tag{24}$$

Here Auxiliary Lemma G.3 supplies the factor 2, while $\dot{\mu} \leq 1/4$ supplies the factor $1/4$; at most K positions are observed in a round. The comparison with $\mathbf{H}_{t+1}^{\mathrm{good}}$ is made only after the prefix identity

has converted the sum to actual observations. For the term in braces in (22), use the weights in (20). The same argument with the squared-norm part of Auxiliary Lemma G.3 bounds its contribution by

$$\begin{aligned}
&\mathbb{E}\sum_{t=1}^{T}\sum_{\substack{k\le\min\{\tau_t,K\}\\(t,k)\in\mathcal{I}_{\mathrm{good}}\setminus\mathcal{J}}}\dot\mu((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\left\{\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_t^{\mathrm{good}})^{-1}}-\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\right\}\\
&\le K\,\mathbb{E}\sum_{t=1}^{T}\{\mathrm{tr}((\mathbf{H}_t^{\mathrm{good}})^{-1})-\mathrm{tr}((\mathbf{H}_{t+1}^{\mathrm{good}})^{-1})\}\le\frac{Kd}{\lambda}.
\end{aligned}\tag{25}$$

Bounding the remaining term in (19). It remains to bound the contribution from the first term on the right-hand side of (21), with the weights in (19). Using $\dot\mu((\varphi_{t,k}^{\pi_t})^\top\theta_\star)=p_{t,k}^\star(1-p_{t,k}^\star)$, Cauchy–Schwarz gives, for each round t,

$$\begin{aligned}
&\left[\sum_{\substack{1\le k\le K\\(t,k)\in\mathcal{I}_{\mathrm{good}}\setminus\mathcal{J}}}\left\{\prod_{\ell<k}(1-p_{t,\ell}^\star)\right\}\dot\mu((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\left\{\prod_{m>k}(1-p_{t,m}^\star)\right\}\right]^2\\
&\le\left[\sum_{\substack{1\le k\le K\\(t,k)\in\mathcal{I}_{\mathrm{good}}\setminus\mathcal{J}}}p_{t,k}^\star\prod_{m>k}(1-p_{t,m}^\star)\right]\\
&\quad\times\left[\sum_{\substack{1\le k\le K\\(t,k)\in\mathcal{I}_{\mathrm{good}}\setminus\mathcal{J}}}\left\{\prod_{\ell<k}(1-p_{t,\ell}^\star)\right\}^2\left\{\prod_{m>k}(1-p_{t,m}^\star)\right\}p_{t,k}^\star(1-p_{t,k}^\star)^2\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\right].
\end{aligned}\tag{26}$$

The first factor is at most $\sum_{k=1}^K p_{t,k}^\star\prod_{m>k}(1-p_{t,m}^\star)=1-\prod_{k=1}^K(1-p_{t,k}^\star)\le 1$. For every summand in the second factor,

$$\left\{\prod_{\ell<k}(1-p_{t,\ell}^\star)\right\}^2\left\{\prod_{m>k}(1-p_{t,m}^\star)\right\}p_{t,k}^\star(1-p_{t,k}^\star)^2\le\bar p^{K-1}\left\{\prod_{\ell<k}(1-p_{t,\ell}^\star)\right\}\dot\mu((\varphi_{t,k}^{\pi_t})^\top\theta_\star).$$

Take square roots in (26), sum over t, and apply Cauchy–Schwarz in round followed by Jensen's inequality. This contribution to the left-hand side of (19) is at most

$$\begin{aligned}
&\bar p^{(K-1)/2}\left[T\,\mathbb{E}\sum_{t=1}^{T}\sum_{\substack{1\le k\le K\\(t,k)\in\mathcal{I}_{\mathrm{good}}\setminus\mathcal{J}}}\left\{\prod_{\ell<k}(1-p_{t,\ell}^\star)\right\}\dot\mu((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\right]^{1/2}\\
&=\bar p^{(K-1)/2}\left[T\,\mathbb{E}\sum_{t=1}^{T}\sum_{\substack{k\le\min\{\tau_t,K\}\\(t,k)\in\mathcal{I}_{\mathrm{good}}\setminus\mathcal{J}}}\dot\mu((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\right]^{1/2}\\
&\le\bar p^{(K-1)/2}\sqrt{2Td\log\left(1+\frac{TK}{4d\lambda}\right)}.
\end{aligned}$$

The equality is by applying Lemma D.3; the final inequality is (23). Combining this with (24) proves (19).

Bounding the remaining term in (20). For the contribution from the first term on the right-hand side of (22), keep the weights in (20), discard the suffix product, and apply Lemma D.3. Then the definition of $\kappa_\star$ gives

$$\begin{aligned}
\mathbb{E}\sum_{t=1}^{T}\sum_{\substack{k\le\min\{\tau_t,K\}\\(t,k)\in\mathcal{I}_{\mathrm{good}}\setminus\mathcal{J}}}\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}&\le\kappa_\star\mathbb{E}\sum_{t=1}^{T}\sum_{\substack{k\le\min\{\tau_t,K\}\\(t,k)\in\mathcal{I}_{\mathrm{good}}\setminus\mathcal{J}}}\dot\mu((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|^2_{(\mathbf{H}_{t,k}^{\mathrm{good}})^{-1}}\\
&\le 2\kappa_\star d\log\left(1+\frac{TK}{4d\lambda}\right).
\end{aligned}$$

Together with (25), this proves (20).

Substituting Corollary D.1 and (19)–(20) into (16)–(18) gives

$$\begin{aligned}\mathrm{Regret}_T^{\mathrm{TopicCB}} \le{}& 2\sqrt{2e}\,\gamma_T(\delta)\bar{p}^{(K-1)/2}\sqrt{Td\log\left(1+\frac{TK}{4d\lambda}\right)} + \frac{2\sqrt{e}\,\gamma_T(\delta)Kd}{\sqrt{\lambda}} \\ &+ \left\{\Gamma_T(\delta)^2 + \frac{e}{2}\gamma_T(\delta)^2 + \frac{1}{4}\right\}\frac{Kd}{\lambda} \\ &+ e\kappa_\star\gamma_T(\delta)^2 d\log\left(1+\frac{TK}{4d\lambda}\right) \\ &+ 2\kappa_\star\Gamma_T(\delta)^2 d\log\left(1+\frac{TK}{d\kappa_\star\lambda}\right) + 2d\log\left(1+\frac{TK}{4d\lambda}\right) + \delta T.\end{aligned} \tag{27}$$

We now use the choices $\lambda = 1/(4S^2)$ and $\delta = 1/(T+1)$ in Theorem 1. Throughout the following estimates, S is fixed and the implicit constants may depend on S. By the definitions of the radii,

$$\begin{aligned}\beta_T(\delta)^2 &= \log(T+1) + d\log\left(e \vee \frac{2eS(T-1)K}{d}\right) = \mathcal{O}(d\log(1+TK))\,, \\ \gamma_T(\delta) &= \sqrt{2(1+S)\{1+\beta_T(\delta)^2\}} = \mathcal{O}\left(\sqrt{d\log(1+TK)}\right), \\ \Gamma_T(\delta)^2 &= 2(1+2S)^2\beta_T(\delta)^2 + 1 = \mathcal{O}(d\log(1+TK))\,.\end{aligned}$$

Also, $\lambda^{-1/2} = 2S$, $\lambda^{-1} = 4S^2$, and $\delta T \le 1$. Since $\kappa_\star \ge 4$ and $d \ge 1$,

$$\log\left(1+\frac{TK}{d\kappa_\star\lambda}\right) \le \log\left(1+\frac{TK}{4d\lambda}\right) = \log\left(1+\frac{S^2TK}{d}\right) = \mathcal{O}(\log(1+TK))\,.$$

Substituting these estimates into (27) gives

$$\begin{aligned}\mathrm{Regret}_T^{\mathrm{TopicCB}} \le \mathcal{O}\Big(&\bar{p}^{(K-1)/2}d\sqrt{T}\log(1+TK) + Kd^{3/2}\sqrt{\log(1+TK)} \\ &+ Kd^2\log(1+TK) + \kappa_\star d^2\log^2(1+TK)\Big) + 1.\end{aligned}$$

Because $d, K \ge 1$ and $\log(1+TK) \ge \log 2$, the second term and the additive constant are absorbed into $\mathcal{O}(Kd^2\log(1+TK))$. Suppressing logarithmic factors and using $d = d_h d_a$, we obtain

$$\mathrm{Regret}_T^{\mathrm{TopicCB}} = \widetilde{\mathcal{O}}\left(\bar{p}^{(K-1)/2}d_h d_a\sqrt{T} + Kd_h^2 d_a^2 + d_h^2 d_a^2\kappa_\star\right),$$

which proves Theorem 1.

D.3 Proofs of Lemmas

D.3.1 Proof of Lemma D.1

Proof. Cauchy–Schwarz and Corollary C.2 imply $U_{t,k}(s,a) \ge \varphi_t(s,a)^\top\theta_\star$, and hence $\widetilde{p}_{t,k}(s,a) \ge p_t^\star(s,a)$. The optimistic Bellman recursion gives $0 \le \widetilde{V}_{t,k}(s) \le 1$ by backward induction. Since $p + (1-p)v$ is nondecreasing in p and v on $[0,1]^2$, the same backward induction gives $\widetilde{V}_{t,k}(s) \ge V_{t,k}^\star(s)$. The same Cauchy–Schwarz bound also gives

$$0 \le U_{t,k}(s_{t,k}^{\pi_t}, a_{t,k}) - (\varphi_{t,k}^{\pi_t})^\top\theta_\star \le 2\gamma_t(\delta)\|\varphi_{t,k}^{\pi_t}\|_{\mathbf{H}_t(\hat{\theta}_t)^{-1}}.$$

Apply the scalar Taylor bound in Auxiliary Lemma G.1 with this nonnegative logit increment. Its linear term is at most $2\gamma_t(\delta)\dot{\mu}((\varphi_{t,k}^{\pi_t})^\top\theta_\star)\|\varphi_{t,k}^{\pi_t}\|_{\mathbf{H}_t(\hat{\theta}_t)^{-1}}$, and its quadratic term is at most $\frac{1}{8}(2\gamma_t(\delta)\|\varphi_{t,k}^{\pi_t}\|_{\mathbf{H}_t(\hat{\theta}_t)^{-1}})^2$. □

D.3.2 Proof of Lemma D.2

Proof. Along the selected maximizing no-click trajectory, the Bellman recursions give

$$\widetilde{V}_{t,1}(s_{\mathrm{init}}) = 1 - \prod_{k=1}^{K}(1-\widetilde{p}_{t,k}), \qquad V_{t,1}^{\pi_t}(s_{\mathrm{init}}) = 1 - \prod_{k=1}^{K}(1-p_{t,k}^\star).$$

Pointwise optimism and the product telescoping identity imply

$$\begin{aligned} V_{t,1}^{\star}(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}}) &\le \widetilde{V}_{t,1}(s_{\mathrm{init}}) - V_{t,1}^{\pi_t}(s_{\mathrm{init}}) \\ &= \sum_{k=1}^{K} \left\{ \prod_{\ell<k} (1 - p_{t,\ell}^{\star}) \right\} (\widetilde{p}_{t,k} - p_{t,k}^{\star}) \left\{ \prod_{m>k} (1 - \widetilde{p}_{t,m}) \right\} \\ &\le \sum_{k=1}^{K} \left\{ \prod_{\ell<k} (1 - p_{t,\ell}^{\star}) \right\} (\widetilde{p}_{t,k} - p_{t,k}^{\star}) \left\{ \prod_{m>k} (1 - p_{t,m}^{\star}) \right\}. \end{aligned}$$

The last inequality uses $\widetilde{p}_{t,k} \ge p_{t,k}$. □

D.3.3 Proof of Lemma D.3

Proof. Conditional on $\mathcal{H}_t$ and π_t, position k is observed if and only if all preceding outcomes are zero. Conditional independence therefore gives

$$\mathbb{P}(k \le \min\{\tau_t, K\} \mid \mathcal{H}_t, \pi_t) = \prod_{\ell<k} (1 - p_{t,\ell}^{\star}).$$

Since $Z_{t,k}$ is determined by $\mathcal{H}_t$ and π_t,

$$\mathbb{E}[\mathbf{1}\{k \le \min\{\tau_t, K\}\} Z_{t,k} \mid \mathcal{H}_t, \pi_t] = Z_{t,k} \prod_{\ell<k} (1 - p_{t,\ell}^{\star}).$$

Taking expectations and summing proves its stated consequence. The identity is applied under the original probability law, after removing the confidence-event indicator. □

D.3.4 Proof of Lemma D.4

Proof. Fix $t \in [T]$ and work on $\mathcal{G}_\delta$. By the definition of $\kappa_\star$, on every path,

$$\mathbf{H}_t(\theta_\star) \succeq \kappa_\star^{-1} \mathbf{V}_t.$$

Fix an observed past position $(s, j) \in \mathcal{I}_{\mathrm{good}}$ with $s < t$. Cauchy–Schwarz, Corollary C.3, $\mathbf{V}_t \succeq \mathbf{V}_s$, and monotonicity of $\Gamma_t(\delta)$ give

$$\begin{aligned} |(\varphi_{s,j}^{\pi_s})^\top (\hat{\theta}_t - \theta_\star)| &\le \Gamma_t(\delta) \|\varphi_{s,j}^{\pi_s}\|_{\mathbf{H}_t(\theta_\star)^{-1}} \\ &\le \sqrt{\kappa_\star} \Gamma_T(\delta) \|\varphi_{s,j}^{\pi_s}\|_{\mathbf{V}_t^{-1}} \\ &\le \sqrt{\kappa_\star} \Gamma_T(\delta) \|\varphi_{s,j}^{\pi_s}\|_{\mathbf{V}_s^{-1}} \le 1. \end{aligned}$$

Lemma F.2 therefore implies

$$\dot{\mu}((\varphi_{s,j}^{\pi_s})^\top \hat{\theta}_t) \ge e^{-1} \dot{\mu}((\varphi_{s,j}^{\pi_s})^\top \theta_\star).$$

Retain only these good summands in $\mathbf{H}_t(\hat{\theta}_t)$ and use $\lambda \mathbf{I}_d \succeq e^{-1} \lambda \mathbf{I}_d$ to obtain the matrix inequality in Lemma D.4. Inverting the positive definite matrices and using $\gamma_t(\delta) \le \gamma_T(\delta)$ gives the bonus bound in the same lemma. □

D.3.5 Proof of Lemma D.5

Proof. At every observed position $k \le \min\{\tau_t, K\}$,

$$\mathbf{M}_t \preceq \mathbf{M}_t + \sum_{j<k} y_{t,j} y_{t,j}^\top \preceq \mathbf{M}_{t+1}.$$

For $r \ge q \ge 0$, $0 \le \min\{1, r\} - \min\{1, q\} \le r - q$. Together with Auxiliary Lemma G.3, this gives

$$\begin{aligned} &\min\{1, \|y_{t,k}\|_{\mathbf{M}_t^{-1}}^2\} - \min\Big\{1, \|y_{t,k}\|_{(\mathbf{M}_t + \sum_{j<k} y_{t,j} y_{t,j}^\top)^{-1}}^2\Big\} \\ &\quad \le \|y_{t,k}\|_{\mathbf{M}_t^{-1}}^2 - \|y_{t,k}\|_{\mathbf{M}_{t+1}^{-1}}^2 \\ &\quad \le L^2 \{\mathrm{tr}(\mathbf{M}_t^{-1}) - \mathrm{tr}(\mathbf{M}_{t+1}^{-1})\}. \end{aligned}$$

Sum over the observed positions and use at most K observations per round. Auxiliary Lemma G.4 applies pathwise to the observed sequence, whose length is at most TK. Hence

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{k=1}^{\min\{\tau_t,K\}} \min\{1, \|y_{t,k}\|^2_{\mathbf{M}_t^{-1}}\} \\
&\le 2d\log\left(1+\frac{TKL^2}{da}\right) + KL^2\sum_{t=1}^{T}\{\mathrm{tr}(\mathbf{M}_t^{-1}) - \mathrm{tr}(\mathbf{M}_{t+1}^{-1})\} \\
&\le 2d\log\left(1+\frac{TKL^2}{da}\right) + KL^2\,\mathrm{tr}((aI_d)^{-1}).
\end{aligned}$$

□

D.3.6 Proof of Corollary D.1

Proof. Because $\kappa_\star\Gamma_T(\delta)^2 \ge 8$, the definition of $\mathcal{I}_{\mathrm{good}}$ implies, on every path,

$$\mathbf{1}\{(t,k)\notin\mathcal{I}_{\mathrm{good}}\} \le \kappa_\star\Gamma_T(\delta)^2\min\{1, \|\varphi_{t,k}^{\pi_t}\|^2_{\mathbf{V}_t^{-1}}\}.$$

Apply Lemma D.5 with $a=\kappa_\star\lambda, L=1, y_{t,k}=\varphi_{t,k}^{\pi_t}, \mathbf{M}_t=\mathbf{V}_t$, and multiply its bound by $\kappa_\star\Gamma_T(\delta)^2$. This gives the first inequality. For the second inequality, apply Lemma D.5 with $a=\lambda, L=\frac{1}{2}, y_{t,k}=\mathbf{1}\{(t,k)\in\mathcal{I}_{\mathrm{good}}\}\sqrt{\dot\mu((\varphi_{t,k}^{\pi_t})^\top\theta_\star)}\varphi_{t,k}^{\pi_t}$. Then $\mathbf{M}_t=\mathbf{H}_t^{\mathrm{good}}$, $\|y_{t,k}\|_2\le 1/2$, and

$$\mathbf{1}\{(t,k)\in\mathcal{J}\} \le \min\{1, \|y_{t,k}\|^2_{\mathbf{M}_t^{-1}}\}.$$

Summing the two indicator inequalities over the observed positions completes the proof. □

E Local Minimax Lower Bound

We prove Theorem 2 by constructing a finite TopicCB instance with item features fixed across rounds. The argument combines a hypercube family of logistic models with a comparison of cascade rewards that preserves the topic order selected by the learner.

Local reference quantities. Fix a reference logit $b\in\mathbb{R}$ and let $\kappa_0 := 1/\dot\mu(b)$. We take $\rho := \log 2$ and $W := [b-\rho, b+\rho]$. Indeed, the logistic derivative satisfies $|\frac{d}{du}\log\dot\mu(u)| = |1-2\mu(u)| \le 1$. Consequently, for every $u\in W$, $e^{-\rho}\dot\mu(b)\le\dot\mu(u)\le e^{\rho}\dot\mu(b)$, so

$$\frac{1}{2\kappa_0} \le \dot\mu(u) \le \frac{2}{\kappa_0}, \qquad u\in W. \tag{28}$$

Define $p_{\mathrm{loc}} := 1-\mu(b+\rho)$ and, for a radius $\varepsilon_T>0$, let $\mathcal{N}_T(\Theta_0) := \{\Theta : \|\Theta-\Theta_0\|_F\le\varepsilon_T\}$. Recall from Definition 2 that, for fixed item features, the inverse link-derivative constant can be written in the cascade notation of Section 2 as

$$\kappa_\star(\Theta) := \sup_{\substack{C=(i_1,\dots,i_K)\in\Pi\\ k\in[K]}} \frac{1}{\dot\mu\big(\phi(z_{<k}(C))^\top\Theta x_{i_k}\big)}. \tag{29}$$

The construction below yields $\kappa_\star(\Theta_0)=\kappa_0$ and ensures that all feasible logits in $\mathcal{N}_T(\Theta_0)$ belong to W for sufficiently large T.

E.1 Hard-Instance Construction

Hard positions and hidden rows. We encode unknown binary signs in rows of the parameter matrix. When $d_h\ge 2$, one row is reserved so that positions outside the hard set do not depend on these signs. When $d_h=1$, the same hidden row is used at every cascade position. Let m be the number of rows containing hidden signs, $\mathcal{K}$ the hard positions, and L the maximum number of positions affected by a single hidden row per round. Specifically,

1. If $d_h\ge 2$, take $m=d_h-1$, $\mathcal{K}=[m]$, and $L=1$.
2. If $d_h=1$, take $m=1$, $\mathcal{K}=[K]$, and $L=K$.

Item and topic-history features. Set $q := d_a - 1$ and introduce exactly K topics. For every topic $g \in [K]$ and sign vector $\sigma \in \{-1, +1\}^q$, introduce an item $i(g, \sigma)$ with fixed feature

$$x_\sigma := 2^{-1/2}\big(1, \sigma^\top/\sqrt{q}\big)^\top \in \mathbb{R}^{d_a}.$$

Thus $\|x_\sigma\|_2 = 1$, and each topic offers every sign vector. In particular, any ordering of the K distinct topics is feasible.

Set $\delta := \min\{1/4, \rho/[4(|b| + 1)]\}$. For every feasible topic history ζ, including terminal histories of length K, choose coefficients $c_\zeta \in [1 - \delta, 1]$ that are pairwise distinct, with $c_\emptyset = 1$, independently of T. Define the known topic-history representation by

$$\phi(\zeta) = \begin{cases} c_\zeta e_{\min\{|\zeta|+1, m\}}, & |\zeta| < K,\ |\zeta| + 1 \in \mathcal{K}, \\ c_\zeta e_{d_h}, & \text{otherwise.} \end{cases} \tag{30}$$

For a feasible cascade $C = (i_1, \ldots, i_K)$, this definition specifies $\phi(z_{<k}(C))$ for $k \in [K+1]$, exactly as in the problem setting. When $d_h \geq 2$, histories before hard position k use direction e_k; all other histories use e_{d_h}. When $d_h = 1$, every history uses e_1. The coefficients c_ζ make ϕ injective, and $\|\phi(\zeta)\|_2 \leq 1$.

Reference and hard parameters. Define $\Theta_0 \in \mathbb{R}^{d_h \times d_a}$ by

$$(\Theta_0)_{r,\cdot} = \sqrt{2}\,(b, 0_q^\top), \qquad r \in [d_h].$$

Every feasible logit at Θ_0 is thus $\phi(z_{<k}(C))^\top \Theta_0 x_\sigma = c_{z_{<k}(C)} b$. Fix once and for all a positive universal constant $c_\omega \leq 1/(4\sqrt{2})$, and for sufficiently large T set $\omega := c_\omega \sqrt{\kappa_0/(LT)}$. Let $\mathcal{V} := (\{-1, +1\}^q)^m$. For $v = (v_1, \ldots, v_m) \in \mathcal{V}$, define Θ_v by

$$(\Theta_v)_{r,\cdot} = \begin{cases} \sqrt{2}\,(b, \omega\sqrt{q}\, v_r^\top), & r \leq m, \\ (\Theta_0)_{r,\cdot}, & m < r \leq d_h. \end{cases}$$

For $r \leq m$, the row–item product equals $(\Theta_v)_{r,\cdot} x_\sigma = b + \omega\langle v_r, \sigma\rangle$. Thus the logit at position k of a feasible cascade $C = (i_1, \ldots, i_K)$, with sign σ_k at position k, satisfies

$$\phi(z_{<k}(C))^\top \Theta_v x_{\sigma_k} = \begin{cases} c_{z_{<k}(C)}\big(b + \omega\langle v_{\min\{k,m\}}, \sigma_k\rangle\big), & k \in \mathcal{K}, \\ c_{z_{<k}(C)} b, & k \notin \mathcal{K}. \end{cases} \tag{31}$$

The relevant hidden sign vector is v_k at each hard position when $d_h \geq 2$ and is always v_1 when $d_h = 1$. We use the shrinking radius

$$\varepsilon_T := \sqrt{2}\, c_\omega q \sqrt{\frac{d_h \kappa_0}{T}}.$$

E.2 Main Proof of Theorem

We now prove Theorem 2. The argument first shows that the constructed hypercube is admissible, then establishes a statistical lower bound on the learner's sign choices, and finally converts that bound into TopicCB regret. The proofs of the intermediate results stated here are collected in Section E.3.

Admissibility and full-feedback reduction. The hypercube lower bound must concern parameters inside the shrinking neighborhood while keeping the logistic derivatives uniformly controlled. The following lemma establishes these properties, together with the required boundedness condition.

Lemma E.1 (Local admissibility)**.** *For all sufficiently large T, the constructed instance satisfies:*

(i) $\{\Theta_v : v \in \mathcal{V}\} \subseteq \mathcal{N}_T(\Theta_0)$.

(ii) Every $\Theta \in \mathcal{N}_T(\Theta_0)$ induces logits in $[b - \rho/2, b + \rho/2] \subseteq W$ at every feasible topic-history–item pair.

(iii) The instance satisfies Assumption 1 throughout $\mathcal{N}_T(\Theta_0)$, with a parameter-norm bound independent of T.

For a lower bound it suffices to consider a more informative learner: after round t, reveal the entire latent click vector $Y_t = (Y_{t,1}, \dots, Y_{t,K})$ rather than only the outcomes up to the first click. Conditional on the history available before the round and the selected cascade, the latent outcomes are independent Bernoulli variables with their specified means. The original censored feedback is a deterministic function of Y_t, so any censored-feedback learner can be simulated by a full-feedback learner that discards the additional observations.

Information-theoretic lower bound. Under full feedback, we can assess how well the learner identifies the hidden sign vectors. For the played cascade $C_t = (i(g_{t,1}, \sigma_{t,1}), \dots, i(g_{t,K}, \sigma_{t,K}))$, the sign $v_{\min\{k,m\}}$ maximizes the click probability at hard position k when its topic history is fixed. Indeed, $\max_\sigma \langle v_{\min\{k,m\}}, \sigma \rangle = q$, and every topic contains an item with this sign. To measure the resulting suboptimality, define the auxiliary row-wise loss

$$\begin{aligned}\mathrm{Regret}_T^{\mathrm{row}}(v) := \sum_{t=1}^{T} \sum_{k \in \mathcal{K}} \Big[& \mu\big(c_{z_{<k}(C_t)}(b + \omega q)\big) \\ & - \mu\big(c_{z_{<k}(C_t)}(b + \omega \langle v_{\min\{k,m\}}, \sigma_{t,k} \rangle)\big) \Big].\end{aligned} \tag{32}$$

This is the sum of position-wise click-probability gaps; it is not yet the TopicCB regret, which is defined through the probability of at least one click.

We next establish that full feedback still cannot reliably distinguish hard parameters differing in a single hidden sign. Fix an arbitrary full-feedback algorithm Alg and let $\mathcal{T}_T = (C_1, Y_1, \dots, C_T, Y_T)$ be its transcript. For $v \in \mathcal{V}$, let P_v denote the law of $\mathcal{T}_T$ under Θ_v. For $a = (r, j) \in [m] \times [q]$, let $v^{(a)}$ be obtained from v by flipping only the jth sign of v_r.

Lemma E.2 (Neighboring transcript KL)**.** *For all sufficiently large T, every full-feedback algorithm, every $v \in \mathcal{V}$, and every $a \in [m] \times [q]$,*

$$\mathrm{KL}(P_v \| P_{v^{(a)}}) \le \frac{1}{8}.$$

By Pinsker's inequality, this KL bound implies that the corresponding transcript laws have total variation distance at most $1/4$. Applying an Assouad-type testing argument to each hidden sign therefore gives a lower bound on the auxiliary row-wise loss.

Corollary E.1 (Hypercube row-wise lower bound)**.** *For all sufficiently large T, in the full-feedback problem,*

$$\inf_{\mathsf{Alg}} \sup_{v \in \mathcal{V}} \mathbb{E}_v^{\mathsf{Alg}} \left[\mathrm{Regret}_T^{\mathrm{row}}(v) \right] \ge c_0 \frac{|\mathcal{K}|}{\sqrt{L}} q \sqrt{\frac{T}{\kappa_0}},$$

where $c_0 = 9c_\omega/32 > 0$ is a universal constant.

Conversion to TopicCB regret. The preceding bound concerns position-wise click-probability gaps, whereas TopicCB reward is the probability of at least one click. To compare these quantities without restricting the learner's topic order, we retain the selected topics at every position and replace only the signs at hard positions. This gives the following order-preserving comparison.

Lemma E.3 (Order-preserving cascade comparison)**.** *For all sufficiently large T, every $v \in \mathcal{V}$, and every sequence of feasible cascades $C_1, \dots, C_T$,*

$$\sum_{t=1}^{T} \big(f_t(C_t^\star(v); \Theta_v) - f_t(C_t; \Theta_v) \big) \ge p_{\mathrm{loc}}^{K-1} \, \mathrm{Regret}_T^{\mathrm{row}}(v),$$

where $C_t^\star(v) \in \arg\max_{C \in \Pi} f_t(C; \Theta_v)$.

Combining the order-preserving comparison and the row-wise lower bound, for every full-feedback algorithm we have

$$\mathrm{Regret}_T^{\mathrm{TopicCB}}(\Theta_v) \ge p_{\mathrm{loc}}^{K-1} \mathbb{E}_v^{\mathsf{Alg}} \left[\mathrm{Regret}_T^{\mathrm{row}}(v) \right].$$

Here the inequality follows by taking expectations in the pathwise comparison of Lemma E.3. Taking the supremum over v and the infimum over full-feedback algorithms, and applying Corollary E.1,

yields

$$\begin{aligned}\inf_{\mathsf{Alg}\,(\text{full})} \sup_{v\in\mathcal{V}} \mathrm{Regret}_T^{\mathrm{TopicCB}}(\Theta_v) &\geq p_{\mathrm{loc}}^{K-1} \inf_{\mathsf{Alg}\,(\text{full})} \sup_{v\in\mathcal{V}} \mathbb{E}_v^{\mathsf{Alg}}\left[\mathrm{Regret}_T^{\mathrm{row}}(v)\right] \\ &\geq c_0 p_{\mathrm{loc}}^{K-1} \frac{|\mathcal{K}|}{\sqrt{L}} q\sqrt{\frac{T}{\kappa_0}}.\end{aligned} \tag{33}$$

Our construction satisfies

$$\frac{|\mathcal{K}|}{\sqrt{L}} = \begin{cases} d_h - 1 \geq d_h/2, & d_h \geq 2, \\ \sqrt{K} \geq 1 = d_h, & d_h = 1, \end{cases} \qquad q = d_a - 1 \geq \frac{d_a}{2}.$$

Every Θ_v lies in $\mathcal{N}_T(\Theta_0)$ by Lemma E.1(i), and the class of censored-feedback algorithms is contained in the full-feedback class. It follows that

$$\begin{aligned}\inf_{\mathsf{Alg}\,(\text{censored})} \sup_{\Theta\in\mathcal{N}_T(\Theta_0)} \mathrm{Regret}_T^{\mathrm{TopicCB}}(\Theta) &\geq \inf_{\mathsf{Alg}\,(\text{full})} \sup_{v\in\mathcal{V}} \mathrm{Regret}_T^{\mathrm{TopicCB}}(\Theta_v) \\ &\geq \frac{c_0}{4}\, p_{\mathrm{loc}}^{K-1} d_h d_a \sqrt{\frac{T}{\kappa_0}}.\end{aligned} \tag{34}$$

Finally, at Θ_0 every feasible logit equals $c_{z_{<k}(C)}b$ with $c_{z_{<k}(C)} \in [1-\delta, 1]$. Since the logistic derivative is even and nonincreasing in $|u|$, we have $\dot{\mu}(c_{z_{<k}(C)}b) \geq \dot{\mu}(b)$. Equality is attained at $k = 1$ because $z_{<1}(C) = \emptyset$ and $c_\emptyset = 1$. Thus, by (29),

$$\kappa_\star(\Theta_0) = \frac{1}{\dot{\mu}(b)} = \kappa_0.$$

The reference parameter Θ_0 and the item and history features are independent of T, while $\varepsilon_T \to 0$ by construction. By Lemma E.1 and $\kappa_0 = \kappa_\star(\Theta_0)$, inequality (34) proves the theorem with the universal constant $c = c_0/4 = 9c_\omega/128$.

E.3 Proofs of Lemmas for Theorem 2

E.3.1 Proof of Lemma E.1

Proof. Only the first m rows of Θ_v differ from Θ_0, so for every $v \in \mathcal{V}$,

$$\|\Theta_v - \Theta_0\|_F = \sqrt{2}\,\omega q\sqrt{m} = \sqrt{2}\,c_\omega q\sqrt{\frac{m\kappa_0}{LT}} \leq \varepsilon_T.$$

Here $m/L \leq d_h$: it equals $d_h - 1$ when $d_h \geq 2$ and $1/K$ when $d_h = 1$. This proves (i).

For all $T \geq 32c_\omega^2 q^2 d_h \kappa_0/\rho^2$, our choice of radius satisfies $\varepsilon_T \leq \rho/4$. At Θ_0, every feasible logit is $c_{z_{<k}(C)}b$, and hence

$$\left|\phi(z_{<k}(C))^\top \Theta_0 x_{i_k} - b\right| = (1 - c_{z_{<k}(C)})|b| \leq \delta|b| \leq \frac{\rho}{4}.$$

For any $\Theta \in \mathcal{N}_T(\Theta_0)$,

$$\left|\phi(z_{<k}(C))^\top (\Theta - \Theta_0) x_{i_k}\right| \leq \|\phi(z_{<k}(C))\|_2 \|\Theta - \Theta_0\|_F \|x_{i_k}\|_2 \leq \varepsilon_T \leq \frac{\rho}{4}.$$

The triangle inequality then gives $|\phi(z_{<k}(C))^\top \Theta x_{i_k} - b| \leq \rho/2$ for every feasible cascade C and $k \in [K]$. This proves (ii).

Finally, the construction gives $\|x_{i_k}\|_2 = 1$ and $\|\phi(\zeta)\|_2 \leq 1$ for every feasible history ζ. For all $\Theta \in \mathcal{N}_T(\Theta_0)$,

$$\|\Theta\|_F \leq \|\Theta_0\|_F + \varepsilon_T \leq \sqrt{2d_h}|b| + \frac{\rho}{4} \leq \sqrt{2d_h}(|b| + \rho).$$

Thus Assumption 1 holds with $S = \sqrt{2d_h}(|b| + \rho)$, finite and independent of T. This proves (iii). □

E.3.2 Proof of Lemma E.2

Proof. Fix an arbitrary full-feedback algorithm Alg and let $\mathcal{T}_{t-1}$ denote the transcript before round t. Under Θ_v, the transcript law factors as

$$P_v(\mathcal{T}_T) = \prod_{t=1}^{T} \mathbb{P}^{\mathsf{Alg}}(C_t \mid \mathcal{T}_{t-1}) P_v(Y_t \mid \mathcal{T}_{t-1}, C_t).$$

Here $\mathbb{P}^{\mathsf{Alg}}(C_t \mid \mathcal{T}_{t-1})$ is the conditional probability assigned to the selected cascade by the algorithm. This action-selection rule is the same under Θ_v and $\Theta_{v^{(a)}}$ for any fixed past transcript, so its factors cancel in the likelihood ratio. The KL chain rule therefore yields

$$\mathrm{KL}(P_v \| P_{v^{(a)}}) = \sum_{t=1}^{T} \mathbb{E}_v \Big[\mathrm{KL}\big(P_v(Y_t \mid \mathcal{T}_{t-1}, C_t) \| P_{v^{(a)}}(Y_t \mid \mathcal{T}_{t-1}, C_t)\big) \Big]. \tag{35}$$

Conditional on the selected cascade and past information, $Y_{t,1}, \ldots, Y_{t,K}$ are independent Bernoulli variables. Thus the conditional feedback KL is the sum of position-wise Bernoulli divergences:

$$\mathrm{KL}(P_v \| P_{v^{(a)}}) = \sum_{t=1}^{T} \mathbb{E}_v \left[\sum_{k=1}^{K} \mathrm{KL}\left(\mathrm{Bern}(p_{v,t,k}) \| \, \mathrm{Bern}(p_{v^{(a)},t,k}) \right) \right], \tag{36}$$

where $p_{v,t,k} := \mu\big(\phi(z_{<k}(C_t))^\top \Theta_v x_{\sigma_{t,k}}\big)$.

Fix $a = (r, j)$. Only hard positions $k \in \mathcal{K}$ with $\min\{k, m\} = r$ can have different click probabilities under Θ_v and $\Theta_{v^{(a)}}$. When $d_h \geq 2$, only position $k = r$ is affected; when $d_h = 1$, all K positions are affected. In either case, at most L positions contribute to the feedback KL in each round. For an affected position and the same played cascade, let u, u' denote the logits under the two parameters. By (31),

$$|u - u'| = c_{z_{<k}(C_t)} \omega \big| \langle v_r - v_r^{(a)}, \sigma_{t,k} \rangle \big| = 2 c_{z_{<k}(C_t)} \omega \leq 2\omega.$$

For logistic Bernoulli observations, put $A(u) := \log(1 + e^u)$, so that $A'(u) = \mu(u)$ and $A''(u) = \dot{\mu}(u)$. A direct calculation and Taylor's theorem give

$$\begin{aligned} &\mathrm{KL}\big(\mathrm{Bern}(\mu(u)) \| \, \mathrm{Bern}(\mu(u'))\big) \\ &\quad = A(u') - A(u) - A'(u)(u' - u) = \frac{1}{2} \dot{\mu}(\xi)(u' - u)^2, \end{aligned}$$

where ξ lies between u and u'. Both logits lie in W by Lemma E.1(ii), so $\xi \in W$. By (28),

$$\mathrm{KL}\big(\mathrm{Bern}(\mu(u)) \| \, \mathrm{Bern}(\mu(u'))\big) \leq \frac{(u' - u)^2}{\kappa_0} \leq \frac{4\omega^2}{\kappa_0}.$$

Summing at most L nonzero position-wise KL terms per round, and using $\omega = c_\omega \sqrt{\kappa_0/(LT)}$, gives

$$\mathrm{KL}(P_v \| P_{v^{(a)}}) \leq \frac{4LT\omega^2}{\kappa_0} = 4c_\omega^2 \leq \frac{1}{8}.$$

□

E.3.3 Proof of Corollary E.1

Proof. For $v, \sigma \in \{-1, +1\}^q$, $q - \langle v, \sigma \rangle = 2 d_{\mathrm{H}}(v, \sigma)$, where d_{H} denotes Hamming distance. At each hard position k, the logit gap between the optimal and selected signs equals $2c_{z_{<k}(C_t)}\omega$ times $d_{\mathrm{H}}(v_{\min\{k,m\}}, \sigma_{t,k})$. Both logits lie in W by Lemma E.1(ii). The mean-value theorem, (28), and $c_\zeta \geq 1 - \delta \geq 3/4$ therefore imply

$$\mathrm{Regret}_T^{\mathrm{row}}(v) \geq \frac{3\omega}{4\kappa_0} \sum_{t=1}^{T} \sum_{k \in \mathcal{K}} d_{\mathrm{H}}(v_{\min\{k,m\}}, \sigma_{t,k}). \tag{37}$$

We use an Assouad-type coordinatewise testing argument [30, Lemma 2.12]. By Lemma E.2 and Pinsker's inequality, $\mathrm{TV}(P_v, P_{v^{(a)}}) \leq 1/4$. Fix a round t, a hard position $k \in \mathcal{K}$, and a coordinate

$j \in [q]$. Since the optimal sign vector at position k is $v_{\min\{k,m\}}$, the selected sign $\sigma_{t,k,j}$ serves as an estimate of $v_{\min\{k,m\},j}$. Let $r = \min\{k, m\}$, $a = (r, j)$, and define the event $E := \{\sigma_{t,k,j} = v_{r,j}\}$. Since $v_{r,j}^{(a)} = -v_{r,j}$, the definition of total variation gives the two-point testing inequality

$$\begin{aligned} P_v(\sigma_{t,k,j} \neq v_{r,j}) + P_{v^{(a)}}(\sigma_{t,k,j} \neq v_{r,j}^{(a)}) &= 1 - P_v(E) + P_{v^{(a)}}(E) \\ &\geq 1 - \mathrm{TV}(P_v, P_{v^{(a)}}) \\ &\geq \frac{3}{4}. \end{aligned}$$

For fixed $a = (r, j)$, the elements of $\mathcal{V}$ can be partitioned into $|\mathcal{V}|/2$ disjoint pairs $\{v, v^{(a)}\}$ that differ only in coordinate a. Pairing the terms of the uniform-prior average yields

$$\begin{aligned} \mathbb{E}_{v\sim\mathrm{Unif}(\mathcal{V})} P_v(\sigma_{t,k,j} \neq v_{r,j}) &= \frac{1}{|\mathcal{V}|} \sum_{\substack{v\in\mathcal{V} \\ v_{r,j}=+1}} \Big[P_v(\sigma_{t,k,j} \neq v_{r,j}) + P_{v^{(a)}}(\sigma_{t,k,j} \neq v_{r,j}^{(a)}) \Big] \\ &\geq \frac{1}{|\mathcal{V}|} \frac{|\mathcal{V}|}{2} \frac{3}{4} = \frac{3}{8}. \end{aligned}$$

By the definition of Hamming distance,

$$d_{\mathrm{H}}(v_{\min\{k,m\}}, \sigma_{t,k}) = \sum_{j=1}^{q} \mathbf{1}\{\sigma_{t,k,j} \neq v_{\min\{k,m\},j}\}.$$

Using linearity of expectation and summing over all rounds, hard positions, and coordinates gives

$$\begin{aligned} &\mathbb{E}_{v\sim\mathrm{Unif}(\mathcal{V})} \mathbb{E}_v \left[\sum_{t=1}^{T} \sum_{k\in\mathcal{K}} d_{\mathrm{H}}(v_{\min\{k,m\}}, \sigma_{t,k}) \right] \\ &= \sum_{t=1}^{T} \sum_{k\in\mathcal{K}} \sum_{j=1}^{q} \mathbb{E}_{v\sim\mathrm{Unif}(\mathcal{V})} P_v(\sigma_{t,k,j} \neq v_{\min\{k,m\},j}) \\ &\geq \frac{3}{8} T|\mathcal{K}|q. \end{aligned}$$

Combining this with (37), for every full-feedback algorithm Alg,

$$\mathbb{E}_{v\sim\mathrm{Unif}(\mathcal{V})} \mathbb{E}_v^{\mathsf{Alg}} \left[\mathrm{Regret}_T^{\mathrm{row}}(v) \right] \geq \frac{3\omega}{4\kappa_0} \frac{3}{8} T|\mathcal{K}|q = \frac{9\omega}{32\kappa_0} T|\mathcal{K}|q.$$

The supremum over v is at least the uniform-prior average, and the inequality holds for every algorithm. Hence

$$\inf_{\mathsf{Alg}} \sup_{v\in\mathcal{V}} \mathbb{E}_v^{\mathsf{Alg}} \left[\mathrm{Regret}_T^{\mathrm{row}}(v) \right] \geq \frac{9\omega}{32\kappa_0} T|\mathcal{K}|q = \frac{9c_\omega}{32} \frac{|\mathcal{K}|}{\sqrt{L}} q \sqrt{\frac{T}{\kappa_0}}.$$

Thus $c_0 = 9c_\omega/32$. □

E.3.4 Proof of Lemma E.3

Proof. Fix a round t and write the learner's cascade as $C_t = (i(g_{t,1}, \sigma_{t,1}), \ldots, i(g_{t,K}, \sigma_{t,K}))$. Define a comparison cascade $\widetilde{C}_t(v)$ by preserving the topic at every position and replacing only the hard signs:

$$\widetilde{\sigma}_{t,k} = \begin{cases} v_{\min\{k,m\}}, & k \in \mathcal{K}, \\ \sigma_{t,k}, & k \notin \mathcal{K}. \end{cases}$$

Each topic has an item with every sign vector, and the topics in C_t are distinct. Hence $\widetilde{C}_t(v) \in \Pi$. Furthermore, $z_{<k}(\widetilde{C}_t(v)) = z_{<k}(C_t)$ for every k, so the topic-history features are unchanged despite the sign replacements. Let $p_{t,k}$ and $\widetilde{p}_{t,k}$ be the click probabilities at position k of C_t and $\widetilde{C}_t(v)$ under Θ_v. By (31), these probabilities coincide when $k \notin \mathcal{K}$. For $k \in \mathcal{K}$,

$$\begin{aligned} \widetilde{p}_{t,k} - p_{t,k} = \mu\big(c_{z_{<k}(C_t)}(b + \omega q)\big) \\ - \mu\big(c_{z_{<k}(C_t)}(b + \omega\langle v_{\min\{k,m\}}, \sigma_{t,k}\rangle)\big) \geq 0. \end{aligned} \tag{38}$$

The final inequality follows from $\langle v_{\min\{k,m\}}, \sigma_{t,k}\rangle \le q$ and monotonicity of μ.

Recall $f_t(C;\Theta_v) = 1 - \prod_{k=1}^{K}(1 - p_{t,k}(C;\Theta_v))$. The product telescoping identity yields

$$\begin{aligned} f_t(\widetilde{C}_t(v);\Theta_v) - f_t(C_t;\Theta_v) &= \prod_{k=1}^{K}(1-p_{t,k}) - \prod_{k=1}^{K}(1-\widetilde{p}_{t,k}) \\ &= \sum_{k\in\mathcal{K}}(\widetilde{p}_{t,k} - p_{t,k}) \prod_{\ell<k}(1-\widetilde{p}_{t,\ell}) \prod_{\ell>k}(1-p_{t,\ell}). \end{aligned} \tag{39}$$

Both cascades are feasible, so their logits lie in W by Lemma E.1(ii). Since μ is increasing and $p_{\mathrm{loc}} = 1-\mu(b+\rho)$, all no-click factors in (39) are at least p_{loc}. Each product has $K-1$ such factors, with empty products interpreted as one. Thus

$$f_t(\widetilde{C}_t(v);\Theta_v) - f_t(C_t;\Theta_v) \ge p_{\mathrm{loc}}^{K-1} \sum_{k\in\mathcal{K}}(\widetilde{p}_{t,k} - p_{t,k}).$$

Finally, $C_t^\star(v)$ maximizes expected cascade reward and therefore has reward at least that of $\widetilde{C}_t(v)$. Summing over t and using (32) proves the claim. □

F Auxiliary Lemmas

This section collects the auxiliary results used in the concentration and regret proofs.

Lemma F.1 (Logistic derivatives and Taylor bound)**.** *For $\mu(z) = (1+e^{-z})^{-1}$ and every $z \in \mathbb{R}$,*

$$0 < \dot\mu(z) \le \tfrac{1}{4}, \qquad |\ddot\mu(z)| \le \dot\mu(z) \le \tfrac{1}{4}.$$

Consequently, for $h \ge 0$,

$$\mu(z+h) - \mu(z) \le \dot\mu(z)h + \tfrac{1}{8}h^2.$$

Proof. Differentiation gives $\dot\mu(z) = \mu(z)(1-\mu(z))$ and $\ddot\mu(z) = \dot\mu(z)(1-2\mu(z))$, proving the derivative bounds. For $h \ge 0$, Taylor's theorem gives

$$\mu(z+h) - \mu(z) - \dot\mu(z)h = \int_0^h (h-v)\ddot\mu(z+v)\,dv \le \frac{1}{4}\int_0^h (h-v)\,dv = \frac{1}{8}h^2.$$

□

Lemma F.2 (Curvature comparison)**.** *For all real u, v,*

$$e^{-|u-v|}\dot\mu(v) \le \dot\mu(u) \le e^{|u-v|}\dot\mu(v).$$

This is [2, Lemma 9], applied to the logistic link, which satisfies $|\ddot\mu| \le \dot\mu$.

Lemma F.3 (Norm shrinkage)**.** *For symmetric matrices $0 \prec A \preceq B$ and $\|x\|_2 \le 1$,*

$$0 \le \|x\|_{A^{-1}} - \|x\|_{B^{-1}} \le 2\{\mathrm{tr}(A^{-1/2}) - \mathrm{tr}(B^{-1/2})\}.$$

This is [15, Lemma 3]: write the positive semi-definite matrix $B - A$ as a sum of outer products and extend from unit vectors by scaling.

Lemma F.4 (Squared-norm shrinkage)**.** *For symmetric matrices $0 \prec A \preceq B$ and $\|x\|_2 \le 1$,*

$$0 \le \|x\|^2_{A^{-1}} - \|x\|^2_{B^{-1}} \le \mathrm{tr}(A^{-1}) - \mathrm{tr}(B^{-1}).$$

Proof. Since $B \succeq A \succ 0$, we have $A^{-1} - B^{-1} \succeq 0$. Therefore,

$$\begin{aligned} 0 \le \|x\|^2_{A^{-1}} - \|x\|^2_{B^{-1}} &= x^\top(A^{-1} - B^{-1})x \\ &\le \|x\|_2^2 \lambda_{\max}(A^{-1} - B^{-1}) \\ &\le \mathrm{tr}(A^{-1} - B^{-1}) \\ &= \mathrm{tr}(A^{-1}) - \mathrm{tr}(B^{-1}). \end{aligned}$$

□

Lemma F.5 (Clipped elliptical potential)**.** *Let $a > 0$, $\|y_j\|_2 \le L$, and $M_j = aI_d + \sum_{i<j} y_i y_i^\top$. For every finite sequence of length n,*

$$\sum_{j=1}^{n} \min\{1, \|y_j\|^2_{M_j^{-1}}\} \le 2 \log \frac{\det M_{n+1}}{\det M_1} \le 2d \log\left(1 + \frac{nL^2}{da}\right).$$

This is the standard elliptical potential lemma [1, Lemma 11]; a version valid for every $a > 0$ is also stated in [19, Lemma E.2]. The inequality is deterministic and applies to any realized subsequence. The minimum can be removed on any subset for which the squared norms are at most one.

G Details of Experiments

This appendix provides the detailed construction of the controlled environment used in our experiments, including the topic-history representation, item-feature generation, base parameter matrix, and seeded perturbation scheme. Throughout this appendix, π denotes the usual circle constant, not the policy notation used in the main text.

Maintained parameterization. All the experiments use the following parameters:

$$\alpha_{\text{pref}} = 1.4, \quad \lambda_{\text{topic}} = 0.65, \quad \rho = 1.8, \quad \eta_{\text{code}} = 0.03, \quad \sigma_x = 0.15. \quad \varepsilon_\Theta = 0.1$$

Topic basis. For each topic $g \in [G]$, define

$$\vartheta_g := \frac{2\pi(g-1)}{G}, \qquad b_g := \text{normalize}\left(\begin{bmatrix}\cos(\vartheta_g)\\ \sin(\vartheta_g)\\ \cos(2\vartheta_g)\end{bmatrix}\right) \in \mathbb{R}^3.$$

We assign each topic g an equally spaced angular position ϑ_g on the unit circle, with topics indexed modulo G. We then define b_g as the normalized angle-based embedding induced by ϑ_g, yielding a deterministic low-dimensional representation of the cyclic topic structure where the topics are placed on a circle that each topic has circular neighbors such as $g-1$ and $g+1$, with wrap-around between $G-1$ and 0. Below, the vectors $\{b_g\}_{g=1}^G$ are used in the history feature map, in the item-feature construction, and in the construction of the base parameter matrix.

Topic-history feature map. We use a deterministic topic history feature map ϕ_s that converts each topic history into a 3-dimensional state vector. The construction is dominated by the last topic in the history: for a non-empty history $p = (z_1, \ldots, g_\ell)$, the feature is centered at the basis vector of the last topic g_ℓ, together with a small shared baseline component. To keep distinct history distinguishable even when the history have the same last topic, we add a small history-specific code. Hence, the map remains injective over history.

We now give the explicit construction of the topic history feature map. Let $\mathcal{P}_K$ be the set of all topic history of length at most $K-1$, and let $M := |\mathcal{P}_K|$. Assign each history $p \in \mathcal{P}_K$ a index $\text{ind}(p) \in \{0, \ldots, M-1\}$. For $j = 1, 2, 3$, define the history-specific code

$$c(\text{ind}(p))_j := \cos\left(\frac{\pi(\text{ind}(p)+1)j}{M+1}\right).$$

Let $e_0 := (0, 0, 1)^\top$, and define $\text{clip}(u) := \frac{u}{\max\{1, \|u\|_2\}}$. Then the history feature map is

$$\phi(p) = \text{clip}\left(\begin{cases} e_0 + \eta_{\text{code}} c(\text{ind}(p)), & p = \emptyset, \\ \alpha_{\text{pref}} b_{g_\ell} + 0.35 e_0 + \eta_{\text{code}} c(\text{ind}(p)), & p = (z_1, \ldots, g_\ell) \neq \emptyset. \end{cases}\right)$$

The nonzero code scale $\eta_{\text{code}} > 0$ ensures that history are distinguishable, rather than depending only on the last topic.

Item topics and item features. Item topics are assigned in a balanced way so that every topic appears at least once, and then shuffled according to the random seed. Item features are generated around topic-specific representative directions, but include small item-specific and episode-specific

variation. This keeps the environment structured but not trivial. For each item $i \in [N]$ with topic z_i, define a topic-cluster center

$$m_i = \lambda_{\text{topic}} b_{z_i} + (1 - \lambda_{\text{topic}}) e_0 + 0.35 \sigma_x u_i,$$

where u_i is a normalized item-specific Gaussian offset. The center m_i is normalized to unit norm. At episode t, the item feature is sampled as

$$x_{t,i} = v_{t,i} \,\text{normalize}\left(m_i + \sigma_x \xi_{t,i}\right),$$

where $\xi_{t,i}$ is Gaussian noise and $v_{t,i} \sim \text{Unif}[0.8, 1.0]$. Thus $\|x_{t,i}\|_2 \leq 1$.

Base parameter matrix. We design the base parameter matrix so that the best next topic depends on the current topic history. In particular, after topic g, the model tends to prefer topic $g + 1$ over g, with a weaker preference toward $g + 2$. Let $\bar{b}$ be the normalized average topic direction, and define

$$a_0 := 1.2 b_1 + 0.3 \bar{b}.$$

For each topic $g \in [G]$, define

$$v_g := \rho b_{g+1} - 0.65 b_g + 0.25 b_{g+2} + \eta_{\text{code}} \bar{b},$$

where topic indices are taken modulo G. The base parameter matrix is

$$\Theta_{\text{base}} := e_0 a_0^\top + \sum_{g=1}^{G} b_g v_g^\top.$$

The coefficient ρ controls the strength of the history-dependent preference. If necessary, Θ_{base} is rescaled so that $\|\Theta_{\text{base}}\|_{\text{F}} \leq S$.

Seeded perturbation of $\Theta_\star$. In robustness experiments, we use a seeded perturbation of the base matrix:

$$\Theta_\star = \Theta_{\text{base}} + \varepsilon_\Theta \Delta_{\text{seed}}.$$

In the seeded setting, each random seed uses a slightly perturbed version of the same base matrix $\Theta_\star$. The purpose is to verify that the results are not tied to a single hand-crafted parameter matrix, while still preserving the same overall favorable structure. The perturbation Δ_{seed} is obtained from a Gaussian matrix by row normalization followed by Frobenius normalization. If the perturbed matrix violates the norm bound, it is rescaled to satisfy $\|\Theta_\star\|_{\text{F}} \leq S$.

Compute Resources. All experiments were conducted on a single MacBook Pro equipped with an Apple M1 Pro chip (8 CPU cores; 6 performance and 2 efficiency cores) and 32 GB unified memory, running macOS 15.7.3. Our implementation is CPU-only and was executed in a Conda environment with Python 3.10.20, NumPy 2.2.5, SciPy 1.15.3, Matplotlib 3.10.8, and scikit-learn 1.7.2. All reported runtimes were measured on this machine.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the paper's main contributions: the CTCB model with topic-history-dependent click probabilities, the topic-history MDP reformulation and optimism-based algorithm, the regret upper bound with a decreasing leading dependence on the cascade length, the local lower bound, and the numerical experiments. These claims are supported by the formal results and experiments presented in the paper; the K-dependence claim is stated for the leading horizon-dependent regret term, consistent with the theorem statement.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations through its modeling assumptions and scope, including the known topic labels, the no-topic-repetition constraint.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states the assumptions for each theoretical result, including the boundedness assumptions for the regret upper bound and the prefix-rich, non-contextual subclass assumptions for the local lower bound. Complete proofs of the main theorems and supporting lemmas are provided in the appendix, with proof sketches and technical ingredients summarized in the main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the synthetic environment, baselines, evaluation metrics, and parameter choices used in the experiments, with further construction details deferred to the appendix. These details are sufficient to reproduce the main experimental results supporting the empirical claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The supplementary material includes code and instructions for reproducing the main experimental results. The experiments use synthetic environments whose construction is fully specified, so no separate external dataset access is required.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental setup, synthetic environment construction, baselines, hyperparameters, and evaluation metrics needed to understand the results. Additional implementation details and reproducibility instructions are provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The experiments report results averaged over multiple random seeds with standard-deviation error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The supplementary material reports the compute environment and approximate runtime needed to reproduce the experiments. The experiments are lightweight synthetic simulations and can be reproduced on standard CPU-based hardware.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is theoretical and experimental work on bandit algorithms using synthetic environments, and does not involve human subjects, private data, or deployed decision systems. We have reviewed the NeurIPS Code of Ethics and do not identify any violations.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Since this paper is primarily a theoretical study of bandit algorithms with supporting synthetic experiments, there is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Since this paper is primarily a theoretical study of bandit algorithms with supporting synthetic experiments, this paper does not release high-risk data or models such as pre-trained language models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM is used only for writing, editing, or formatting purposes

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.