Bandit Learning with Implicit Feedback

Anonymous Author(s)
Affiliation
Address
email

Abstract

Implicit feedback, such as user clicks, although abundant in online information service systems, does not provide substantial evidence on users' evaluation of system's output. Such incomplete supervision inevitably misleads model estimation, especially in a bandit learning setting where the feedback is acquired on the fly. In this work, we study a contextual bandit problem with implicit feedback by modeling the feedback as a composition of user result examination and relevance judgment. Since users' examination behavior is unobserved, we introduce latent variables to model it. We perform Thompson sampling on top of variational Bayesian inference for arm selection and model update. Rigorous upper regret bound analysis of the proposed algorithm proves its feasibility of learning from implicit feedback; and extensive empirical evaluations on click logs collected from a major MOOC platform further demonstrate its learning effectiveness in practice.

1 Introduction

Contextual bandit algorithms [3, 18, 17] provide modern information service systems an effective solution to adaptively find good mappings between available items and users. This family of algorithms sequentially select items to serve users using side information about user and item, while adapting their selection strategies based on the immediate user feedback to maximize users' long-term satisfaction. They have been popularly deployed in practical systems for content recommendation [18, 4, 24] and display advertising [5, 20].

However, the most dominant form of user feedback in such systems is implicit feedback, such as clicks, which is known to be biased and incomplete about users' evaluation of system's output [15, 10]. For example, a user skips a recommended item might not be because he/she does not like the item, but he/she just does not examine that display position, i.e., position bias [12]. Unfortunately, a common practice in contextual bandit applications simply treats no click as a form of negative feedback [18, 23, 5]. This introduces inconsistency to model update, since the skipped items might not be truly irrelevant, and it inevitably leads to suboptimal outputs of bandit algorithms over time.

In this work, we focus on learning contextual bandits with user click feedback, and model such implicit feedback as a composition of user result examination and relevance judgment. Examination hypothesis [7], which is a fundamental assumption in click modeling, postulates that a user clicks on a system's returned result *if and only if* that result has been examined by the user and it is relevant to the user's information need at the moment. Because a user's examination behavior is unobserved, we model it as a latent variable, and realize the examination hypothesis in a probabilistic model. We define the conditional probabilities of result examination and relevance judgment via logistic functions over the corresponding contextual features. To perform model update, we take a variational Bayesian approach to develop closed form approximation to the posterior distribution of model parameters on the fly. This approximation also paves the way for an efficient Thompson sampling strategy for arm selection in bandit learning. Our rigorous finite time analysis proves that, despite the increased complexity in parameter estimation introduced by the latent variable, our Thompson

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

sampling policy based on the variational lower bound is guaranteed to achieve a sub-linear Bayesian regret with a high probability. We also prove that when one fails to model result examination in click feedback, a linearly increasing regret is possible, as the model cannot differentiate examination driven skip from relevance driven skip in negative feedback.

We tested the algorithm in a major Massive Open Online Course (MOOC) platform for personalized education. To personalize students' learning experience in this platform, we recommend quiz-like questions in a form of banners on top of the lecture videos when students are watching the videos. The algorithm needs to decide where in a video to display which question to a target student. If the student feels the displayed question is helpful for him/her to understand the video lecture, he/she could click on the banner to read the answer and more related online content about the question. Therefore, our goal is to maximize the click through rate (CTR) on the selected questions. There are several properties of this application that amplifies the bias and incompleteness of click feedback. First, based on the consideration of user experience, to minimize the risk of annoying any student, the displayed time of a banner is limited to a few seconds. Second, as this feature is newly introduced to the platform, many users might not realize that they can click on the question to read more related content about it. Hence, no click on a question does not necessarily indicate its irrelevance. We tested the algorithm in this application in a four-month period, where a total of 500 questions are manually compiled for the algorithm to select over 503 videos within more than 200 thousands student video-watching sessions. Based on the unbiased offline evaluation policy [19], our algorithm achieved a 7.4% CTR lift compared to standard contextual bandits [18, 8] which do not model users' examination behavior. By sharing user examination model across videos, the improvement is further boosted to over 8.8%.

2 Related Works

As having been extensively studied in click modeling of user search results [6], various factors affect users' click decisions, and among them result examination plays a central role [12, 7]. Unfortunately, most applications of bandit algorithms simply treat user clicks as feedback for model update [18, 23, 5, 24], where no click on a selected result is considered as negative feedback. This inevitably leads to inaccurate model update and sub-optimal arm selection.

There is a line of research that develops click model based bandit algorithms for learning to rank problems. For example, by assuming that skipped documents are less attractive than later clicked ones in a ranked list, Kveton et al. [16] develop a cascading bandit model to learn from both clicks and skips in search results. To enable learning from multiple clicks in the same result ranking list, they adopt the dependent click model [9] to infer user satisfaction after a sequence of clicks [13], and later further extend to broader types of click models [25]. However, such algorithms aim at estimating the best ranking of results in a per-query basis, without specifying any ranking function nor generalizing to unseen queries. This limits their application scenario in practice.

Another line of related research is bandit learning with latent variables. Maillard and Mannor studied the problem of latent bandit [21], which assumes reward distributions are clustered and the clusters are determined by some latent variables. They only studied the problem in a context-free setting, and a very weak performance guarantee is provided when the reward distribution is unknown in those clusters. Kawale et al. developed a Thompson sampling scheme for online matrix-factorization [14]. Latent features are extracted via an online low-rank matrix completion based on samples selected from Thompson sampling on the fly. Due to the ad-hoc combination of factorization method and bandit method, little theoretical analysis is provided. Wang et al. studied the problem of latent feature learning for contextual bandits [23]. They extended arms' context vectors with latent features under a linear reward structure, and applied the upper confidence bound principle over coordinate descent to iteratively estimate the hidden features and model parameters. The linear reward structure prohibits it from recognizing the nonlinear dependency between result examination and relevance judgment in click feedback.

3 Problem Setup

We consider a contextual bandit problem with finite, but possibly large, number of arms $\mathcal{A}$. At each trial $t = 1, ..., T$, the learner observes a subset of candidate arms $\mathcal{A}_t$ with $\mathcal{A}_t \subset \mathcal{A}$, where each arm

a is associated with a context vector $\mathbf{x}^a$ summarizing the side information about the arm. Once an arm $a_t \in \mathcal{A}_t$ is chosen according to some policy π, corresponding implicit binary feedback C_{a_t}, e.g., user click, will be given to the learner as the reward. The learner's goal is to adjust its arm selection strategy to maximize its cumulative reward over time. What makes this problem unique and challenging is that C_{a_t} does not truly reflect users' evaluation of the selected arm a_t. Based on the examination hypothesis, when $C_{a_t} = 1$, the choice a_t must be relevant to the user's information need at time t; but when $C_{a_t} = 0$, a_t might be relevant but the user just does not examine it. Unfortunately, the unobserved result examination condition is unobserved to the learner.

We model user examination through a binary latent variable E_{a_t} and assume that the context vector $\mathbf{x}_t^a$ of arm a can be decomposed into $(\mathbf{x}_{C,t}^a, \mathbf{x}_{E,t}^a)$, where the dimension of $\mathbf{x}_{C,t}^a$ and $\mathbf{x}_{E,t}^a$ are d_C and d_E respectively. Accordingly, users' result examination and relevance judgment decisions are assumed to be governed by a conjecture of $(\mathbf{x}_{C,t}^a, \mathbf{x}_{E,t}^a)$ and the bandit parameter $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_C^*, \boldsymbol{\theta}_E^*)$. In the rest of this paper, when no ambiguity is introduced, we drop the index a to simplify the notations. As a result, we make the following generative assumption about an observed click C_t on arm a_t,

$$\begin{aligned}\mathbb{P}(C_t = 1|E_t = 0, \mathbf{x}_{C,t}) &= 0\\ \mathbb{P}(C_t = 1|E_t = 1, \mathbf{x}_{C,t}) &= \rho(\mathbf{x}_{C,t}^\mathsf{T}\boldsymbol{\theta}_C^*)\\ \mathbb{P}(E_t = 1|\mathbf{x}_{E,t}) &= \rho(\mathbf{x}_{E,t}^\mathsf{T}\boldsymbol{\theta}_E^*)\end{aligned}$$

where $\rho(x) = \frac{1}{1+e^{-x}}$. Based on this assumption, we have $\mathbb{E}[C_t|\mathbf{x}_t] = \rho(\mathbf{x}_{C,t}^\mathsf{T}\boldsymbol{\theta}_C^*)\rho(\mathbf{x}_{E,t}^\mathsf{T}\boldsymbol{\theta}_E^*)$. As a result, the observed click feedback C_t is a sample from this generative process.

4 Algorithm

The learner needs to estimate the bandit parameters $\boldsymbol{\theta}_C^*$ and $\boldsymbol{\theta}_E^*$ based on its interactively obtained click feedback $\{C_i\}_{i=1}^t$ over time. Ideally, this estimation can be obtained by maximizing the data likelihood. However, the inclusion of examination as a latent variable to model implicit feedback in our bandit setting poses serious challenges to both parameter estimation and arm selection policy design. Neither conventional least square estimator nor maximum likelihood estimator can be easily obtained, let alone computational efficiency, due to the non-convexity of the corresponding optimization problem. Even worse, the two popular bandit learning paradigms, upper confidence bound principle [1] and Thompson sampling [2], both demand an accurate estimation of bandit parameters and its uncertainty. In this section, we present an elegant solution to tackle these two challenges, which makes use of variational Bayesian inference technique to learn parameters approximately on the fly, as well as to bridge parameter estimation and policy design.

4.1 Variational Bayesian for parameter estimation

To complete the generative process defined in Section 3, we further assume $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_E$ follow Gaussian distribution $N(\hat{\boldsymbol{\theta}}_C, \boldsymbol{\Sigma}_C)$ and $N(\hat{\boldsymbol{\theta}}_E, \boldsymbol{\Sigma}_E)$ respectively. We are interested in developing a closed form approximation to their posteriors, when a newly obtained observation $(\mathbf{x}_C, \mathbf{x}_E, C)$ becomes available. By applying Bayes' rule in the log space, we have,

$$\begin{aligned}\log \mathbb{P}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_E|\mathbf{x}_C, \mathbf{x}_E, C) &= \log \mathbb{P}(C|\boldsymbol{\theta}_C, \boldsymbol{\theta}_E, \mathbf{x}_C, \mathbf{x}_E) + \log \mathbb{P}(\boldsymbol{\theta}_C, \boldsymbol{\theta}_E) + \log \text{const}\\ &= C \log \rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E) + (1 - C) \log \left(1 - \rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)\right)\\ &- \frac{1}{2}(\boldsymbol{\theta}_C - \hat{\boldsymbol{\theta}}_C)^\mathsf{T}\boldsymbol{\Sigma}_C^{-1}(\boldsymbol{\theta}_C - \hat{\boldsymbol{\theta}}_C) - \frac{1}{2}(\boldsymbol{\theta}_E - \hat{\boldsymbol{\theta}}_E)^\mathsf{T}\boldsymbol{\Sigma}_E^{-1}(\boldsymbol{\theta}_E - \hat{\boldsymbol{\theta}}_E) + \log \text{const}\end{aligned}$$

The key idea is to develop a variational lower bound in the quadratic form of $\boldsymbol{\theta}_C$ and $\boldsymbol{\theta}_E$ for the log-likelihood function $\log \mathbb{P}(C|\boldsymbol{\theta}_C, \boldsymbol{\theta}_E, \mathbf{x}_C, \mathbf{x}_E)$. Because of the convexity of $\log \rho(x) - \frac{x}{2}$ with respect to x^2 (See Appendix B.1) and a Jensen's inequality of $\log x$ (See Appendix B.2), the lower bound in required form is achievable. When $C = 1$, by Equation 15 in Appendix B.3, we have,

$$l_{C=1}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}) := \log \left(\rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)\right) \geq g(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}, \xi_C) + g(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}, \xi_E) \tag{1}$$

where $g(x, \xi) := \frac{x}{2} - \frac{\xi}{2} + \log \rho(\xi) - \lambda(\xi)(x^2 - \xi^2)$, $\lambda(\xi) = \frac{\tanh \frac{\xi}{2}}{4\xi}$, $x, \xi \in \mathcal{R}$. More specifically, $g(x, \xi)$ is a polynomial of degree 2 with respect to x. When $C = 0$, by Equation 16 in Appendix B.3,

Algorithm 1 Thompson sampling for E-C Bandit

1: Initiate $\mathbf{\Sigma}_C = \lambda I, \mathbf{\Sigma}_E = \lambda I, \hat{\boldsymbol{\theta}}_C = \boldsymbol{\theta}_{C,0}, \hat{\boldsymbol{\theta}}_E = \boldsymbol{\theta}_{E,0}$.
2: **for** $k = 0, 1, 2...$ **do**
3: Observe the available arm set $\mathcal{A}_k \subset \mathcal{A}$ and its corresponding context set $\mathcal{X}_k := \{(\mathbf{x}_C^a, \mathbf{x}_E^a) : a \in \mathcal{A}_k\}$.
4: Randomly sample $\tilde{\boldsymbol{\theta}}_C \sim N(\hat{\boldsymbol{\theta}}_C, \mathbf{\Sigma}_C)$, $\tilde{\boldsymbol{\theta}}_E \sim N(\hat{\boldsymbol{\theta}}_E, \mathbf{\Sigma}_E)$.
5: Select:

$$a_k = \arg\max_{a \in \mathcal{A}_k} \rho((\mathbf{x}_C^a)^\mathsf{T} \tilde{\boldsymbol{\theta}}_C) \rho((\mathbf{x}_E^a)^\mathsf{T} \tilde{\boldsymbol{\theta}}_E)$$

6: Play the selected arm a_k and Observe the reward C_k.
7: Update $\mathbf{\Sigma}_C, \hat{\boldsymbol{\theta}}_C, \mathbf{\Sigma}_E, \hat{\boldsymbol{\theta}}_E$ according to Eq (3), (4), (5), (6) respectively.
8: **end for**

we have,

$$\begin{aligned} l_{C=0}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}) &:= \log\left(1 - \rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)\right) \\ &\geq H(q) + qg(-\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}, \xi_C) + qg(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}, \xi_{E,1}) + (1-q)g(-\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}, \xi_{E,2}) \end{aligned} \tag{2}$$

where $H(q) := -q \log q - (1-q)\log(1-q)$. Once the lower bound in the quadratic form is established, we can use a Gaussian distribution to approximate our target posterior, whose mean and covariance matrix are determined by the following equations:

$$\mathbf{\Sigma}_{C,\text{post}} = \mathbf{\Sigma}_C + 2q^{1-C}\lambda(\xi_C)\mathbf{x}_C\mathbf{x}_C^\mathsf{T} \tag{3}$$

$$\hat{\boldsymbol{\theta}}_{C,\text{post}} = \mathbf{\Sigma}_{C,\text{post}}^{-1}(\mathbf{\Sigma}_C\hat{\boldsymbol{\theta}}_C + \frac{1}{2}(-q)^{1-C}\mathbf{x}_C) \tag{4}$$

$$\mathbf{\Sigma}_{E,\text{post}} = \mathbf{\Sigma}_E + 2\lambda(\xi_E)\mathbf{x}_E\mathbf{x}_E^\mathsf{T} \tag{5}$$

$$\hat{\boldsymbol{\theta}}_{E,\text{post}} = \mathbf{\Sigma}_{E,\text{post}}^{-1}(\mathbf{\Sigma}_E\hat{\boldsymbol{\theta}}_E + \frac{1}{2}(2q-1)^{1-C}\mathbf{x}_E) \tag{6}$$

where the subscript "post" denotes the parameters in the Gaussian distributions that approximate the posteriors. Consecutive observations can be incorporated into the approximated posterior recursively. There is one thing left to decide, i.e., the choice of variational parameters (ξ_C, ξ_E, q). A reasonable criterion is to choose the values such that likelihood on the observation is maximized. Similar to the choice made by [11], we choose the closed form update formulas of those variational parameters as:

$$\begin{aligned} \xi_C &= \sqrt{\mathbf{E}_{\boldsymbol{\theta}_C}[\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C]^2} \\ \xi_E &= \sqrt{\mathbf{E}_{\boldsymbol{\theta}_E}[\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E]^2} \\ q &= \frac{\exp\left(g(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C, \xi_C) + g(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E, \xi_E) - g(-\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E, \xi_E)\right)}{1 + \exp\left(g(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C, \xi_C) + g(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E, \xi_E) - g(-\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E, \xi_E)\right)} \end{aligned}$$

where all the expectations are taken under the approximated posteriors. Empirically, we found the iterative update of the approximated posterior and the variational parameters converge quite rapidly, such that it usually only needs several rounds of iterations in our experiments.

4.2 Thompson sampling with approximated lower bound

Thompson sampling, also known as probability matching, is widely used in bandit learning to balance exploration and exploitation, and shows great empirical performance [5]. Thompson sampling requires a distribution of the model parameters to sample from. In a standard Thompson sampling [2], one is required to sample from the true posterior of model parameters. But as logistic regression does not have a conjugate prior, the model defined in our problem does not have an exact posterior. We decide to sample from the approximated posterior. Later we will demonstrate this is a very tight posterior approximation in our experiment. Once the sampling of $(\tilde{\boldsymbol{\theta}}_C, \tilde{\boldsymbol{\theta}}_E)$ is complete, we can select the corresponding arm $a_k \in \mathcal{A}_k$ which maximizes $\rho(\mathbf{x}_C^\mathsf{T}\tilde{\boldsymbol{\theta}}_C)\rho(\mathbf{x}_E^\mathsf{T}\tilde{\boldsymbol{\theta}}_E)$. We name the resulting bandit algorithm as examination-click bandit, or E-C Bandit in short, and summarize it in Algorithm 1.

5 Regret Analysis

Define $f_{\boldsymbol{\theta}}(\mathbf{x}) := \mathbb{E}[C|\mathbf{x}, \boldsymbol{\theta}] = \rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)$. The accumulated regret of a bandit policy π up to time T is formally defined as,

$$\text{Regret}(T, \pi, \boldsymbol{\theta}^*) = \sum_{t=1}^{\mathsf{T}} \max_{a \in \mathcal{A}_t} f_{\boldsymbol{\theta}^*}(\mathbf{x}^a) - f_{\boldsymbol{\theta}^*}(\mathbf{x}^{a_t})$$

where $\mathbf{x}^{a_t} := (\mathbf{x}_C^{a_t}, \mathbf{x}_E^{a_t})$ is the context vector of the arm $a_t \in \mathcal{A}_t$ selected by the policy π at time t based on the history $\mathcal{H}_t := \{(\mathcal{A}_i, \mathbf{x}_i, C_i)\}_{i=1}^{t-1}$. The Bayesian regret is defined by $\mathbb{E}\big[\text{Regret}(T, \pi, \theta^*)\big]$, where the expectation is taken with respect to the prior distribution over $\boldsymbol{\theta}^*$, and can be written as,

$$\text{BayesRegret}(T, \pi) = \sum_{t=1}^{\mathsf{T}} \mathbb{E}\big[\max_{a \in \mathcal{A}_t} f_{\boldsymbol{\theta}^*}(\mathbf{x}^a) - f_{\boldsymbol{\theta}^*}(\mathbf{x}^{a_t})\big]$$

Our algorithm, which is based on a maximum likelihood estimator, is equivalent to an estimator that minimizes a log-loss with binary random variables. In this section, we will first bound the aggregate empirical discrepancy of the log-loss estimator used in our model in Proposition 1, which prepares for the upper bound of the generic Bayeisan regret under log-loss estimators with Thompson sampling policy in Theorem 1. Based on this generic Bayesian regret bound under a log-loss estimator, we study the upper bound of Bayesian regret for our proposed E-C Bandit. Due to space limit, we provide all the detailed proofs in the Appendix.

To further simplify our notations, we use f for $f_{\boldsymbol{\theta}}$, which is the reward function based on a model's estimated bandit parameter $\boldsymbol{\theta}$, and f_k for $f_{\boldsymbol{\theta}}(\mathbf{x}^{a_k})$. We use f^* for $f_{\boldsymbol{\theta}^*}$, which is the reward function based on the ground-truth bandit parameter, and f_k^* for $f_{\boldsymbol{\theta}^*}(\mathbf{x}^{a_k})$. We assume that f^* lies in a known function space $\mathcal{F}$, where any $f \in \mathcal{F}$ is a function mapping from the arm set $\mathcal{A}$ to the range $(0, 1)$. Define the log-loss estimator by $\hat{f}_t^{\text{LOGLOSS}} \in \arg\max_{f \in \mathcal{F}} L_{2,t}(f)$ where $L_{2,t}(f)$ is the aggregate log-loss written as $\sum_{k=1}^{t-1} l_k(f)$ where $l_k(f) = -\big(C_k \log f_k + (1 - C_k)\log(1 - f_k)\big)$. We have the following lemma,

Proposition 1. Denote the aggregate empirical discrepancy $\sum_{k=1}^{t}(f_k - f_k^*)^2$ by $\|f - f^*\|_{E,t}$. For all $\delta > 0$ and $\alpha > 0$, if $\mathcal{F}_t = \{f \in \mathcal{F} : \left\|f - \hat{f}_t^{\text{LOGLOSS}}\right\|_{E,t} \leq \sqrt{\beta_t^*(\mathcal{F}, \delta, \alpha)}\}$ for all $t \in N$,

$$\mathbb{P}\big(f^* \in \cap_{t=1}^{\infty} \mathcal{F}_t\big) > 1 - \delta, \tag{7}$$

where $\beta_t^*(\mathcal{F}, \delta, \alpha)$ is an appropriately constructed confidence parameter and is defined as $\beta_t^*(\mathcal{F}, \delta, \alpha) := \frac{2}{\lambda_0}\log(N(\mathcal{F}, \alpha, \|\cdot\|_\infty)/\delta) + 2\alpha\eta_t$, where $N(\mathcal{F}, \alpha, \|\cdot\|_\infty)$ denotes the alpha-covering number of $\mathcal{F}$, $\lambda_0 = \frac{3}{(\frac{1}{m_f} + \frac{1}{1-M_f})^2}$ and $\eta_t = \big(4M_f + \frac{1}{\min\{m_f, 1-M_f\}}\big)t$, in which $m_f, M_f \in \mathcal{R}$ are upper and lower bounds of f such at $0 < m_f \leq f \leq M_f < 1$ for any $f \in \mathcal{F}$.

Remark 1. The proof is provided in Appendix C. Here we discuss two important issues related to our later proof about E-C Bandit's regret. First, the precise optimization of $\hat{f}_t^{\text{LOGLOSS}} \in \arg\max_{f \in \mathcal{F}} L_{2,t}(f)$ could be hard to achieve in some instances of $\mathcal{F}$, for example, when $\mathcal{F}$ is a set of non-convex functions. Nevertheless, we can always resort to approximation method to solve the optimization problem as long as the approximation error can be bounded. Indeed, in our E-C Bandit, we resort to variational inference to estimate $\hat{f}_t^{\text{LOGLOSS}}$ on the fly and find it works quite well in practice. Second, when $f^* \notin \mathcal{F}$, this corresponds to the problem of mis-specification of model. In this situation, the regret bound could be very poor, and the real regret could be linear with respect to time. To show this clearly in our case, in Appendix F we constructed a situation in which the regret is inevitably linear if one fails to model the examination condition in click feedback and simply treats no click as negative feedback.

With Proposition 1, we have the following Theorem which bounds the Bayesian regret of the Thompson Sampling strategy under a log-loss estimator.

Theorem 1. For all $T \in N$, $\alpha > 0$ and $\delta < \frac{1}{2T}$, if π^{TS} denotes the policy derived from the log-loss estimator and a Thompson sampling strategy along the time steps, then

$$\text{BayesRegret}(T, \pi^{TS}) \leq 1 + \big(\dim_E^{\mathcal{A}}(\mathcal{F}, T^{-1}) + 1\big)C + 4\sqrt{\dim_E^{\mathcal{A}}(\mathcal{F}, T^{-1})\beta_T^*(\mathcal{F}, \alpha, \delta)T} \tag{8}$$

where $C = \sup_{f \in \mathcal{F}}\{\sup f\}$, $\dim_E^{\mathcal{A}}(\mathcal{F}, T^{-1})$ is the eluder dimension (see Definition 3 in Russo and Van Roy [22]) of $\mathcal{F}$ with respect to $\mathcal{A}$.

Remark 2. We can choose $C = 1$ in our click feedback case since $f \in (0, 1)$. C is kept in the theorem to show the same form compared to the Proposition 8 in Russo and Van Roy [22]. In fact, the proof is almost the same once we have Proposition 1. Hence, we omit the proof in our paper.

Now we turn to provide an upper regret bound of our E-C Bandit, based on the above generic Bayeisan regret analysis under a log-loss estimator. We add the following assumptions which are standard in the literature of contextual bandits.

Assumption 1. The optimal $\boldsymbol{\theta}^*$ lies in $\mathcal{B}_s := \{\boldsymbol{\theta} \in \mathcal{R}^d : \|\boldsymbol{\theta}\|_2 \leq s\}$, and s is known as a prior.

Assumption 2. The norm of context vectors are bounded by x, i.e., $(\mathbf{x}_C, \mathbf{x}_E) \in \mathcal{B}_x$, where $\mathcal{B}_x := \{\mathbf{x} \in \mathcal{R}^d : \|\mathbf{x}\|_2 \leq x\}$ and x is known as a prior.

Based on these two assumptions, it is straightforward to verify that $\rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)$, $\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)$ and $f_{\boldsymbol{\theta}}(\mathbf{x})$ are bounded. Let $M_\rho = \max_{\boldsymbol{\theta}\in\mathcal{B}_s, \mathbf{x}\in\mathcal{B}_x} \rho(\mathbf{x}^\mathsf{T}\boldsymbol{\theta})$ and $m_\rho = \min_{\boldsymbol{\theta}\in\mathcal{B}_s, \mathbf{x}\in\mathcal{B}_x} \rho(\mathbf{x}^\mathsf{T}\boldsymbol{\theta})$. Hence, $0 < m_\rho \leq M_\rho < 1$. Similarly, denoting the maximum of $f_{\boldsymbol{\theta}}(\mathbf{x})$ by M_f and the minimum of it by m_f, we have $0 < m_f \leq M_f < 1$. Once the arm set is restricted to a finite cardinality, $\dim_E^{\mathcal{A}}(\mathcal{F}, T^{-1}) \leq |\mathcal{A}|$ by Appendix C.1. in Russo and Van Roy [22]. Choosing the function class as that in our E-C Bandit, i.e., $\mathcal{F} = \{f : \mathcal{B}_x \to \mathcal{R} | f = \rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E), \boldsymbol{\theta} \in \mathcal{B}_s\}$, by Lemma 8 (See Appendix 8 for its proof), we have $N(\mathcal{F}, \alpha, \|\cdot\|_\infty) = (\gamma/\alpha)^d$ where $\gamma = 2M_\rho k_\rho x$ (k_ρ is the Lipschitz constant of ρ, see Lemma 4). Hence, choosing $\alpha = 1/t^2$ and $\delta = 1/t$ leads to

$$\beta_t^*(\mathcal{F}, 1/t, 1/t^2) = \frac{2}{\lambda_0} d \log(\gamma t^3) + \frac{1}{t}(4M_f + \frac{1}{m_f}). \tag{9}$$

Therefore, the upper bound of Bayesian regret of our proposed E-C Bandit takes the following form,

$$\text{BayesRegret}(T, \pi^{TS}) = O(|\mathcal{A}| + \sqrt{d|\mathcal{A}|T \log T}). \tag{10}$$

When $T \gg |\mathcal{A}|$ and $T \gg d$, which is the typical case in practice, it becomes $O(\sqrt{T \log T})$.

6 Experiment

We perform empirical comparisons in simulation and on the student click logs collected from our MOOC platform to verify the effectiveness of our proposed algorithm, against those that fail to model the examination condition and directly use click as feedback.

6.1 Algorithms for comparison

We list the models used for empirical comparisons below. We briefly explain why we choose them as baseline models and how we adjust them in our scenario. **Logistic Bandit.** This model has been extensively used for advertisement CTR optimization. In [5, 19], the authors model user clicks by a regularized logistic regression model over observed context features and makes decisions by Thompson sampling. In particular, no click is treated as negative feedback. We also used the Laplace approximation and Gaussian prior presented in [5] to update the model parameters on the fly. We also want to highlight that despite of mountains of works focusing on generalized linear bandits, most of them are not truly online algorithms, because the estimation of parameters at each iteration has to involve all historical observations iteratively, which incurs a time complexity at least $O(t)$ (e.g., Filippi et al. [8] require exact optimum of logistic regression on all historical observations at each round). **E-C Bandit.** This is the algorithm we present in Algorithm 1. We highlight here that in the experiment on real-world data, the manual separation of examination feature $\mathbf{x}_E$ and click feature $\mathbf{x}_C$ in the context vector $\mathbf{x}$ offers a principled approach to incorporate one's domain knowledge about what affects user examination and what affects user result relevance into the model. We explain in detail what features are chosen for which in Appendix G. Thanks to the approximation achieved by Bayesian variational inference presented in section 4, truly online model update is feasible in this algorithm. This provides both computational and storage efficiency. **hLinUCB.** This is the bandit algorithm proposed by Wang et al. [23]. It is related to our model in a sense that both models learn hidden features. In particular, hLinUCB extends the linear contextual bandit by inclusion of hidden features and operates under a UCB-like strategy. However, it still treats click as direct feedback, but aims at learning more expressive features to describe the observed clicks.

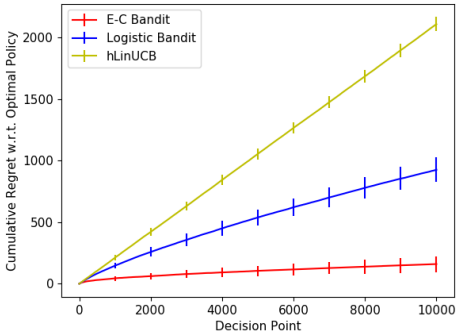

Figure 1: Cumulative regret over 100 simulations.

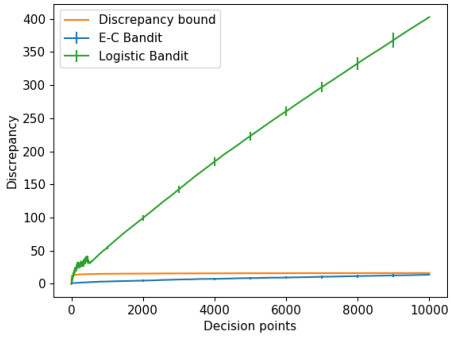

Figure 2: Discrepancy bound given by Proposition 1

6.2 Experiments on simulations

First we demonstrate the effectiveness of our algorithms by experiment with simulated data. The experiment is set as follows. The context vector's dimension d_C and d_E are set to 5, and thus $d = d_C + d_E = 10$. We set $|\mathcal{A}| = 100$, each of which is associated with a unique context vector $(\mathbf{x}_C, \mathbf{x}_E)$. The ground-truth parameter $(\boldsymbol{\theta}^*_C, \boldsymbol{\theta}^*_E)$ and the specific value of $(\mathbf{x}_C, \mathbf{x}_E)$ are all randomly sampled from the unit ball $\mathcal{B} = \{\mathbf{x} \in \mathcal{R}^d : \|\mathbf{x}\|_2 \leq 1\}$. Since $(\boldsymbol{\theta}^*_C, \boldsymbol{\theta}^*_E)$ and $(\mathbf{x}_C, \mathbf{x}_E)$ are both sampled from $\mathcal{B}$, m_f and M_f can be obtained by taking the minimum and maximum of $\rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)$ on $\mathcal{B}$, respectively, i.e., $m_f = \frac{1}{(1+e)^2}$ and $M_f = \frac{1}{(1+e^{-1})^2}$.

At each time t, an available arm set $\tilde{\mathcal{A}}_t$ is randomly sampled from $\mathcal{A}$ such that $|\tilde{\mathcal{A}}_t| = 10$, i.e., each time we offer 10 randomly selected arms for the algorithm to choose from. Any algorithm selects an arm from $\mathcal{A}_t$ and observes the corresponding reward C_t^{alg} generated by the Bernoulli distribution $B(\rho(\mathbf{x}_{C,t}^\mathsf{T}\boldsymbol{\theta}^*_C)\rho(\mathbf{x}_{E,t}^\mathsf{T}\boldsymbol{\theta}^*_E))$. The actual regret for this algorithm at time t is defined by the received click, i.e., $\text{regret}(t) = C_{a_t^*} - C_{a_t}$, where a_t^* is the optimum arm to be chosen based on ground-truth bandit parameters $(\boldsymbol{\theta}^*_C, \boldsymbol{\theta}^*_E)$.

We repeat the experiment 100 times using the same simulation setting, each containing 10,000 iterations. The average cumulative regret over 100 runs and the corresponding standard deviation (plotted per thousand decision point) are illustrated in Figure 1. One can clearly notice that while the Logistic Bandit suffers from a linear regret with respect to time t, as it mistakenly treats no click as negative feedback, our E-C Bandit achieves a fast converging sub-linear regret. The result that hLinUCB performs the worst is expected, since it assumes a linear relation between click and context feature vectors. We further investigate how the aggregate empirical discrepancy of E-C bandit and Logistic Bandit increase with respect to time. Figure 2 illustrates that the aggregate empirical discrepancy of E-C Bandit is well bounded by the upper bound given by Proposition 1, while the Logistic Bandit's aggregate empirical discrepancy increase linearly. This directly explains their accumulative regret in this experiment comparison.

6.3 Experiments on MOOC video watching data

The MOOC data we used for evaluation is collected from a single course in a 4-month period. The course has 503 lecture videos in total. About 500 high-quality quiz-like questions has been manually crafted and each video is assigned with a subset of them based on human-judged relatedness and fitness. On average a video has 7 questions, each of which is associated with six possible displaying positions within the video, leading to a total of 56 arms in average.

We picked one video as an example to analyze the students' click behavior. 11 arms are randomly picked and projected to a plane according to their relative distance in Figure 3. The number in the parenthesis is the empirical CTR of the corresponding arm. It can be clearly seen that while arm b and arm j have the same empirical CTR, the arms between them, such as are e or d, have lower CTR. Logistic Bandit is never able to capture this non-monotonicity relation, since its reward prediction increases monotonically with respect to a linear predictor. We construct a more general case in Appendix F to illustrate the scenario that failing to model examination condition would

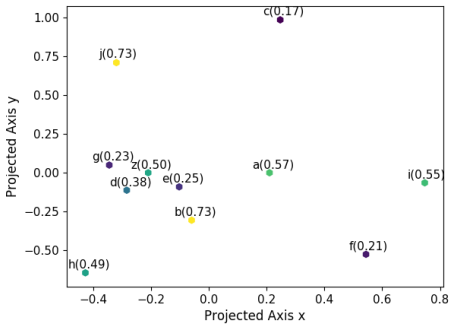

Figure 3: 11 arms' feature vectors projected onto a plane, such that the relative distances between points are kept. The number in the parenthesis is the arm's relative CTR.

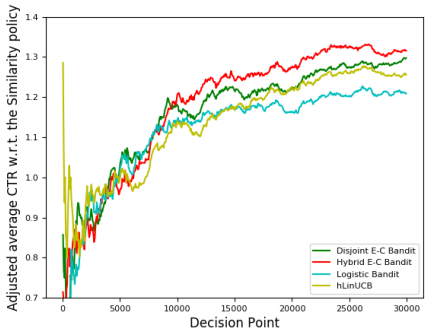

Figure 4: Performance comparison on MOOC videos' data

lead to a linear regret. In reality, arm b and arm j are two different questions displayed at the same position in the video, while arm b and arm e are the same question displayed at different positions. This phenomenon strongly suggests a bias in users' implicit feedback, which again justifies our decomposition of examination and relevance in click feedback.

We followed [19] to develop our online data collection policy in our MOOC platform to prepare our evaluation data set. In particular, any related questions with respect to a video will have an equal probability to be selected and displayed at all positions in this video. We name this policy as $Similarity$. We create an instance of a bandit model for each video to learn its own optimal question placing policy. Further, we follow [18] to create a hybrid model of our E-C Bandit across multiple videos. We make the examination component in E-C Bandit, i.e., $\boldsymbol{\theta}_E$, shared by all videos. Our underlying assumption is that while how the relevance quality of each question to the host video affects student clicks may vary across videos, the students' examination decision should be mostly affected by the platform or their own watching habits, e.g., our user interface design. (See Appendix G for detailed explanations of our examination feature choice.) We denote the non-hybrid algorithm by disjoint E-C Bandit and the hybrid one by Hybrid E-C Bandit. It is worth noting that there is no straightforward way to construct hybrid algorithm for Logistic Bandit, as it only has one set of parameter for modeling click. And hLinUCB is inherently hybrid since different videos share information through the learned hidden features, i.e., a collaborative filtering scheme.

Li et al. [19] proposed a method to calculate a near-unbiased estimate of the average CTR by any bandit algorithm based on the collected history data, so that offline evaluation and performance comparison are possible. We take their offline evaluation protocol here and report the estimated average CTR in Figure 4. To avoid disclosing any proprietary information about the platform, all algorithms' CTRs are normalized by that from the $Similarity$ policy. As shown in the figure, independently estimating E-C Bandits across videos achieves an average 29.9% increase in CTR over the $Similarity$ baseline, while the joint estimation further boosts it to 31.6%. These two versions of E-C Bandit consistently outperform other two baseline bandits, i.e., hLinUCB and Logistic Bandit. This clearly suggests the necessity of modeling examination condition in user clicks for improving the online recommendation performance.

7 Conclusion

Motivated by the examination hypothesis in user click modeling, in this paper we developed E-C Bandit, which differentiates result examination and content relevance in user clicks and actively learns from such implicit feedback. We developed an efficient and effective learning algorithm based on variational inference and demonstrated its effectiveness on both simulated and real-world datasets. We proved that despite the complexity of underlying reward generation assumption and the resulting parameter estimation procedure, the proposed learning algorithm enjoys a sub-linear regret bound. Currently we only studied click feedback on single items; it is important for us to study it in a more general setting, e.g., a list of ranked items, where sequential result examination and relevance judgment introduce richer inter-dependency.

References

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, pages 2312–2320. 2011.

[2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[3] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002. URL `http://www.jmlr.org/papers/v3/auer02a.html`.

[4] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. A contextual-bandit algorithm for mobile context-aware recommender system. In *Neural Information Processing*, pages 324–331. 2012.

[5] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[6] Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke. Click models for web search. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 7(3):1–115, 2015.

[7] 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. ACM, 2008.

[8] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594, 2010.

[9] Fan Guo, Chao Liu, and Yi Min Wang. Efficient multiple-click models in web search. In *Proceedings of the Second ACM International Conference on WSDM*, pages 124–131. ACM, 2009.

[10] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 263–272. Ieee, 2008.

[11] Tommi S Jaakkola and Michael I Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

[12] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, volume 51, pages 4–11. Acm, 2017.

[13] 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, 2016.

[14] Jaya Kawale, Hung H Bui, Branislav Kveton, Long Tran-Thanh, and Sanjay Chawla. Efficient thompson sampling for online matrix-factorization recommendation. In *NIPS*, pages 1297–1305, 2015.

[15] Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, volume 37, pages 18–28. ACM, 2003.

[16] 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.

[17] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, pages 817–824, 2008.

[18] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[19] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

[20] Wei Li, Xuerui Wang, Ruofei Zhang, Ying Cui, Jianchang Mao, and Rong Jin. Exploitation and exploration in a performance based contextual advertising system. In *Proceedings of 16th SIGKDD*, pages 27–36. ACM, 2010.

[21] Odalric-Ambrym Maillard and Shie Mannor. Latent bandits. In *International Conference on Machine Learning*, pages 136–144, 2014.

[22] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[23] Huazheng Wang, Qingyun Wu, and Hongning Wang. Learning hidden features for contextual bandits. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1633–1642. ACM, 2016.

[24] Yisong Yue and Carlos Guestrin. Linear submodular bandits and their application to diversified retrieval. In *NIPS*, pages 2483–2491, 2011.

[25] Masrour Zoghi, Tomas Tunys, Mohammad Ghavamzadeh, Branislav Kveton, Csaba Szepesvari, and Zheng Wen. Online learning to rank in stochastic click models. In *International Conference on Machine Learning*, pages 4199–4208, 2017.

Appendix A Preliminaries

In this section, we present some basic definitions and inequalities for later use.
First, let $p, q \in (0,1)$, $H(p) := -p\log p - (1-p)\log(1-p)$ and $KL(p|q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ denote the entropy and the KL-divergence function of Bernoulli variable respectively. Note that for later derivations' convenience, the $\log$ here is with respect to e instead of 2, which is slightly different from a standard definition.

The first inequality is Pinsker's inequality in Bernoulli case.

Lemma 1 (Pinsker's inequality)**.** $KL(p|q) - 2(p-q)^2 \geq 0$.

Proof. Let $c(q) := KL(p|q) - 2(p-q)^2$. We have $c'(q) = (p-q)(\frac{1}{q(1-q)} - 4)$. Since $q \in (0,1)$, $q(1-q) \leq 1/4$ by the arithmetic average inequality. Thus, when $q < p, c'(q) <= 0$; when $q > p$, $c'(q) >= 0$. Hence, $c(q) \geq c(p) = 0$. □

The next inequality provides a Lipschitz condition-like equality for $\log$ function defined on bounded set.

Lemma 2. Given $a, b \in (0,1)$, for any $x, y \in [a,b]$, $\frac{1}{b}|x-y| \leq |\log x - \log y| \leq \frac{1}{a}|x-y|$.

Proof. According to the mean value theorem, for any $x, y \in [a,b]$, there exists a $\eta \in [x,y]$ such that

$$\log x - \log y = \frac{1}{\eta}(x-y).$$

The inequality follows by taking absolute value on two side and the fact that η is bounded in $[a,b]$. □

The next is Hoeffding's lemma.

Lemma 3 (Hoeffding's lemma)**.** Let X be any real-valued random variable such that $E(X) = 0$ and $a \leq X \leq b$ almost surely. Then, for all $\lambda \in \mathcal{R}$,

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \lambda^2(b-a)^2/8.$$

We omit its proof here, since its proof should be available on any standard textbook about the theory of probability. The last inequality deals with the Lipschitz condition of the logistic function.

Lemma 4. For the logistic function $\rho(x) = \frac{1}{1+e^{-x}}$, we have:

$$|\rho(x) - \rho(y)| < k_\rho|x-y|$$

where $k_\rho = 1/4$.

Proof. According to te mean value theorem, there exists a $\eta \in (x,y)$ such that

$$\rho(x) - \rho(y) = \rho(\eta)(1-\rho(\eta))(x-y)$$

where we have used $\dot{\rho}(x) = \rho(x)(1-\rho(x))$. Since $\rho(\eta) \in (0,1)$, we have $\rho(\eta)(1-\rho(\eta)) \leq 1/4$. Thus,

$$|\rho(x) - \rho(y)| = |\rho(\eta)(1-\rho(\eta))| \times |(x-y)| \leq k_\rho|x-y|.$$

□

Appendix B Variational lower bound for E-C Bandit

The variational lower bound in the quadratic form of our E-C bandit is constructed from two basic variational lower bounds. They are the lower bound to a logistic regression and the variational lower bound to a log-sum function. In this section, we first provide these two basic lower bounds in Section B.1 and Section B.2 respectively. We end this section by providing the desired lower bound for our E-C bandit in Section B.3.

B.1 Variational lower bound for the log-logistic function

Jaakkola and Jordan [11] provide for the log-logistic function a variational lower bound taking the form of a 2-degree polynomial based on the observation that $\log \rho(x) - \frac{x}{2}$ is convex with respect to x^2. We report their result here. Defining the related functions as follows:

$$\rho(x) := \frac{1}{1+e^{-x}} (x \in \mathcal{R})$$
$$\lambda(\xi) := \frac{\tanh \frac{\xi}{2}}{4\xi}$$
$$g(x, \xi) := \frac{x}{2} - \frac{\xi}{2} + \log \rho(\xi) - \lambda(\xi)(x^2 - \xi^2) (x, \xi \in \mathcal{R})$$

we have the variational lower bound:

$$\log(\rho(x)) \geq g(x, \xi) \tag{11}$$

When $\xi^2 = x^2$, the lower bound is exact. Note that $1 - \rho(x) = \rho(-x)$. Thus:

$$\log(1 - \rho(x)) \geq g(-x, \xi) \tag{12}$$

The equation holds when $\xi^2 = x^2$.

B.2 Variational lower bound for a log-sum function

We are interested in developing a bound for the following log-sum function, which is exactly the likelihood function of a single sample with reward $C = 0$ in E-C Bandit:

$$\log \big((1 - \rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C))\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E) + 1 - \rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)\big). \tag{13}$$

Since $\log(x)$ is convex, we have the following inequality holds for any $q \in (0, 1)$:

$$\begin{aligned} &\log \Big(\big((1 - \rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C))\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)\big) + 1 - \rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E) \Big) \\ &\geq H(q) + q \log \big((1 - \rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C))\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)\big) + (1 - q) \log \big(1 - \rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)\big) \end{aligned} \tag{14}$$

where $H(q) = -q \log q - (1 - q) \log(1 - q)$. It can be easily verified using Jensen's inequality for a convex function.

The equality holds whenever $q = \frac{\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E) - \rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)\rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C)}{1 - \rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)\rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C)}$, which always lies in the domain of q, i.e., $(0, 1)$, since $\rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C) \in (0, 1)$ and $\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E) \in (0, 1)$.

B.3 Variational lower bound for the log-likelihood of E-C Bandit

We provide the variational lower bound of a quadratic form for the log-likelihood of a single observation in E-C Bandit in this section. When $C = 1$, the log-likelihood of a single sample is:

$$l_{C=1}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E) = \log(\rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E))$$

Since $\log(x)$ is additive, given Equation 11, we have:

$$\begin{aligned} l_{C=1}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E) &= \log(\rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C)) + \log(\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)) \\ &\geq g(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C, \xi_C) + g(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E, \xi_E) \\ &=: \tilde{l}_{C=1}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E, \xi_C, \xi_E) \end{aligned} \tag{15}$$

where $=:$ means "denoted by". When $C = 0$, the log-likelihood of a single sample is a log-sum function, which is the first lower-bounded by Equation 14. Then each log-logistic function is lower-bounded by Equation 11 and Equation 12. The derivation is as follows:

$$\begin{aligned} l_{C=0}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E) &= \log(1 - \rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)) \\ &= \log((1 - \rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C))\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E) + 1 - \rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)) \\ &\geq H(q) + q \log((1 - \rho(\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C))\rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)) + (1 - q) \log(1 - \rho(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E)) \\ &\geq H(q) + qg(-\mathbf{x}_C^\mathsf{T} \boldsymbol{\theta}_C, \xi_C) + qg(\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E, \xi_{E,1}) + (1 - q)g(-\mathbf{x}_E^\mathsf{T} \boldsymbol{\theta}_E, \xi_{E,2}) \end{aligned} \tag{16}$$

where $\xi_C, \xi_{E,1}, \xi_{E,2}$ are the variational parameter of $\log(1-\rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)), \log(\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)), \log(1-\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E))$ respectively. Note that the equality holds when $\xi_{E,1}^2 = (\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)^2$ and $\xi_{E,2}^2 = (-\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E)^2$. We can always impose $\xi_{E,1} = \xi_{E,2}$ without harming the attainability of equality in the lower-bound. Thus, we can get a simplified lower-bound:

$$\begin{aligned} l_{C=0}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E) &\geq H(q) + qg(-\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C, \xi_C) + qg(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E, \xi_E) + (1-q)g(-\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E, \xi_E) \\ &=: \tilde{l}_{C=0}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E, \xi_C, \xi_E, q) \end{aligned} \tag{17}$$

We can unify Equation 15 and Equation 17 as:

$$\tilde{l}_C(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E) := C\tilde{l}_{C=1}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E) + (1-C)\tilde{l}_{C=0}(\mathbf{x}_C, \mathbf{x}_E, \boldsymbol{\theta}_C, \boldsymbol{\theta}_E, \xi_C, \xi_E, q)$$

which is a quadratic form of $(\boldsymbol{\theta}_C, \boldsymbol{\theta}_E)$ as desired (Plug in the definition of $g(x)$ to see it clearly).

Appendix C Proof of confidence bound: Proposition 1

Proposition 1 is established by a union lower bound of aggregate empirical discrepancy of any function $f \in \mathcal{F}$ (Lemma 5) with respect to the optimal f^* and a discretization error bound based on an α-covering of the function space $\mathcal{F}$ (Lemma 6). In this section, we first present the proof of these two lemmas first, and end this section by a proof of Proposition 1 based on these two lemmas.

C.1 Bound of log-loss estimator by aggregate empirical discrepancy

We lower-bound the log-loss of any function $f \in \mathcal{F}$ in terms of the empirical log-loss of the true function f^* and the aggregate empirical discrepancy $\|f-f^*\|_{E,t} := \sum_{i=1}^t (f_k - f_k^*)^2$. The result is stated as the following lemma.

Lemma 5. For all $\delta > 0$ and $f \in \mathcal{F}$, with probability at least $1-\delta$,

$$\mathbb{P}\left(L_{2,t+1}(f) \geq L_{2,t+1}(f^*) + \frac{1}{2}\|f-f^*\|_{E,t} - \frac{1}{\lambda_0}\log\frac{1}{\delta} \right) \geq 1-\delta \tag{18}$$

holds simultaneously for all natural number $t \in N$, where $\lambda_0 := \frac{3}{(\frac{1}{m_f} + \frac{1}{1-M_f})^2}$.

We first recall some definitions and properties that are helpful here. Let $\mathcal{H}_t^x = \mathcal{H}_t \cup \{\mathcal{A}_t, \mathbf{x}_t\}$ be a new filtration. First, $C_k \in \{0,1\}$ is a binary variable whose expectation conditioned on $\mathcal{H}_k^x$ is f_k^*. Hence, the random variable $\zeta_k := C_k - f_k^*$ is sub-Gaussian according to Lemma 7 such that:

$$\begin{aligned} \mathbb{E}\big[\zeta_k | \mathcal{H}_k^x\big] &= 0 \\ \mathbb{E}\big[e^{\lambda\zeta_k} | \mathcal{H}_k^x\big] &\leq \frac{\lambda^2}{2} \end{aligned}$$

The log-loss of a single sample is defined as $l_k(f) = -\big(C_k \log f_k + (1-C_k)\log(1-f_k)\big)$. The KL-divergence of two binary variables whose expectation are p and q respectively is defined by $KL(p|q) := p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$, and satisfied Pinsker's Inequality(Lemma 1):

$$KL(p|q) \geq 2(p-q)^2$$

Proof. We add the following definition for simplicity of notation:

$$\begin{aligned} h_k &:= \log f_k - \log f_k^* + \log(1-f_k^*) - \log(1-f_k) \\ Z_k &:= l_k(f^*) - l_k(f) \\ \Phi_k(\lambda) &:= \log \mathbb{E}\big[e^{\lambda(Z_t - E(Z_t))} | \mathcal{H}_k^x\big]. \end{aligned}$$

By definition, $\sum_{i=1}^t Z_k = L_{2,t+1}(f^*) - L_{2,t+1}(f)$. Plugging in the equalities $C_k = \zeta_k + f_k^*$ and $l_k(f) = -\big(C_k \log f_k + (1-C_k)\log(1-f_k)\big)$, we can get that:

$$Z_k = h_k\zeta_k - KL(f_k^*|f_k). \tag{19}$$

Therefore, $\mathbb{E}\big[Z_k|\mathcal{H}_k^x\big] = -KL(f_k^*|f_k)$ and $\Phi_k(\lambda) = \log \mathbb{E}\big[e^{\lambda h_k \zeta_k}|\mathcal{H}_k^x\big]$. Hence, by Martingale Exponential Inequality (See Appendix B.1. in Russo and Van Roy [22]), for all $x \geq 0, \lambda \geq 0$,

$$\mathbb{P}\left(\lambda \sum_{i=1}^{t} Z_k \leq x - \lambda \sum_{i=1}^{t} KL(f_k^*|f_k) + \sum_{i=1}^{t} \log \mathbb{E}\big[e^{\lambda h_k \zeta_k}|\mathcal{H}_k^x\big]\right) \geq 1 - e^{-x}. \tag{20}$$

Since $KL(p|q) \geq 2(p-q)^2$ according to Lemma 1, and $\mathbb{E}\big[e^{\lambda \zeta_k}|\mathcal{H}_k^x\big] \leq \frac{\lambda^2}{2}$ according to Lemma 7, we have,

$$\mathbb{P}\left(\lambda \sum_{i=1}^{t} Z_k \leq x - 2\lambda \sum_{i=1}^{t} (f_k - f_k^*)^2 + \sum_{i=1}^{t} \lambda^2 h_k^2/2\right) \geq 1 - e^{-x}. \tag{21}$$

Again, according to Lemma 2, $|h_k| \leq |\log f_k - \log f_k^*| + |\log(1-f_k^*) - \log(1-f_k)| \leq (\frac{1}{m_f} + \frac{1}{1-M_f})|f_k - f_k^*|$. Hence,

$$\mathbb{P}\left(\lambda \sum_{i=1}^{t} Z_k \leq x - 2\lambda \sum_{i=1}^{t} (f_k - f_k^*)^2 + \sum_{i=1}^{t} \lambda^2 (\frac{1}{m_f} + \frac{1}{1-M_f})^2 (f_k - f_k^*)^2/2\right) \geq 1 - e^{-x}. \tag{22}$$

Plugging in the definition of aggregate empirical discrepancy and Z_t, and rearranging the terms we get,

$$\mathbb{P}\left(L_{2,t+1}(f^*) - L_{2,t+1}(f) \leq \frac{x}{\lambda} + \big((\frac{1}{m_f} + \frac{1}{1-M_f})^2 \frac{\lambda}{2} - 2\big)\|f - f^*\|_{E,t}\right) \geq 1 - e^{-x}. \tag{23}$$

Choosing $\lambda = \lambda_0 := \frac{3}{(\frac{1}{m_f} + \frac{1}{1-M_f})^2}$ and $x := \log \frac{1}{\delta}$ results in

$$\mathbb{P}\left(L_{2,t+1}(f) \geq L_{2,t+1}(f^*) + \frac{1}{2}\|f - f^*\|_{E,t} - \frac{1}{\lambda_0} \log \frac{1}{\delta}\right) \geq 1 - \delta. \tag{24}$$

Note that $\lambda_0 > 0$ thus satisfies the assumption of Martingale Exponential Inequality's assumption. Also note that the deduction above makes no assumption about the value of $t \in N$, thus the desired lemma is proved. □

C.2 Discretization error

Lemma 6. If f^α satisfies $\|f - f^\alpha\|_\infty \leq \alpha$, and $m_f > 0, M_f < 1$, we have

$$\left|\frac{1}{2}\|f^\alpha - f^*\|_{E,t} - \frac{1}{2}\|f - f^*\|_{E,t} + L_{2,t}(f) - L_{2,t}(f^\alpha)\right| \leq \alpha \eta_t \tag{25}$$

where $\eta_t := (4M_f + \frac{1}{\min\{m_f, 1-M_f\}})t$.

Proof. It is sufficient to consider a single sample's discretization error and then sum them up over time t. Thus we omit the subscript t in the following derivation.

We have

$$\begin{aligned} \big|(f^\alpha - f^*)^2 - (f - f^*)^2\big| &= |(f^\alpha - f)(f^\alpha + f) + 2f^*(f - f^\alpha)| \\ &\leq |(f^\alpha - f)(f^\alpha + f)| + |2f^*(f - f^\alpha)| \\ &\leq 2M_f\alpha + 2M_f\alpha \\ &= 4M_f\alpha \end{aligned} \tag{26}$$

and

$$\begin{aligned} |l(f) - l(f^\alpha)| &= |C(\log f^\alpha - \log f) + (1-C)(\log(1-f^\alpha) - \log(1-f))| \\ &\leq C|(\log f^\alpha - \log f)| + |(1-C)(\log(1-f^\alpha) - \log(1-f))| \\ &\leq \frac{C}{m_f}|f^\alpha - f| + \frac{1-C}{1-M_f}|f^\alpha - f| \\ &\leq \frac{1}{\min\{m_f, 1-M_f\}}\alpha \end{aligned} \tag{27}$$

Summing over t and using triangular inequality, we get the desired result. □

C.3 Proof of Proposition 1

Proof. Let $\mathcal{F}^\alpha$ be an α-covering of $\mathcal{F}$ with respect to the sup norm such that for any $f \in \mathcal{F}$, there exists an $f^\alpha \in \mathcal{F}^\alpha$ such that $\|f - f^\alpha\|_\infty \leq \alpha$. By a union bound to Lemma 5, with probability at least $1-\delta$,

$$L_{2,t}(f^\alpha) - L_{2,t}(f^*) \geq \frac{1}{2}\|f^\alpha - f^*\|_{E,t} - \frac{1}{\lambda_0}\log(|\mathcal{F}|^\alpha/\delta)\ \forall t \in N, f^\alpha \in \mathcal{F}^\alpha.$$

Thus, with probability at least $1-\delta$ for all $t \in N$ and $f \in \mathcal{F}$,

$$\begin{aligned} L_{2,t}(f) - L_{2,t}(f^*) \geq & \frac{1}{2}\|f - f^*\|_{E,t} - \frac{1}{\lambda_0}\log(|\mathcal{F}^\alpha|/\delta) \\ & + \underbrace{\min_{f^\alpha \in \mathcal{F}^\alpha}\left\{\frac{1}{2}\|f^\alpha - f^*\|_{E,t} - \frac{1}{2}\|f - f^*\|_{E,t} + L_{2,t}(f) - L_{2,t}(f^\alpha)\right\}}_{\text{Discretization error}} \end{aligned} \tag{28}$$

Since $\hat{f}_t^{\text{LOGLOSS}} \in \arg\min_{f\in\mathcal{F}} L_{2,t}(f)$ and $f^* \in \mathcal{F}$, we have

$$L_{2,t}(\hat{f}_t^{\text{LOGLOSS}}) - L_{2,t}(f^*) \leq 0. \tag{29}$$

Using the two-side bound of the dicretization error term established by Lemma 6, we find with probability at least $1-\delta$,

$$\frac{1}{2}\left\|\hat{f}_t^{\text{LOGLOSS}} - f^*\right\|_{E,t} \leq \frac{1}{\lambda_0}\log(|\mathcal{F}|^\alpha/\delta) + \alpha\eta_t \tag{30}$$

Taking the infimum over the size of α-covers implies

$$\left\|f^* - \hat{f}_t^{\text{LOGLOSS}}\right\|_{E,t} \leq \sqrt{\beta_t^*(\mathcal{F},\delta,\alpha)}$$

where $\beta_t^*(\mathcal{F},\delta,\alpha) := \frac{2}{\lambda_0}\log(N(\mathcal{F},\alpha,\|\cdot\|_\infty)/\delta) + 2\alpha\eta_t$. □

Appendix D The sub-Gaussian property of $C_t - f_t^*$

In this section, we prove $C_t - f_t^*$ is a sub-Gaussian random variable.

Lemma 7. Let $\mathcal{H}_t^x = \mathcal{H}_t \cup \{\mathcal{A}_t, \mathbf{x}_t\}$ be a new filtration constructed on top of $\mathcal{H}_t$, where $\mathcal{H}_t := \{(\mathcal{A}_i, \mathbf{x}_i, C_i)\}_{i=1}^{t-1}$ and C_t is a binary random variable such that $\mathbb{E}\big[C_t|\mathbf{x}_t\big] = f_t^*$. The random variable $\zeta_t := C_t - f_t^*$ is σ-sub-Gaussian conditioned on $\mathcal{H}_t^x$. That is to say, we have

$$\mathbb{E}\big[\zeta_t|\mathcal{H}_t^x\big] = 0$$

and there exists a $\sigma > 0$ such that

$$\mathbb{E}\big[e^{\lambda\zeta_t}|\mathcal{H}_t^x\big] \leq e^{\frac{\lambda^2\sigma^2}{2}}$$

for any $\lambda \in \mathcal{R}$.

Proof. Since $\mathbb{E}\big[C_t|\mathbf{x}_t\big] = f_t^*$ and $f_t^* \in \mathcal{H}_t^x$, we have $\mathbb{E}\big[\zeta_t|\mathcal{H}_t^x\big] = \mathbb{E}\big[C_t - f_t^*|\mathcal{H}_t^x\big] = 0$. Since ζ_t is bounded in $(-1,1)$, by Hoeffding's lemma,

$$\mathbb{E}\big[e^{\lambda\zeta_t}|\mathcal{H}_t^x\big] \leq e^{\frac{\lambda^2}{2}}.$$

Thus conditioned on $\mathcal{H}_t^x$, ζ_t is σ-sub-Gaussian, where one can choose $\sigma = 1$. □

Appendix E Alpha-covering number of function space

We provide an α-covering number for the function space used in E-C Bandit.

Lemma 8. Let $\boldsymbol{\theta} := [\boldsymbol{\theta}_C^\mathsf{T}, \boldsymbol{\theta}_E^\mathsf{T}]^\mathsf{T} \in \mathcal{B}_s, \mathbf{x} := [\mathbf{x}_C^\mathsf{T}, \mathbf{x}_E^\mathsf{T}]^\mathsf{T} \in \mathcal{B}_x$, where $\mathcal{B}_s := \{\boldsymbol{\theta} \in \mathcal{R}^d : \|\boldsymbol{\theta}\|_2 \leq s\}$ and $\mathcal{B}_x := \{\mathbf{x} \in \mathcal{R}^d : \|\mathbf{x}\|_2 \leq x\}$. Let $\mathcal{F} = \{f : \mathcal{B}_x \to \mathcal{R} | f = \rho(\mathbf{x}_C^\mathsf{T}\boldsymbol{\theta}_C)\rho(\mathbf{x}_E^\mathsf{T}\boldsymbol{\theta}_E), [\boldsymbol{\theta}_C^\mathsf{T}, \boldsymbol{\theta}_E^\mathsf{T}]^\mathsf{T} \in \mathcal{B}_s\}$. Let $\gamma = 2M_\rho k_\rho x$. We have that

$$N(\mathcal{F},\alpha,\|\cdot\|_\infty) = (\gamma/\alpha)^d$$

holds for any $\alpha > 0$.

Proof. For any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{B}_s$, we have

$$
\begin{aligned}
&\|f(\boldsymbol{\theta}_1) - f(\boldsymbol{\theta}_2)\| \\
&= \left\|\rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,1})\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,1}) - \rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,1})\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,2}) + \rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,1})\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,2}) - \rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,2})\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,2})\right\| \\
&\le \left\|\rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,1})\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,1}) - \rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,1})\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,2})\right\| + \left\|\rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,1})\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,2}) - \rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,2})\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,2})\right\| \\
&\le M_\rho\left(\left\|\rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,1}) - \rho(\mathbf{x}_E^\top \boldsymbol{\theta}_{E,2})\right\| + \left\|\rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,1}) - \rho(\mathbf{x}_C^\top \boldsymbol{\theta}_{C,2})\right\|\right) \\
&\le M_\rho k_\rho \left\|\mathbf{x}_E^\top(\boldsymbol{\theta}_{E,1} - \boldsymbol{\theta}_{E,2})\right\| + \left\|\mathbf{x}_C^\top(\boldsymbol{\theta}_{C,1} - \boldsymbol{\theta}_{C,2})\right\| \\
&\le 2M_\rho k_\rho x \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \\
&= \gamma\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|
\end{aligned}
\tag{31}
$$

where all norms are infinity norm. Thus a α-covering of $\mathcal{F}$ can be achieved by a (α/γ)-covering of $\mathcal{B}_s$. Evenly divide $\mathcal{B}_s$ in each dimension, we get the desired result that $N(\mathcal{F}, \alpha, \|\cdot\|_\infty) = (\gamma/\alpha)^d$, where we omit the ceiling function for simplicity. □

Appendix F Linear regret of a mis-specified model

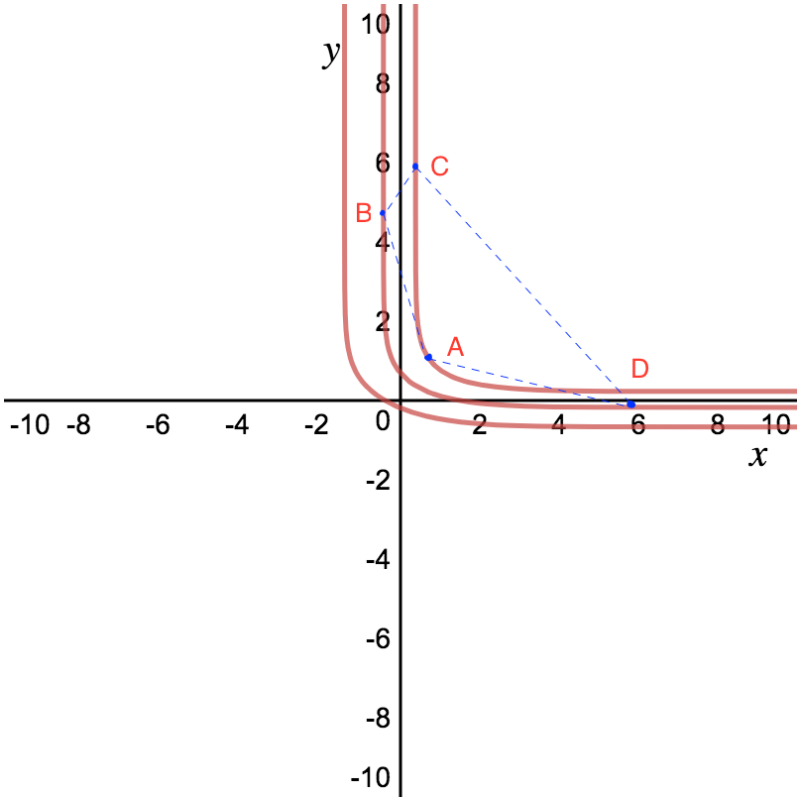

Figure 5: A contour plot of $f(x, y) = \rho(-x)\rho(-2y)$. Four arms are labeled by A, B, C, D such that the four points forms a convex quadrilateral and both pairs of opposite vertexes have the same function value(reward). One can always find such four points due to the non-convexity of f.

In this section, we construct an example when a Logistic Bandit suffers a linear regret under the examination's hypothesis. It is enough to consider a simple case with $d_C = d_E = 1$ and $|\mathcal{A}| = 4$. Figure 5 plots a simple but representative case when $\boldsymbol{\theta} = [1, 2]^T$. One can think of the x-axis as the context feature of result examination and y-axis as the context feature of relevance judgment. Denote the reward at arm a by R_a, and the corresponding coordinate by (x_a, y_a), where $a \in \{A, B, C, D\}$. Then the figure indicates that

$$R_A = R_C > R_B = R_D. \tag{32}$$

Note that for a Logistic Bandit, which does not consider the decomposition of implicit feedback thus combines x and y in a linear predictor, will not be able to learn the correct order between those arms, i.e., Equation 32. This can be easily verified through a necessary condition. In fact, assume there is $[a, b]^T$ such that

$$\begin{aligned} ax_A + by_A &= ax_C + by_C \\ ax_B + by_B &= ax_D + by_D \end{aligned} \tag{33}$$

which is a necessary condition for a Logistic Bandit to learn Equation 32 exactly. As long as $x_B < x_A < x_C < x_D$ and $y_A < y_B < y_D < y_C$ (the case we plot in the figure), the determinant of coefficient matrix is less than 0 so that the only solution is $a = b = 0$. In this case, Logistic Bandit fails to differentiate A, C from B, D, and would suffer a regret of $R_A - R_B$ with probability $\frac{1}{2}$ when ties break evenly. Without of loss of generality, we assume $a \neq 0$. If $a > 0, b \geq 0$, then $ax_D + by_D > ax_A + by_A$, thus a regret of $R_A - R_D$ is inevitable when the availble arm set is $\{A, D\}$, since logistic function is monotonically increasing with respect to its linear predictor. So is case when $a < 0, b \leq 0$ with the inequality $ax_B + by_B > ax_C + by_C$. The cases when $a > 0, b < 0$ with the inequality $ax_D + by_D > ax_C + by_D$, and when $a < 0, b > 0$ with the inequality $ax_B + by_B > ax_A + by_A$ follow the same logic. Thus, we have construct a case with an adversary environment when a linear regret is inevitable for a Logistic Bandit lacking in the decomposition of implicit feedback.

Appendix G Features used for the modeling of click and examination

We list the features we use for the modeling of examination and click in E-C Bandit in Table 1. Table 2 gives detailed description of these features. The main criteria of separating features between examination and relevance is that whether this feature is question-related or not.

Table 1: Feature usage in E-C Bandit

Feature	Used for examination?	Used for click(relevance)?
Playing speed	$\surd$	$\times$
Device	$\surd$	$\times$
Position	$\surd$	$\times$
Content Related	$\times$	$\surd$

Table 2: Feature description

Features	Description
Playing speed	Students are able to play the video at five different speeds. They are 1, 1.25, 1.5, 1.75 and 2 timing the default speed, respectively. The faster the playing speed, the less probable the examination.
Device	A student may watch a video using different devices, such as a PC, iPad or smart cellphones. Different device should have different probability of examination.
Position	The question may be displayed at the beginning of the video, or somewhere middle of the video. Different CTRs for the same question at different positions are observed in real data, which we assume is caused by examination.
Content Related	A question's semantic meaning and its similarity with the subtitle of a video. Distributed embedding or tf-idf score may be used to gain a numeric representation of these features.