Multi-level Feedback Web Links Selection Problem: Learning and Optimization

Kechao Cai[1], Kun Chen[2], Longbo Huang[2], John C.S. Lui[1]
[1]Department of Computer Science & Engineering, The Chinese University of Hong Kong
[2]Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University

***Abstract*—Selecting the right web links for a website is important because appropriate links not only can provide high attractiveness but can also increase the website's revenue. In this work, we first show that web links have an intrinsic *multi-level feedback structure*. For example, consider a 2-level feedback web link: the 1st level feedback provides the Click-Through Rate (CTR) and the 2nd level feedback provides the potential revenue, which collectively produce the compound 2-level revenue. We consider the context-free links selection problem of selecting links for a homepage so as to maximize the total compound 2-level revenue while keeping the total 1st level feedback above a preset threshold. We further generalize the problem to links with n $(n \geq 2)$-level feedback structure. The key challenge is that the links' multi-level feedback structures are unobservable unless the links are selected on the homepage. To our best knowledge, we are the first to model the links selection problem as a constrained multi-armed bandit problem and design an effective links selection algorithm by learning the links' multi-level structure with provable *sub-linear* regret and violation bounds. We uncover the multi-level feedback structures of web links in two real-world datasets. We also conduct extensive experiments on the datasets to compare our proposed LExp algorithm with two state-of-the-art context-free bandit algorithms and demonstrate that LExp algorithm is the most effective in links selection while satisfying the constraint.**

I. Introduction

Websites nowadays are offering many web links on their homepages to attract users. For example, news websites such as Flipboard, CNN, and BBC constantly update links to the news shown on their homepages to attract news readers. Online shopping websites such as Amazon and Taobao frequently refresh various items on their homepages to attract customers for more purchase.

Each link shown on a homepage is intrinsically associated with a *"multi-level feedback structure"* which provides valuable information on users' behaviors. Specifically, based on the user-click information, the website can estimate the probability (or Click-Through Rate (CTR)) that a user clicks that link, and we refer to this as the 1*st level feedback*. Moreover, by tracking the behaviors of users after clicking the link (e.g., whether users will purchase products associated with that link), the website can determine the revenue it can collect on that web page, we refer to this as the 2*nd level feedback*. The *compound* 2-*level feedback* is a function of the 1st level feedback and the 2nd level feedback. Naturally, the 1st level feedback measures the *attractiveness* of the link, while the 2nd level feedback measures the *potential revenue* of the link given that the link is clicked, and the compound 2-level feedback measures the *compound revenue* of the link. In summary, for a given homepage, its total attractiveness is the sum of CTRs of all links on that homepage, and its total compound revenue is the sum of the compound revenue of all links on that homepage. Both the total attractiveness and the total compound revenue of a homepage are important measures for investors to assess the value of a website [1].

Due to the limited screen size of mobile devices or the limited size of an eye-catching area on a web page, homepages usually can only contain a finite number of web links (e.g., Flipboard only shows 6 to 8 links for its users on its homepage frame without sliding the frame.). Moreover, contextual information (e.g., the users' preferences) is not always available due to visits from casual users, cold start [2] or cookie blocking [3]. Furthermore, a website with an unattractive homepage would be difficult to attract investments. In this case, it is important for website operators to consider the *context-free* web links selection problem: how to select a finite number of links from a large pool of web links to show on its homepage so to maximize the total compound revenue, while keeping the total attractiveness of the homepage above a preset threshold?

The threshold constraint on the attractiveness of the homepage makes the above links selection problem challenging. On the one hand, selecting those links with the highest CTRs ensures that the attractiveness of the homepage is above the threshold, but it does not necessarily guarantee the homepage will have high total compound revenue. On the other hand, selecting links with the highest compound revenue cannot guarantee that the total attractiveness of the homepage satisfies the threshold constraint. Further complicating the links selection problem is the multi-level feedback structures of web links, i.e., the CTRs (1st level feedback) and the potential revenues (2nd level feedback), are *unobservable* if the links are not selected into the homepage.

To tackle this challenging links selection problem, we formulate a stochastic constrained multi-armed bandit with multi-level rewards. Specifically, arms correspond to links in the pool, and each arm is associated with a 1st level reward corresponding to the 1st level feedback (the CTR) of a link, a 2nd level reward corresponding to the 2nd level feedback (the potential revenue) of the link, and a compound 2-level reward corresponding to the compound 2-level feedback of the same link. Our objective is to select a finite number of links on the homepage so as to maximize the cumulative compound 2-level rewards (or minimizing the regret) subject to a threshold constraint while learning/mining of links' multi-level feedback structures. To achieve this objective, we design a constrained bandit algorithm **LExp**, which is not only effective in links selection, but also has the attractive property that it achieves provable *sub-linear* regret and violation bounds.

Contributions: (i) We show that a web link is intrinsically associated with a multi-level feedback structure. (ii) To our best knowledge, we are the first to model the links selection problem with multi-level feedback structures as a stochastic constrained bandit problem (Sec. II). (iii) We design an bandit algorithm **LExp** that selects L arms from K arms ($L \leq K$) with provable *sub-linear* regret and violation bounds (Sec. III). (iv) We show that **LExp** is more effective in links selection than two state-of-the-art context-free bandit algorithms, CUCB [4] and EXP3.M [5], via extensive experiments on two real-world datasets (Sec. IV).

II. MODEL

In this section, we first introduce the context-free web links selection problem with a 2-level feedback structure. Then we show how to formulate it as a stochastic constrained bandit problem, and illustrate how it can model the links selection problem. Finally, we *generalize* the links selection problem to links selection problems with n-level feedback with $n \geq 2$.

A. Bandit Formulation (Constrained 2-level Feedback)

Consider a website structure with a homepage frame and a pool of K web pages, $\mathcal{W} = \{w_1, \dots, w_K\}$. Each web page $w_i \in \mathcal{W}$ is addressed by a URL link and so we have K links in total. The homepage frame can only accommodate up to $L \leq K$ links. When we select the link associated with web page $w_i, 1 \leq i \leq K$, and put it into the homepage frame, we can observe the following information when users browse the homepage:

1) $A_i \geq 0$, the probability that a user clicks the link to w_i, which is also referred to as the click-through rate (CTR);
2) $B_i \geq 0$, the potential revenue received from the user who clicks the link and then purchases products (or browses ads) on the web page w_i.

Therefore, for the link associated with web page w_i, the compound revenue is A_iB_i, $1 \leq i \leq K$. Our task is to select L links from the pool of K links for the homepage frame. The objective is to maximize the total compound revenue of the selected L links, subject to the constraint that the total CTR of these selected L links is greater than or equal to a preset threshold $h > 0$. Let $\mathcal{I} = \{i | w_i \in \mathcal{W}\}$ and $|\mathcal{I}| = L$ be the set of indices of any L links. Denote the feasible set of the above links selection problem as $\mathcal{S}$, which contains all possible subsets of indices of any L links such that satisfy the total CTR requirement h. Specifically, the optimal set of the L links for the described links selection problem is the solution to the following constrained knapsack problem,

$$\begin{aligned} &\arg\max_{\mathcal{I}\in\mathcal{S}} \sum\nolimits_{i\in\mathcal{I}} A_iB_i, \\ &\mathcal{S} = \big\{\mathcal{I} = \{i|w_i \in \mathcal{W}\} \big| |\mathcal{I}| = L, \sum\nolimits_{i\in\mathcal{I}} A_i \geq h\big\}. \end{aligned} \tag{1}$$

Problem (1) is known to be NP-hard [6]. To tackle this problem, we relax (1) to a probabilistic linear programming problem (2) as follows,

$$\begin{aligned} &\arg\max_{\boldsymbol{x}\in\mathcal{S}'} \sum\nolimits_{i=1}^{K} x_iA_iB_i, \\ &\mathcal{S}' = \big\{\boldsymbol{x} \in [0,1]^K \big| \sum\nolimits_{i=1}^{K} x_iA_i \geq h, \sum\nolimits_{i=1}^{K} x_i = L\big\}, \end{aligned} \tag{2}$$

where $\boldsymbol{x} = (x_1, \dots, x_i, \dots, x_K)$ and x_i represents the probability of selecting the web page $w_i, 1 \leq i \leq K$. Note that problem (2) is still non-trivial to solve because A_i and B_i are only *observable* if the web page w_i is *selected* to the homepage frame. If w_i is not selected, one *cannot* observe A_i or B_i.

To answer problem (2), we formulate the links selection problem as a stochastic constrained multi-armed bandit problem with 2-level rewards and design a constrained bandit algorithm. Formally, let $\mathcal{K} = \{1, \dots, K\}$ denote the set of arms, where each arm corresponds to a specific link to a web page in $\mathcal{W}$. Each arm $i \in \mathcal{K}$ is associated with two *unknown* random processes, $A_i(t)$ and $B_i(t)$, $t = 1, \dots, T$. Specifically, $A_i(t)$ characterizes the arm i's $1st$ *level reward* which corresponds to link i's 1st level feedback (CTR), and $B_i(t)$ characterizes arm i's $2nd$ *level reward* which corresponds to link i's 2nd level feedback (potential revenue) that can be collected from w_i. We assume that $A_i(t)$ are stationary and independent across i, and the probability distribution of $A_i(t)$ has a finite support. As for $B_i(t)$, they *are not necessarily stationary* due to the heterogeneity of users but are bounded across i. Without loss of generality, we normalize $A_i(t) \in [0,1]$ and $B_i(t) \in [0,1]$. We also assume that $A_i(t)$ is independent of $B_i(t)$ for $i \in \mathcal{K}$, $t \geq 1$. Note that this assumption is reasonable as we have observed and validated that different level feedbacks of links in the multi-level feedback structure do not have strong correlations in the real-world datasets (please refer to Sec. IV).

The stationary random process $A_i(t)$, is assumed to have *unknown* mean $a_i = \mathbb{E}[A_i(t)]$ for $1 \leq i \leq K$. Let $\boldsymbol{a} = (a_1, \dots, a_K)$.[1] Let $\boldsymbol{a}_t = (a_1^t, \dots, a_K^t)$ and $\boldsymbol{b}_t = (b_1^t, \dots, b_K^t)$ denote the realization vectors for the random processes $A_i(t)$ and $B_i(t)$, respectively for $1 \leq i \leq K$. Let $\boldsymbol{x}_t = (x_1^t, \dots, x_i^t, \dots, x_K^t)$ be the *probabilistic selection vector* of the K arms at time t, where $x_i^t \in [0,1]$ is the probability of selecting of the arm i at time t. The number of selected arms is L at each time t, i.e., $\mathbf{1}^\intercal \boldsymbol{x}_t = L$, where $\mathbf{1} = (1, \dots, 1)$ is the one vector. At time t, a set of $L \leq K$ arms $\mathcal{I}_t \in \mathcal{K}$ is selected via a dependent rounding procedure [7], which guarantees the probability that $i \in \mathcal{I}_t$ is x_i^t at time t (see Sec. III). For each arm $i \in \mathcal{I}_t$, the algorithm observes a $1st$ *level reward* a_i^t generated by $A_i(t)$ as well as a $2nd$ *level reward* b_i^t generated by $B_i(t)$, and receives a *compound* 2*-level* reward. Specifically, the compound 2-level reward, g_i^t, of an arm i at time t is generated by the random process $G_i(t) = A_i(t)B_i(t)$. Let $g_i^t = a_i^tb_i^t, 1 \leq i \leq K$ and $\boldsymbol{g}_t = (g_1^t, \dots, g_i^t, \dots, g_K^t)$. In addition, there is a preset threshold $h > 0$ such that the average of the sum of the 1st level rewards needs to be above

[1] All vectors defined in this paper are column vectors.

this threshold, i.e., $\boldsymbol{a}^\intercal \mathbb{E}[\boldsymbol{x}_t] \geq h$.[2] At time t, the expected total compound 2-level reward is $\mathbb{E}[\sum_t \boldsymbol{g}_t^\intercal \boldsymbol{x}_t]$ with the probabilistic selection vector $\boldsymbol{x}_t$, $t = 1, \ldots, T$.

Our objective is to design an algorithm to choose the selection vectors $\boldsymbol{x}_t$ for $t = 1, \ldots, T$ such that the *regret*, which is also referred to as loss compared with the oracle $\max_{\boldsymbol{a}^\intercal \boldsymbol{x} \geq h} \sum_{t=1}^T \boldsymbol{g}_t^\intercal \boldsymbol{x}$, is minimized. Specifically, the regret for an algorithm π is,

$$\text{Reg}_\pi(T) = \max_{\boldsymbol{a}^\intercal \boldsymbol{x} \geq h} \sum\nolimits_{t=1}^T \boldsymbol{g}_t^\intercal \boldsymbol{x} - \mathbb{E}\left[\sum\nolimits_{t=1}^T \boldsymbol{g}_t^\intercal \boldsymbol{x}_t^\pi\right], \tag{3}$$

where $\boldsymbol{x}_t^\pi$ is the probabilistic selection vector calculated by the algorithm π at time t. Note that $\boldsymbol{x}_t^\pi$ may violate the constraint initially especially when we have little information about the arms. To measure the overall violations of the constraint at time T, the *violation* of the algorithm π is defined as,

$$\text{Vio}_\pi(T) = \mathbb{E}\left[\sum\nolimits_{t=1}^T (h - \boldsymbol{a}^\intercal \boldsymbol{x}_t^\pi)\right]_+, \tag{4}$$

where $[x]_+ = \max(x, 0)$. Note that if the regret and violation of an algorithm are linear, the algorithm is *not learning*. A simple example of such algorithms is the uniform arm selection algorithm where any L arms are selected with equal probability. Such a random policy would result in both linear regret and linear violation as there is a constant loss compared to the optimal policy at each time t.

B. Generalization to n-level Feedback, where $n \geq 2$

We can further extend the constrained multi-armed bandit model with 2-level reward to the constrained multi-armed bandit model with n-level ($n \geq 2$) reward, and this allows us to model links selection problem with n-level feedback structure. Specifically, we can take each web page $w_i \in \mathcal{W}$ as a pseudo homepage frame. For the pseudo homepage frame, there is a pool of web pages $\mathcal{W}'$, $|\mathcal{W}| = K'$. Then we consider selecting a subset of L' links $\mathcal{I}'$, $|\mathcal{I}'| = L'$ (that each links to a web page in $\mathcal{W}'$) for the pseudo home page frame, with the constraint that the total CTR on the pseudo homepage frame is above the threshold h'. Formally, for each web page $w_i \in \mathcal{W}$, we consider the potential revenue of B_i in a much more precise way, i.e., $B_i = \sum_{j \in \mathcal{I}'} A_j' B_j'$, where A_j' is the CTR of the link associated with the web page $w_j' \in \mathcal{I}'$ and B_j' is the potential revenue collected from the web page $w_j' \in \mathcal{I}'$. As such, we extend the links selection problem with 2-level feedback (A_i and B_i where $i \in \mathcal{I}$) to a problem with 3-level feedback And similarly, we can further extend the problem to the problems with n-level feedback structure where $n \geq 2$.

III. Algorithm & Analysis

In this section, we first elaborate the design of our constrained bandit algorithm **LExp** (which stands for "$\binom{K}{L}$-Lagrangian Exponential weights") and present the algorithmic details. Then we provide both regret and violation analysis and show that our algorithm has the attractive property of being *sub-linear* in both regret and violation.

Algorithm 1 LExp (γ, δ)

```
Init: $\boldsymbol{\eta}^1 = \mathbf{1}, \lambda_1 = 0, h > 0, \beta = (1/L - \gamma/K)/(1-\gamma)$
for $t = 1, \ldots, T$ do
    $\mathcal{A}_t = \emptyset,\ \mathcal{I}_t = \emptyset,\ \alpha_t = 0.$
    if $\max_{i \in \mathcal{K}} \eta_i^t \geq \beta \sum_{i=1}^K \eta_i^t$ then
        Find $\alpha_t$ such that
            $\alpha_t / \left(\sum_{i=1, \eta_i^t \geq \alpha_t}^K \alpha_t + \sum_{i=1, \eta_i^t < \alpha_t}^K \eta_i^t\right) = \beta$
        $\mathcal{A}_t = \{i : \eta_i^t \geq \alpha_t\}$
    for $i = 1, \ldots, K$ do
        $\tilde{\eta}_i^t = \alpha_t$ if $i \in \mathcal{A}_t$; otherwise, $\tilde{\eta}_i^t = \eta_i^t$
    for $i = 1, \ldots, K$ do
        $\tilde{x}_i^t = L[(1-\gamma)\tilde{\eta}_i^t / \sum_{i=1}^K \tilde{\eta}_i^t + \gamma/K]$
    $\mathcal{I}_t = \text{DependentRounding}(L, \tilde{\boldsymbol{x}}_t)$
    for $i \in \mathcal{I}_t$ do receive $a_i^t$, and $b_i^t$
    for $i = 1, \ldots, K$ do
        $\hat{a}_i^t = a_i^t / \tilde{x}_i^t \mathbb{1}(i \in \mathcal{I}_t),\ \hat{g}_i^t = a_i^t b_i^t / \tilde{x}_i^t \mathbb{1}(i \in \mathcal{I}_t)$
    for $i = 1, \ldots, K$ do
        $\eta_i^{t+1} = \begin{cases} \eta_i^t & \text{if } i \in \mathcal{A}_t; \\ \eta_i^t \exp[\zeta(\hat{g}_i^t + \lambda_t \hat{a}_i^t)] & \text{if } i \notin \mathcal{A}_t \end{cases}$
    $\lambda_{t+1} = [(1 - \delta\zeta)\lambda_t - \zeta(\frac{\hat{\boldsymbol{a}}_t^\intercal \tilde{\boldsymbol{x}}_t}{1-\gamma} - h)]_+$
function DependentRounding($L, \boldsymbol{x}$)
    while exist $x_i \in (0,1)$ do
        Find $i, j, i \neq j$, such that $x_{i,j} \in (0,1)$
        $p = \min\{1 - x_i, x_j\},\ q = \min\{x_i, 1 - x_j\}$
        $(x_i, x_j) = \begin{cases} (x_i + p, x_j - p) \text{ with prob. } \frac{q}{p+q}; \\ (x_i - q, x_j + q) \text{ with prob. } \frac{p}{p+q}. \end{cases}$
    return $\mathcal{I} = \{i \,|\, x_i = 1, 1 \leq i \leq K\}$
```

A. Constrained Bandit Algorithm

The unique challenge for our algorithmic design is to balance between maximizing the compound multi-level rewards (or minimizing the regret) and at the same time, satisfying the threshold constraint. To address this challenge, we incorporate the theory of Lagrange method in constrained optimization into the design of **LExp**. We consider minimizing a modified regret function that includes the violation with an *adjustable* penalty coefficient that increases the regret when there is any non-zero violation. Specifically, **LExp** introduces a sub-linear bound for the Lagrange function of $\text{Reg}_\pi(T)$ and $\text{Vio}_\pi(T)$ in the following structure,

$$\text{Reg}_\pi(T) + \rho(T)\text{Vio}_\pi^2(T) \leq T^{1-\theta}, 0 < \theta \leq 1, \tag{5}$$

where $\rho(T)$ plays the role of a Lagrange multiplier. From (5), we can derive a bound for $\text{Reg}_\pi(T)$ and a bound for $\text{Vio}_\pi(T)$ as follows:

$$\text{Reg}_\pi(T) \leq O(T^{1-\theta}), \text{Vio}_\pi(T) \leq \sqrt{O(T^{1-\theta} + LT)/\rho(T)}, \tag{6}$$

where the bound for $\text{Vio}_\pi(T)$ in (6) is for the fact that $-\text{Reg}_\pi(T) \leq O(LT)$ for any algorithm π. Thus, with properly

[2] If $h = 0$, the problem is equivalent to the classic unconstrained multiple play multi-armed bandit problem (MP-MAB) [8].

chosen algorithm parameter $\rho(T)$, both the regret and violation can be bounded by sub-linear functions of T.

The details of **LExp** are shown in Algorithm 1. In particular, **LExp** maintains a weight vector $\boldsymbol{\eta}^t$ at time t, which is used to calculate the probabilistic selection vector $\tilde{\boldsymbol{x}}_t$ (line 3 to line 7). Specially, line 3 to line 6 ensure that the probabilities in $\tilde{\boldsymbol{x}}_t$ are less than or equal to 1. At line 8, we deploy the dependent rounding function (line 13 to line 17) to select L arms using the calculated $\tilde{\boldsymbol{x}}_t$. At line 9, the algorithm obtains the rewards a_i^t and b_i^t, and then gives *unbiased* estimates of $\hat{a}_i^t$ and $\hat{g}_i^t$ at line 10. Specifically, the 1st level reward $\hat{a}_i^t$, and the compound 2-level reward $\hat{g}_i^t$ are estimated by $a_i^t/\tilde{x}_i^t$, and $a_i^t b_i^t/\tilde{x}_i^t$, respectively, such that $\mathbb{E}[\hat{a}_i^t] = a_i^t$, and $\mathbb{E}[\hat{g}_i^t] = a_i^t b_i^t$. Finally, the weight vector $\boldsymbol{\eta}_t$ and the Lagrange multiplier λ_t are updated (line 11 and line 12) using previous estimations.

B. Regret and Violation Analysis

Theorem 1. *Let* $\zeta = \frac{\gamma\delta L}{(\delta+L)K}$, $\gamma = \Theta(T^{-\frac{1}{3}})$ *and* $\delta = \Theta(T^{-\frac{1}{3}})$ *that satisfy* $\delta \geq \frac{4(e-2)\gamma L}{1-\gamma} - L$. *By running the* **LExp** *algorithm* $\tilde{\pi}$, *we achieve sub-linear bounds for both the regret in* (3) *and violation in* (4) *as follows:*

$$\mathrm{Reg}_{\tilde{\pi}}(T) \leq O(LK\ln(K)T^{\frac{2}{3}}) \textit{ and } \mathrm{Vio}_{\tilde{\pi}}(T) \leq O(L^{\frac{1}{2}}K^{\frac{1}{2}}T^{\frac{5}{6}}).$$

Readers can refer to our technical report [9] for the proof.

IV. Experiments

In this section, we first examine the web links' multi-level feedback structures in two real-world datasets from the Kaggle Competitions, *Avito Context Ad Clicks* [10] and *Coupon Purchase Prediction* [11], referred to as "Ad-Clicks" and "Coupon-Purchase" in this paper. Then we conduct a comparative study by applying **LExp** and two state-of-the-art context-free bandit algorithms, CUCB [4] and Exp3.M [5], to show the effectiveness of **LExp** in links selection.

A. Multi-level Feedback Structure Discovery

The Ad-Clicks data is collected from users of the website Avito.ru where a user who is interested in an ad has to first click to view the ad before making a phone request for further inquiries. The data involves the logs of the visit stream and the phone request stream of $71,677,831$ ads. We first perform some data cleaning. For each ad in Ad-Clicks, we count the number of views of the ad and thereafter, count the number of phone requests that ad received. In particular, we filter out the ads that have an abnormally large number of views (greater than 2000), and the ads that receive few numbers of phone requests (smaller than 100). Finally, we obtain 225 ads from Ad-Clicks. For each of these 225 ads, we divide the number of phone requests by the number of views to get the Phone Request Rate. We normalize the numbers of views of each ad to the interval $[0, 1]$ using min-max scaling. The normalized number of views can be taken as the CTRs of the ads. As such, we find the multi-level feedback structure for each ad in the Ad-Clicks data: the CTR corresponds to the ad's 1*st level feedback*, the Phone Request Rate corresponds to the ad's 2*nd level feedback*, and the product of CTR and the Phone Request Rate corresponds to the *compound* 2*-level feedback*.

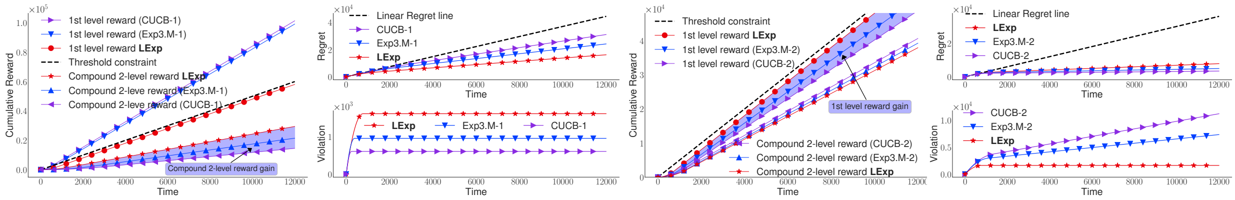
The Coupon-Purchase data is extracted from transaction logs of $32,628$ coupons on the site ponpare.jp, where users first browse a coupon and then decide whether to purchase the coupon or not. We extract 271 coupons from Coupon-Purchase. For each coupon, we divide its number of purchase by the number it was browsed, and take the ratio as the *Coupon Purchase Rate*. We then normalize all the browsed times to $[0, 1]$ using min-max scaling and refer to the normalized browsed times as to the CTRs of the coupons. Thus, for each coupon in Coupon-Purchase, the CTR corresponds to the coupon's 1*st level feedback*, the Coupon Purchase Rate corresponds to the coupon's 2*nd level feedback*, and the product of the CTR and the Coupon Purchase Rate is the *compound* 2*-level feedback*.

Next, we validate our previous claim in Sec. II that the 1st level feedback and the 2nd level feedback in the multi-level feedback structure do not have a strong correlation. In Ad-Clicks, we find that the CTRs and the Phone Request Rates are not strongly correlated with a correlation coefficient -0.47. Similarly, low correlation can also be found in Coupon-Purchase where the correlation coefficient is -0.27 only.

B. Comparative Study on Links Selection

We simulate the multi-level feedback structures in the real-world datasets and model the probabilistic ads/coupons selection process with time-variant rewards. In particular, for each of the 225 ads in Ad-Clicks, we treat its CTR/1st level feedback as a Bernoulli random variable with the mean CTR taking from Ad-Clicks, and we vary the Phone Request Rate/2nd level reward over time in a similar fashion as a sinusoidal wave (similar to [12]): the 2nd level reward starts from a random value drawn uniformly from 0 to the mean Phone Request Rate taking from Ad-Clicks; then in each time slot, it increases or decreases at rate $10/T$ until reaching the mean or 0. This time-variant rewards can model the seasonal fluctuations of the potential revenue of the ads. For each of the 271 coupons in Coupon-Purchase, we simulate its 2-level rewards in the same way.

In our performance comparison, the CUCB algorithm always selects the top-L arms with the highest UCB (upper confidence bound) indices without considering any constraint. For the CUCB algorithm, on the one hand, if we only want to maximize the total 1st level rewards, the L arms of the highest UCB indices $\hat{a}_i^t + \sqrt{3\ln t/(2N_i(t))}$ are selected at each time t (CUCB-1), where $N_i(t)$ is the number of times that the arm i has been selected by time t. On the other hand, if we only want to maximize the total compound 2-level rewards, the L arms of the highest UCB indices $\hat{g}_i^t + \sqrt{3\ln t/(2N_i(t))}$ will be selected (CUCB-2). For Exp3.M, only the 1st level reward estimation $\hat{\boldsymbol{a}}_t$ is considered when we only maximize the total 1st level rewards (Exp3.M-1), and only the compound 2-level reward estimation $\hat{\boldsymbol{g}}_t$ is considered when we only maximize the total compound 2-level rewards (Exp3.M-2). In our experiments, the *cumulative* 1st level reward and *the cumulative* compound



(a) Experiment 1: Cumulative rewards (b) Experiment 2: Regret/Violation (c) Experiment 3: Cumulative reward (d) Experiment 4: Regret/Violation

Fig. 1. Four groups of comparative experiments among **LExp**, EXP3.M and CUCB on Ad-Clicks. $K = 225$, $L = 20$, $h = 5$ and $T = 12000$ ($\gamma = 0.019$, $\delta = 0.021$). Note that both regrets and violations are calculated in an accumulative fashion.

2-level reward at time t are calculated using $\sum_{t'=1}^{t}\sum_{i\in\mathcal{I}_{t'}} a_i^{t'}$ and $\sum_{t'=1}^{t}\sum_{i\in\mathcal{I}_{t'}} g_i^{t'}$, respectively. The *regret* at time t is calculated using $\sum_{t'=1}^{t}\boldsymbol{g}_{t'}^{\intercal}\boldsymbol{x}^* - \sum_{t'=1}^{t}\sum_{i\in\mathcal{I}_{t'}} g_i^{t'}$ where $\boldsymbol{x}^*$ is the optimal probabilistic selection vector by solving $\max_{\boldsymbol{a}^{\intercal}\boldsymbol{x}\geq h}\sum_{t'=1}^{t}\boldsymbol{g}_{t'}^{\intercal}\boldsymbol{x}$ with $\boldsymbol{a}$ taking from the datasets. The *violation* at time t is calculated using $\sum_{t'=1}^{t}(h-\sum_{i\in\mathcal{I}_{t'}} a_i^{t'})_{+}$.

For Ad-Clicks, we run **LExp**, the EXP3.M variants: EXP3.M-1 and EXP3.M-2, and the CUCB variants: CUCB-1 and CUCB-2. The parameter settings are shown in Fig. 1.

- **Experiment 1** (considering total 1st level rewards only): In Fig. 1(a), we compare the cumulative rewards of **LExp**, EXP3.M-1 and CUCB-1. Specifically, the *cumulative* threshold constraint is a linear function of t, i.e., $h \cdot t$, shown by the black-dash line. One can observe that **LExp** satisfies the threshold constraint because its slope is equal to h after time $t = 841$, meaning that there is no violation for **LExp** after $t = 841$. On the contrary, CUCB-1 and EXP3.M-1 initially do not satisfy the threshold constraint and both exceed the threshold constraint. Furthermore, the cumulative compound 2-level rewards of CUCB-1 and EXP3.M-1 are both below that of **LExp** as the ads selected by CUCB-1 and EXP3.M-1 with high 1st level rewards do not necessarily result in a high 2nd level rewards. Thus, by selecting ads that satisfy the threshold constraint, even when the cumulative 1st level rewards is restricted by the threshold, **LExp** gives the highest total compound 2-level rewards with the compound 2-level rewards gain shown in the blue shaded area.
- **Experiment 2** (Regrets & Violations on Experiment 1): Fig. 1(b) shows the regrets and violations of CUCB-1, EXP3.M-1 and **LExp**. First, we draw the linear regret line which shows the largest regret for the ads selection problem where no ads are selected at each time, i.e., all the rewards are lost. The regret of CUCB-1 and the regret of EXP3.M-1 are very close to the linear regret line. They are also greater than the regret of **LExp**, which has a sub-linear property, and this further confirms the fact that cumulative compound 2-level rewards of CUCB-1 and EXP3.M-1 are both below that of **LExp** in Fig. 1(a). For the violations, all the three algorithms, **LExp**, EXP3.M-1 and CUCB-1 first show some increases and then remain constant. This is because they all are in the *exploration phase*: they first select ads with random 1st level rewards when the 1st level rewards are still unknown, and later select ads with high 1st level rewards that satisfy or exceed the threshold constraint. **LExp** is less aggressive than EXP3.M-1 and CUCB-1 which both select the ads with total 1st level rewards that far exceed the threshold. But as we can observe in Fig. 1(a), **LExp** performs much better than these algorithms. In summary, **LExp** takes longer to explore the optimal ads but after some trials, the violation *at each time* diminishes to zero. This can be observed from Fig. 1(b) since the violations remains unchanged after about 850 time slots.
- **Experiment 3** (considering the compound 2-level rewards only): Fig. 1(c) shows the cumulative rewards of **LExp**, EXP3.M-2 and CUCB-2. Specially, the cumulative compound 2-level rewards of CUCB-2 and EXP3.M-2 are both larger than that of **LExp** as they consider maximizing the total compound 2-level rewards only. However, the cumulative 1st level rewards of both EXP3.M-2 and CUCB-2 increase slower than the threshold constraint $h \cdot t$ as their slopes are less than h. This means that they violate the constraint and the gap between their cumulative 1st level rewards and the threshold constraint continues to grow as time goes by. For the cumulative 1st level rewards of **LExp**, the slope increases up to h and maintains at h and therefore the gap becomes a constant. In summary, **LExp** ensures that the threshold constraint is satisfied and produces the additional 1st level reward gain (blue-shaded area) compared with EXP3.M-2 and CUCB-2.
- **Experiment 4** (Regrets & Violations on Experiment 3): Fig. 1(d) shows the regrets and violations of **LExp**, EXP3.M-2 and CUCB-2. For the regrets, the regrets of EXP3.M-2, CUCB-2 and **LExp** are all *sub-linear* and below the linear regret line, as these three algorithms all aimed at minimizing the regret. Among them, **LExp** has a comparable regret but it also has an addition property, which is to satisfy the threshold constraint. As for the violation, the violations of both CUCB-2 and EXP3.M-2 end up linear as their cumulative 1st level rewards increase slower than the cumulative threshold constraint as shown in Fig. 1(c). This implies that both algorithms will never learn from the data to satisfy the constraint. In contrast, the violation of **LExp** increases and then eventually stays constant. This confirms that **LExp** aims to satisfy the constraint and it will not make mistake after some rounds of learning.

For Coupon-Purchase, we obtain similar experimental results and we can draw similar conclusions. Therefore, we omit the detailed descriptions for conciseness. In summary, our experimental results show that **LExp** is the only algorithm which balances the regret and violation in the ads/coupons selection problem.

V. Related Work

One common approach to the links selection problem is to perform A/B testing [13], which splits the traffic for different sets of links on two different web pages, and then evaluate their rewards. However, A/B testing does not have any loss/regret guarantee as it splits equal amounts of traffic to the links regardless of the links' rewards. That said, A/B testing is still widely used in commercial web systems. Our algorithm can be viewed as a complementary approach to A/B testing, e.g., our algorithm can select the set of links with the 1st level reward above a given threshold and facilitate a more efficient A/B testing for the links selection problem.

Another approach is to model the links selection problem as contextual bandit problems. [14] first formulated a contextual bandit problem aiming at selecting articles/links that maximize the total number of clicks based on the user-click feedback. Recently, [15] and [16] incorporated the collaborative filtering method into contextual bandit algorithms using users' profiles to recommend web links. However, contextual information is not always available due to cold start [2] or blocking of cookie tracking [3] and it neglects the multi-level feedback structures of the links and do not consider any constraint.

Our bandit formulation is related to the bandit models with multiple plays, where multiple arms are selected in each round. [5] presented the Exp3.M bandit algorithm that extends the single-played Exp3 algorithm [17] to multiple-played cases using exponential weights. [4] proposed an algorithm that selects multiple arms with the highest upper confidence bound (UCB) indices. [18] presented the multiple-play Thompson Sampling algorithm (MP-TS) for arms with binary rewards. [19] proposed a bandit-based ranking algorithm for ranking search queries. Our bandit model differs from these bandit models as we further consider the constraint on the total 1st level rewards in selecting the multiple arms.

Note that the constraint in our constrained bandit model is very different from that in bandit with budgets [20], [21] and bandit with knapsacks [22]. For these works, the optimal stopping time is considered since no arms can be selected/played if the budget/knapsacks constraints are violated. However, the constraint in our model does not pose such restrictions and the arm selection procedure can continue without stopping. Finally, our constrained bandit problem is related but different from the bandit model considered in [23] which tries to balance regret and violation. They only considered selecting a single arm without any multi-level rewards. While in our work, we consider how to select multiple arms and each arm is associated with multi-level rewards, making our model more challenging and applicable to the web links selection problem.

VI. Conclusion

In this paper, we reveal the intrinsic multi-level feedback structures of web links and formulate the web links selection problem. To our best knowledge, we are the first to model the links selection problem with multi-level feedback structures as a stochastic constrained bandit problem. We propose and design an effective links selection algorithm **LExp** with *provable sub-linear regret* and *violation* bounds. Furthermore, we carry out extensive experiments to compare **LExp** with the state-of-the-art context-free bandit algorithms and demonstrate that **LExp** is superior in selecting web links with constrained multi-level feedback by balancing both regret and violation.

Acknowledgment

The research of John C.S. Lui is supported in part by the GRF 14630815.

References

[1] R. Kohavi, A. Deng, R. Longbotham, and Y. Xu, "Seven rules of thumb for web site experimenters," in *Proceedings of SIGKDD*, 2014.

[2] M. Elahi, F. Ricci, and N. Rubens, "A survey of active learning in collaborative filtering recommender systems," *Computer Science Review*, vol. 20, pp. 29–50, 2016.

[3] W. Meng, B. Lee, X. Xing, and W. Lee, "Trackmeornot: Enabling flexible control on web tracking," in *Proceedings of WWW*, 2016.

[4] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework, results and applications," in *Proceedings of ICML*, 2013.

[5] T. Uchiya, A. Nakamura, and M. Kudo, "Algorithms for adversarial bandit problems with multiple plays," in *Proceedings of ACL'10*, 2010.

[6] B. Korte and R. Schrader, *On the existence of fast approximation schemes*, ser. Reprint series. Inst. für Ökonometrie u. Operations-Research, 1982.

[7] R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan, "Dependent rounding and its applications to approximation algorithms," *Journal of the ACM (JACM)*, vol. 53, no. 3, pp. 324–360, 2006.

[8] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, 1987.

[9] K. Cai, K. Chen, L. Huang, and J. C. Lui, "Multi-level feedback web links selection problem: Learning and optimization," *arXiv preprint arXiv:1709.02664*, 2017.

[10] Kaggle, "Avito context ad clicks," 2015, https://www.kaggle.com/c/avito-context-ad-clicks.

[11] ——, "Coupon purchase prediction," 2016, https://www.kaggle.com/c/coupon-purchase-prediction.

[12] O. Besbes, Y. Gur, and A. Zeevi, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Proceedings of NIPS*, 2014.

[13] A. Deng, J. Lu, and J. Litz, "Trustworthy analysis of online a/b tests: Pitfalls, challenges and solutions," in *Proceedings of WSDM*, 2017.

[14] L. Li, W. Chu, J. Langford, and R. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of WWW*, 2010.

[15] G. Bresler, D. Shah, and L. F. Voloch, "Collaborative filtering with low regret," in *Proceedings of ACM SIGMETRICS*, 2016.

[16] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *Proceedings of SIGIR*, 2016.

[17] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multiarmed bandit problem," *SIAM J. Comput.*, vol. 32, no. 1, Jan. 2003.

[18] J. Komiyama, J. Hondaand, and H. Nakagawa, "Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays," in *ICML*, 2015.

[19] A. Vorobev, D. Lefortier, G. Gusev, and P. Serdyukov, "Gathering additional feedback on search results by multi-armed bandits with respect to production ranking," in *Proceedings of WWW*, 2015.

[20] K. Deng, C. Bourke, S. Scott, J. Sunderman, and Y. Zheng, "Bandit-based algorithms for budgeted learning," in *Proceedings of ICDM*, 2007.

[21] Y. Xia, T. Qin, W. Ma, N. Yu, and T.-Y. Liu, "Budgeted multi-armed bandits with multiple plays," in *Proceedings of IJCAI*, 2016.

[22] S. Agrawal, N. R. Devanur, L. Li, and N. Rangarajan, "An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives," in *Proceedings of COLT*, 2016.

[23] M. Mahdavi, T. Yang, and R. Jin, "Online decision making under stochastic constraints," in *NIPS workshop on Discrete Optimization in Machine Learning*, 2012.