Incentivizing High Quality User Contributions: New Arm Generation in Bandit Learning

Yang Liu
yangl@seas.harvard.edu
Harvard University

Chien-Ju Ho
chienju.ho@wustl.edu
Washington University in St. Louis

Abstract

We study the problem of incentivizing high quality contributions in user generated content platforms, in which users arrive sequentially with unknown quality. We are interested in designing a content displaying strategy which decides which content should be chosen to show to users, with the goal of maximizing user experience (i.e., the likelihood of users liking the content). This goal naturally leads to a joint problem of incentivizing high quality contributions and learning the unknown content quality. To address the incentive issue, we consider a model in which users are strategic in deciding whether to contribute and are motivated by exposure, i.e., they aim to maximize the number of times their contributions are viewed. For the learning perspective, we model the content quality as the probability of obtaining positive feedback (e.g., *like* or *upvote*) from a random user. Naturally, the platform needs to resolve the classical trade-off between exploration (collecting feedback for all content) and exploitation (displaying the best content).

We formulate this problem as a multi-arm bandit problem, where the number of arms (i.e., contributions) is increasing over time and depends on the strategic choices of arriving users. We first show that applying standard bandit algorithms incentivizes a flood of low cost contributions, which in turn leads to linear regret. We then propose `Rand_UCB` which adds an additional layer of randomization on top of the UCB algorithm to address the issue of flooding contributions. We show that `Rand_UCB` helps eliminate the incentives for low quality contributions, provides incentives for high quality contributions (due to bounded number of explorations for the low quality ones), and achieves sub-linear regrets with respect to displaying the current best arms.

Introduction

User generated content (UGC) sites are ubiquitous on the Web – from online Q&A forums (such as Quora and stackoverflow), to reviewing sites (such as yelp and tripadvisor), to content-sharing sites (such as YouTube), and beyond. The success of UGC platforms relies heavily on user satisfaction. Ideally, a platform wants to optimize user experience by displaying the best possible content. Naturally, this objective leads to a joint incentive and learning problem. In particular, how does the platform incentivize users to contribute high quality content, and how does the platform learn the quality of the contributed content and identify the best one?

Let us first consider a simplified version of our problem and assume the platform has access to the true quality of the contributed content. In particular, we employ the model in which users are strategic and aim to maximize the exposure of their contributed content (i.e., the number of times their content are viewed in the future). [1] This simplified problem can then be reduced to a standard mechanism design problem: how to allocate the "rewards" (number of times we show the content to future users) to incentivize high quality contributions? This problem turns out to relatively well-studied in the literature (Ghosh and McAfee 2011; Ghosh and Hummel 2011; Ghosh and McAfee 2012; Ghosh and Hummel 2012).

However, in practice, the content quality is often not known in advance. Instead, the platform needs to rely on user feedback to estimate the content quality, defined as the probabilities of obtaining positive feedback (e.g., *like* or *upvote*) from a random user. This leads to a natural exploration-exploitation trade-off as in the bandit literature; the platform wants to learn the content quality (or the payoffs of *arms* in bandit settings) through exploration while optimizing user satisfaction through exploitation. However, our setting is more complicated than standard bandit settings in two aspects. First, the number of arms is not fixed and is increasing over time. Second, the quality distribution of arms is associated with the design of online learning algorithms, since users' decisions on whether to contribute is related to the number of times the learning algorithm will display their contributions.

In this paper, we explore this joint incentive and learning problem in user generated content platforms. In our setting, users arrive at the platform one at a time, providing feedback (i.e., votes) to the content displayed to them, and deciding whether to contribute new content. We assume users are unbiased in providing feedback and are strategic in deciding whether to contribute (aiming to maximize the exposure of

Copyright © 2018, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] This user incentive model is adopted in the literature (Ghosh and McAfee 2011; Ghosh and Hummel 2011). It captures the popular scenario that many online users contribute content to get attention. In addition, the number of content views could be translated into monetary rewards through, for example, embedding advertisements in the content.

their contributions). The content quality is unknown to the platform but can be learned through user feedback. The goal of the platform is to maximize the overall user satisfaction (i.e., the likelihood of showing users the content they like) by choosing a content-displaying strategy that simultaneously learns the quality of existing content and incentivizes high quality new contributions.

We first show that directly applying standard bandit algorithms (e.g., UCB1 (Auer, Cesa-Bianchi, and Fischer 2002)) generates bad incentives. The intuition is that, in standard bandit algorithms, we need to explore each arm enough number of times to estimate its quality with high confidence. This unavoidable exploration phase provides incentives for users to contribute, regardless of the quality of their content. This will result in a flood of contributed content, increase the number of arms, and further reduce incentives for contributing high quality content and degrade the performance of online learning. We call this phenomena the *curse of exploration*.

To address this issue, we proposed `Rand_UCB`, which randomly "drops" contributed arms with a dropping probability increasing over time. We show such a randomized UCB algorithm will de-incentivize low quality contributions; and further this property will provide incentives for users with high quality content to contribute, as the algorithm needs only explore a smaller number of arms. As one may imagine, this mechanism may drop good contributions as well. However, with carefully chosen dropping probabilities, the algorithm will obtain the near-optimal arm with high probability.

Related Work

This paper is closely related to the body of work on incentivizing high quality user contributions, in the context of online Q&A forums (Jain, Chen, and Parkes 2009), Games with A Purpose (Jain and Parkes 2013), crowdsourcing markets (Ghosh and McAfee 2012; Ho et al. 2015), and general UGC websites (Ghosh and McAfee 2011; Ghosh and Hummel 2011; Ghosh and Hummel 2012). However, most of the works along this line assume that the quality of user contributions are immediately observable, while this paper considers the learning perspective and the interaction between learning and incentives. Ghosh and Hummel (2013) has considered a similar setting as in this paper, in which strategic agents endogenously determines the quality of the arms. However, they consider a one-shot scenario, in which all agents need to determine the quality of the contributions simultaneously at the beginning of the learning process, without knowing other agents' actions. In this paper, we consider a more dynamic setting, in which agents sequentially make decisions based on what previously arrived agents have done in the platform.

The techniques we use are largely borrowed from the bandit literature (Lai and Robbins 1985; Auer, Cesa-Bianchi, and Fischer 2002; Bubeck, Cesa-Bianchi, and others 2012). However, in this paper, we need to address the issue of an increasing number of arms. Whether new arms will be contributed are determined by strategic agents. There has been some recent work discussing the incentive elements in learning. Both Gonen and Pavlov (2007) and Frazier et al. (2014) consider the setting that arms are pulled by selfish and myopic agents. In order to encourage exploration, the principal needs to provide incentives. A couple of later works are also on this line (Mansour, Slivkins, and Syrgkanis 2015; Mansour et al. 2016). Chakrabarti et al. (2009) considers the setting in which the arms can be replaced over time. In their setting, the arm replacements happen stochastically and the total number of arms is fixed. In contrast, in our setting, the number of arms is increasing and the arm contribution/generation is a choice by strategic agents.

Setting

We explore the *content displaying* problem in user generated content platforms. In our setting, users randomly arrive at the platform one at a time. When a user arrives, she first provides feedback (i.e., votes) to content displayed to her and then decide whether to make new contributions. We assume users are unbiased in providing feedback [2]. However, they are strategic in deciding whether to contribute, aiming to maximize the exposure of their own contributed content. The content quality is unknown to the platform in advance but can be learned through user feedback. The platform designer aims to choose a content displaying strategy to (1) learn the content quality and (2) incentivize high quality contributions, with the goal of maximizing the overall user satisfaction.

User and content models. Consider a discrete time setting with $t = 1, \ldots, T$. Let $A(t)$ be the set of existing content at time t. Initially, $A(t) = \emptyset$ when $t = 1$. Each content has an intrinsic quality $q \in [0, 1]$, which represents the probability of getting positive feedback/vote from a random user. At each time step t, a user randomly drawn from some unknown distribution arrives, and the platform chooses a set of content $a(t) \subseteq A(t)$ to display to the arriving user.

We abuse the notation and denote the user arriving at time t as user t. When user t arrives, she reviews the displayed content and provides votes to the content. For each content $i \in a(t)$, let q_i be the quality of content i, user t provides a vote $v_i(t) \in \{0, 1\}$ to the platform, with

$$v_i(t) \sim \texttt{Bernoulli}[q_i].$$

After voting, user t then decides whether to contribute. Each user t possesses a content i_t of quality $q_t \in [0, 1]$ randomly drawn from $F(q)$ and incurs a cost c_t to contribute. Suppose c_t has bounded support $c_t \in [\underline{c}, \overline{c}]$. Further we assume that $\underline{c} > 0$: this is to say that even contributing the least quality content will incur a non-zero cost. In practice, consider the fact that contributing a random answer to Q&A platforms requires effort (e.g., passing through several admin and verification steps). We denote $F(\cdot)$ as the CDF of the distribution of q. We assume that user t observes the true quality of her content and the quality of existing content on

[2] We consider the common scenario that the number of visits from a single user is small with respect to the total number of visits from all users. Therefore, the benefit a user can gain by strategically providing feedback is negligible.

the platform. She is also aware of the quality distribution $F(q)$ and the platform's content-displaying strategy.

Let $\omega_t(t')$ be the event that the contribution of user t is displayed at time t' (if user t chooses to contribute her content i_t):

$$\omega_t(t') = \{i_t \in a(t')\}, \forall t' = t+1, ..., T.$$

The utility for user t to contribute her content writes as

$$U_t := \mathbb{E}\left[\sum_{t'=t+1}^{T} \mathbb{1}(\omega_t(t'))\right] - c_t \ ,$$

where the expectation is over the randomness of the algorithm and the distribution of user quality. Since $U_t = 0$ if user t chooses not to contribute, we know that when $U_t > 0$, user t will choose to contribute. We would like to note that the linear sum is mainly to simplify the presentation. Our results stay valid as long as the users' utilities are monotone in this sum.

Objective of the platform. The goal of the platform is to choose a displaying strategy $\mathcal{A}$ to maximize the user experience, which can be formulated as the total number of positive feedback collected from users. Specifically we assess the performance of strategy $\mathcal{A}$ in the following three aspects:

1. We are interested in the time uniform regret in displaying the so-far best content. A smaller such regret will lead to better user experience over time. Denote by $K^*(t)$ the top-K arms with the highest quality at any time t (from $A(t)$), we define the following regret:

$$\mathsf{Regret}_{\mathcal{A}}(t) = \mathbb{E}_F\left[\sum_{t'=1}^{t} \sum_{i \in K^*(t')} q_i\right] - \mathbb{E}_{\mathcal{A},F}\left[\sum_{t'=1}^{t} \sum_{i \in a(t')} q_i\right]$$

   Our goal is to then achieve $\mathsf{Regret}_{\mathcal{A}}(t) = o(t)$ such that over time the time-average regret $\mathsf{Regret}_{\mathcal{A}}(t)/t \to 0$ [3].

2. We measure the maximum quality of collected arms at the end of mechanism: $\max_{i \in A(T)} q_i$. A better maximum quantity implies better incentives for new arm generation.
3. We also analyze the number of contributed low quality arms (which will be formally defined later). Controlling the number of low quality arms will help the system run more efficiently.

In the following discussion, we explore the design of displaying strategies. We focus on the natural design space that $\mathcal{A}$ can display at most K content at every time step, i.e., $|a(t)| \leq K$. Note that this displaying strategy also approximates users' position biases: users are a lot more likely to view the content in the top page (say it shows up to K content) than the content in subsequent pages.

A Warm-Up Setting: Known Quality

As a warm-up, we start with a simple setting in which the platform can observe the quality of the content contributed by arriving users. We demonstrate that the greedy algorithm (Top-K) incentivizes high quality contributions, while the random display algorithm incentivizes low cost contributions regardless of content quality.

To simplify the analysis, we consider the regime when T is large. In particular, we assume $T \to \infty$. We also assume there is no tie in user quality, i.e., $q_t \neq q_{t'}$ for all $t \neq t'$. Note that these assumptions are just used to simplify the presentations.

Let us first consider the random display algorithm, i.e., the platform randomly chooses K content from $A(t)$ to display at time t. It is easy to show that this algorithm incentivizes low-cost content, regardless of the content' true quality. [4]

Lemma 1. *If the platforms runs the random display algorithm, it is a dominant strategy (which leads to highest expected utility, regardless of other users' actions) for user t to contribute if and only if $|A(t)| < k(c_t)$, where $k(\cdot)$ is a monotonically decreasing function.*

Next we consider a simple greedy algorithm, i.e., Top-K algorithm, which ranks the quality of content in $A(t)$ and chooses the top K to display at time t. We can show that the Top-K algorithm incentivizes high quality contributions, i.e., user t will only contribute if her content will be ranked top K in the next time step and if her content quality is higher than some threshold (to make sure her contribution stays in top K for a long enough period of time).

Lemma 2. *Assume the platform runs the* Top-K *algorithm. Let j be the rank of q_t in $A(t+1)$ if user t contributes. In the symmetric equilibrium (as in standard Bayesian Nash Equilibrium), user t will contribute iff $j \leq K$ and $F_j(q_t) \geq c_t$ for some function F_j monotonically increasing in q_t.*

While these results are intuitive and may not seem surprising, they provide intuitions on the analysis of exploration-exploration-type algorithms in the following discussion. For example, when the platform explores to obtain feedback, it is essentially running random display algorithm. We would want to carefully limit the amount of explorations as it will lead to a flood of content regardless of quality.

A Bandit Approach

In practice, full information would be too ideal to assume. The platform often needs to collect information to learn the quality of each content. Below we first show that this additional learning phase creates bad incentives and motivates a flood of contributions, which in turn makes it infeasible to achieves sub-linear regrets. We then propose a simple, yet novel algorithm `Rand_UCB` to address the problem.

The curse of exploration

Since the quality of the contributed content is not known, the platform needs to show each contributed content to arriving

[3] As an alternative and perhaps a stronger notion, we can define the regret with respect to the best arms among all arms that could have been contributed, instead of the ones that have been contributed. But note that, since we will show later that under our mechanism, the quality of the best contributed arm approaches 1 when T goes large, our regret notion is sensible in characterizing the algorithm performance.

[4] The omitted proofs are included in the supplementary material available on the authors' websites.

users some number of times to learn its quality. At the same time, the platform also wants to only show the best content to users to maximize their satisfaction. This creates a tension between exploration and exploitation as in the classical multi-armed bandit learning. To resolve this exploration-exploitation trade-off, running standard bandit algorithms (for example, the well-celebrated UCB1 algorithm (Auer, Cesa-Bianchi, and Fischer 2002)) seems to be a very natural solution. However, we show that directly applying standard bandit algorithms introduces bad incentives and fails to achieve sub-linear regrets.

Informally, consider a user arriving at time t, where $t < \mathsf{const} \cdot T$ for any $\mathsf{const} < 1$.If she decides to contribute, a bandit-based display algorithm needs to display her content a high number of times (in the order of $\Omega(\log T)$) in order to achieve sub-linear regrets (Lai and Robbins 1985). This huge amount of unavoidable exploration creates bad incentives. In particular, for each user t, the benefit of explorations (in the order of $\Omega(\log T)$) she can obtain will outweigh the cost of contribution c_t when T goes large, regardless of content quality. This implies that applying standard bandit algorithms will create a flood of contributions and lead to unbounded regrets. We call this phenomenon the curse of exploration and formally summarize this negative result below.

Lemma 3. *Let $N_T(\mathcal{A})$ be the total number of contributed arms when the platform runs the content-display algorithm $\mathcal{A}$ for T rounds. Also, let $\mathcal{A}_{\mathcal{B}}$ be the content-display algorithm that chooses which content to display based on a bandit algorithm $\mathcal{B}$. When T is large enough, there does not exist a standard stochastic bandit algorithm $\mathcal{B}$ (that achieves $\Theta(\log T)$ regret with finite number of arms) such that $N_T(\mathcal{A}_{\mathcal{B}}) = o(T/\log T)$.*

Proof. We prove by contradiction. Suppose there exists such a bandit algorithm. Since the total number of arms is in the order of $o(T/\log T)$ by the contradicting assumption, first follow standard argument of bandit, we know the number of times a sub-optimal arm will be selected is at most in the order of $O(\log T)$, and the total number of sub-optimal arm selection is bounded by $o(T)$.

Then for any user arriving early, say before $\mathsf{const} \cdot T$, for any $\mathsf{const} < 1$, he would reason that his arm will be competing only with the optimal ones for the rest of $T' = (1 - \mathsf{const})T - o(T)$ steps. Apply the standard lower bound argument to the k steps, the arm will be explored at least $\Omega(\log T') = \Omega(\log T)$ times, which is larger than the cost of contribution when T is large enough. Therefore, when T is large enough, every user arriving before $\mathsf{const} \cdot T$ for any $\mathsf{const} < 1$ will choose to contribute, and the total number of arms will not be bounded as $o(T/\log T)$. This leads to the contradiction and finishes the proof. □

The above simple, yet striking result points out the caveat of directly running bandit algorithms to learn content quality in user-generated content platforms. Specifically, the number of arms that will be incentivized will approach the order of $\Omega(T/\log T)$, which makes it impossible to achieve sub-linear regrets. Note that when the number of explorations is large, the display algorithm based on bandit algorithms will look similar to a randomized display strategy. Thus this negative result is also hinted by Lemma 1.

Goal: We would like to propose an online algorithm that not only minimizes the regret in selecting the best arm, but also is able to incentivize high quality arms. More formally, recall that q_t is the quality of content by user t and $A(t)$ is the set of existing arms at time t. Let k_t be the rank of q_t in $A(t)$ if user t decided to contribute. If $k_t \leq K$, we call such an arm arriving at time t a high quality arm (since the arm will be among the top K arms in the next round), and we would like to incentivize its contribution. If $k_t > K$, we name such an arm as a low quality arm (the arm is not among the top K arms in the next round), and we would like to de-incentivize its contribution.

Proposed algorithm: `Rand_UCB`

Intuitively, the curse of explorations occurs because a bandit algorithm needs to explore most arms (in particular, any arm arriving at time $t < \mathsf{const} \cdot T$ for any $\mathsf{const} < 1$) a large number of times. This creates a bad incentive. It is therefore natural to ask, can we reduce the explorations for some of the arms without sacrificing the performance of the learning algorithm too much?

To achieve this goal, we propose `Rand_UCB`, which adds an additional layer of randomization on top of the UCB1 algorithm. This additional layer of randomization serves as a device for us to *tune* the amount of explorations at each time step. We show that, with the appropriate choice of the tuning parameters, we can incentivize better arms and achieve sub-linear regrets. [5]

`Rand_UCB` runs in two phases in each time step. In the first phase, the algorithm selects a subset of content $a(t)$ from existing ones according to the UCB1 algorithm (Auer, Cesa-Bianchi, and Fischer 2002). The process is described as follows.

- Let $n_i(t)$ be the number of times arm $i \in A(t)$ has been selected till time t. Let $v_i(n)$ be the n-th feedback (vote) user i has received, where $n = 1, \cdots, n_i(t)$.
- Select the top K arms from $A(t)$ to add to $a(t)$ according to the following index rule, with random tie-breaking

$$I_i(t) = \frac{\sum_{n=1}^{n_i(t)} v_i(n)}{n_i(t)} + d\sqrt{\frac{g(t)}{n_i(t)}}. \qquad (1)$$

Both d and $g(t)$ are configurable parameters. In standard UCB1, d is set to be 1, and $g(t) := 2\log t$.

The second phase is where our algorithm differs from standard bandit algorithms. In the second phase, we add an additional layer of randomization to handle newly contributed arms. In particular, whenever a new arm is contributed at time t, our algorithm flips a coin to decide whether to include the new arm. The newly contributed arm will only be added to the set of arms $A(t)$ with probability p_t and will be dropped [6] with probability $1 - p_t$.

[5] This additional layer of randomization is not the only possible device to tune the amount of explorations. As discussed later in the paper, we can combine existing machine learning tools as a device

Algorithm 1: Rand_UCB

```
Input: $\{p_t : t = 1, \dots, T\}$
for $t = 1, \cdots, T$ do
    select arms to display according to UCB1.
    if a new arm is contributed then
        add the new arm in $A(t+1)$ with probability $p_t$
    end if
end for
```

How should we choose p_t? As discussed previously, applying UCB1 will lead to linear regrets due to the large number of unavoidable explorations when T is large. In Rand_UCB, we propose to decrease p_t over time, i.e., to gradually decrease the chance of adding a newly contributed arm. The intuition is that we want to obtain good arms early with high probability (therefore we start with larger p_t when t is small) while not providing too much incentive for all arms (we decrease p_t when t increases). In particular, we show that when $p_t = \min\{1, M/t\}$ for some constant $M > 0$, Rand_UCB has good incentive properties and achieves sub-linear regrets.

Incentive properties of Rand_UCB

We first analyze the incentive properties of Rand_UCB. Define $S_{[t:T]} := \sum_{t'=t+1}^{T} p_{t'}$. Intuitively, $S_{[t:T]}$ upper bounds the number of arms added to the platform after time t. Denote the action of user arriving at time t as $\textsf{act}_t \in \{\textsf{contribute}, \textsf{don't contribute}\}$ and $\mathcal{H}_t$ as the set of historical statistics. We define dominant strategy as our solution concept for each incoming user t:

Definition 4. For each user arriving at time t, action a is called a dominant strategy if for all $a' \neq a$,

$$\mathbb{E}[U_t|a, \mathcal{H}_t, \{q_i, i \in A(t)\}] > \mathbb{E}[U_t|a', \mathcal{H}_t, \{q_i, i \in A(t)\}].$$

Below we show that Rand_UCB can incentivize high quality arms while discouraging low quality arms. Recall that k_t is the rank of the arm t (i.e., the arm possessed by user t) within the existing arms if it is contributed.

Theorem 5. *Assume the platform runs Rand_UCB with $p_t = \min\{1, M/t\}$ for some configurable constant M. Let $d = 1$ and $g(t) = \log t$ in the UCB1 algorithm. When T is large enough, for any user who arrives at time $t < \textsf{const} \cdot T$ with any $\textsf{const} < 1$, we can characterize whether user t will contribute based on the following conditions:*

- *If user t ($t = o(T)$) has a high quality arm (i.e., $k_t \leq K$), it is a dominant strategy for her to contribute if*
$$S_{[t:T]} \cdot (1 - F(q_t)) < K - k_t + 1.$$
- *If user t has a low quality arm (i.e., $k_t > K$), whether she will contribute depends on her arrival time. In particular, there exists a $f_t(T) = \Theta(\log T)$, where the constants in $f_t(T)$ depends on the realization of the top K contributions, such that*
  - *if $t \leq f_t(T)$, it is a dominant strategy to contribute.*
  - *if $t > f_t(T)$, it is a dominant strategy to not contribute.*

Proof. (Sketch) - We first prove that for a high quality arm s.t. $k_t \leq K$, when $S_{[t:T]} \cdot (1 - F(q_t)) < K - k_t + 1$, there is a positive probability $\kappa > 0$ that arm t will stay in the top-K set until T. Next we bound (i) selection of t, if contributed. (ii) number of total contributed arms.

The number of contributed arms can be bounded as follows with high probability

$$N_{\textsf{Contr.}}(t) := |A(t)| + O(S_{[t:T]}) = O(\log T + S_{[1:T]}).$$

Then we can bound $n_t(T)$ as follows, using standard three-way UCB1 proof (Auer, Cesa-Bianchi, and Fischer 2002):

$$\mathbb{E}\Big[\sum_{t'=t+1}^{T} \mathbb{1}(\omega_t(t'))\Big] \geq T - t - N_{\textsf{Contr.}}(t) \cdot \frac{8\log(T-t)}{(q_t - q_K)^2} - \text{const.},$$

where q_K denotes the K-th largest quality in the current $A(t)$. When set $p_t = M/t$ and $t = o(T)$, the above bound is in the order of $O(T - t - \log^2 T)$. Then agent's utility becomes in the order of $(T-t) \cdot \kappa \cdot p_t - O(\log^2 T \cdot p_t) - c_t$. Suppose $t = O(T^\theta)$, $0 < \theta < 1$. Then the above quantity is lower bounded by $O(\frac{T - T^\theta - \log^2 T}{T^\theta}) - c_t > 0$, when T is large enough.

For a low quality arm, when $t = \Omega(\log T)$, his utility if contribute can be upper bounded by $U_t \leq O(\log T) \cdot p_t - c_t \to 0 - c_t < 0$. Therefore there is no incentive to contribute. On the other hand when $t = o(\log T)$, we will know that $U_t \geq \Omega(\log T) \cdot p_t - c_t > 0$, when T is large enough. So users will contribute. Since a higher quality leads to higher expected utility, we establish the existence of such a threshold. □

There is a tension in selecting p_t: selecting a higher p_t will provide incentives for higher number of contributions, which leads to the curse of exploration; but setting a low p_t will miss out good arms. Our choice of p_t avoids over-explorations while guaranteeing we obtain good arms with high probability.

Note that there exists a gray region of users that we couldn't characterize their equilibrium strategy: agents who have good arms but the quality of their arms do not satisfy the condition we specified in Theorem 5. Intuitively, these agents' qualities are not high enough to ensure that it will stay in the top-K set in the long run. This may look like a concern. However, we show that as long as there are enough high quality contributed arms, the regret can still be bounded whether those users contribute or not.

Is Rand_UCB practical?

In Rand_UCB, we propose to randomly drop new contributions with a probability decreasing over time. While this seems to be an impractical strategy (we might not want to tell users their contributions may not be viewed at all), in practice, we can implement a soft version of Rand_UCB:

to help reduce the explorations and obtain good learning guarantee.

[6] We say a content is dropped if it's not added to the active exploring set of arms.

each arm is guaranteed to be explored a constant number of times before random dropping. As long as the guaranteed exploration is small, we can obtain the same incentive property and achieves sub-linear regret as stated in the previous section. In practice, we can even utilize the information from the guaranteed exploration and drop the arms that receive bad feedback (instead of random dropping).

The design intuition of Rand_UCB is to reduce the amount of explorations for later arms. Randomly dropping new arms is one of the strategies to achieve this. We discuss another approach that combines existing machine learning tools later in the paper. The intuition is that, in addition to user feedback, we might utilize the content features (e.g., the length of the contribution, the reputation of the contributor, etc) to help learn the content quality and reduce the amount of explorations.

Performance Analysis of Rand_UCB

In this section, we present a set of results characterizing the performance of Rand_UCB. In particular, we show that (1) Rand_UCB achieves sub-linear regrets, (2) the best arm collected by Rand_UCB approaches the best possible arm when T is large, and (3) the total number of low quality arms collected by Rand_UCB is bounded in the order of $O(\log T)$.

Regret analysis. We first state the lemma towards characterizing $\textsf{Regret}_{\mathcal{A}}(t)$.

Lemma 6. *At any time t, we have*

$$\textit{Regret}_{\mathcal{A}}(t) \leq 16\sqrt{Mt}\log t + O(\sqrt{t}).$$

Compared to standard stochastic bandit regret, we have an additional $\sqrt{t}$ term. This is mainly due to the fact that the arm quality is drawn from a continuous space – so we cannot differentiate two ϵ-close arms with only $O(\log t)$ number of samples. We are confident that when the quality levels are discrete, we will be able to bring the regret order back to poly-log.

Quality of the best contributed arm. We now bound the quality of the best contributed arm. Bounding this quality will hint on how well the algorithm can do in incentivizing high quality arm. We prove the following lemma showing that when T goes large, the highest quality will approach 1 (the highest quality).

Lemma 7. *When T goes large, we have* $\max_{i \in A(T)} q_i \rightarrow 1$ *w.h.p* $\geq 1 - \Theta(1/T)$.

Proof. (Sketch) - Denote by $c_k^*(t)$ the threshold for contributing a new arm at time t, when the new arm's quality rank is $k \leq K$, for a sub-linear time $t = o(T)$. That is $c_k^*(t)$ is the smallest q that satisfies that $S_{[t:T]} \cdot (1 - F(q)) < K - k + 1$. When $k = 1$ we have that

$$c_1^*(t) = \begin{cases} F^{-1}(1 - \frac{K}{S_{[t:T]}}), \text{ if } 1 - \frac{K}{S_{[t:T]}} > 0 \\ 0, \text{ o.w.} \end{cases}$$

We will show that at time t, if $\max_{i \in A(t)} q_i < c_1^*(t)$, there will be with high probability that an arm that has quality higher than $c_1^*(t)$ will be contributed in the future. Also we can find a $t = o(T)$ such that $\frac{K}{S_{[t:T]}} \gg K$ – this is due to the selection of p_t, i.e., setting $p_t = M/t$, we have $0 < S_{[t:T]} \leq O(\log T \cdot \cdot T^{-\theta})$, for some $0 < \theta < 1$. Therefore there will be such a time $c_1^*(t) \rightarrow 1$. □

Number of contributed low quality arms. We bound the number of low quality arms contributed in Rand_UCB.

Lemma 8. *The number of contributed low quality arms is bounded at the order of* $O(\log T)$.

This can be established straightforwardly from Theorem 5 that after $\Theta(\log T)$ number of rounds, low quality arms will have no incentive to contribute, due to the diminishing probability of the contribution being added to the pool of arms. Though trivially true, this result is of great practical value: de-incentivizing low quality contributions not only will reduce system's load for running and maintaining the algorithm, but also provides fundamental incentives for good arm contribution.

Simulation

In this section, we provide simulation results to demonstrate the intuitions of the design of Rand_UCB. Rand_UCB has two advantages over the standard UCB algorithm. First, it collects a good amount of content in the early stages ($p_t = \min\{1, M/t\}$) and gradually decreases the probability of adding newly contributed content into exploration phase. This allows the platform to obtain a good enough content early with high probability, while not sacrificing on keeping exploring new content. Second, as shown in Theorem 5, Rand_UCB incentivizes high quality contributions. This naturally improves the algorithm performance, since the arms are better. Below we use simulations to demonstrate the effects of these two components.

Decaying p_t over time improves performance. We first examine the effects of different p_t choices. We assume users always contribute and compare the results of decaying $p_t = \min\{1, M/t\}$ and constant $p_t = \{1, 0.1, 0.01, 0.001\}$. We set $K = 1$, $T = 10,000$, and $M = 10$. We also assume the quality distribution F is an uniform distribution in $[0, 1]$. We run each algorithm 100 times and plot the mean performance in Figure 1. The result shows that setting $p_t = \min\{1, M/t\}$ outperforms every other choices. Note that in the figure, the y-axis is the average utility till time t.

Good incentives help. We next examine the effects of good incentives. We assume each arriving user decide whether to contribute based on the characterization in Theorem 5. We compare the results of running Rand_UCB and UCB on strategic users. For comparison, we also plot the results of running Rand_UCB on always-contributing users. As we can see from the results, as shown in Figure 2, providing good incentives significantly improves the performance.

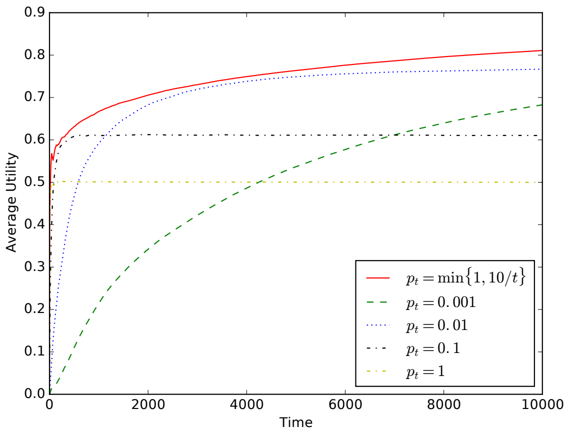

Figure 1: Gradually decaying p_t performs well comparing to fixed p_t.

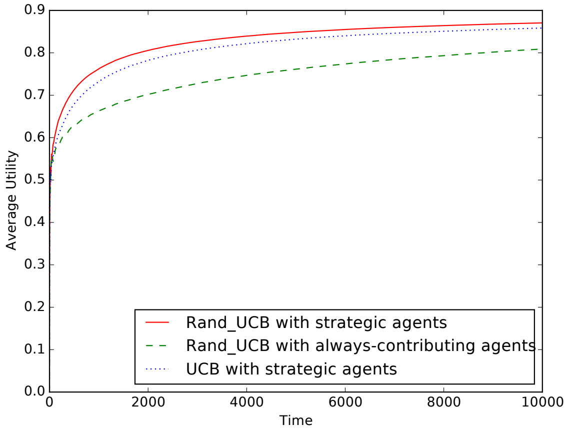

Figure 2: Providing good incentives improves the performance.

Discussions

We hope our work will open the discussion of designing incentivize compatible sequential learning methods to collect high quality contributions for online platforms. Below we first discuss an alternative approach to random dropping and then outline a few interesting extensions of our framework.

An alternative to random dropping

In Rand_UCB, we randomly drop newly contributed arms and use this as a device to control the amount of explorations. In this section we discuss an alternative approach to get around of the random dropping, via leveraging additional information and machine learning (ML) techniques.

Our main idea is inspired by the following observation. User contributions often arrive with features that signal content quality, e.g., content length, contributor's reputation, etc. Can we leverage machine learning techniques to predict the quality of newly contributed content from its feature vectors, without displaying it extensively? In other words, can we use machine learning tools as a device to help reduce the amount of explorations in our problem?

Note that, if there exist a ML algorithm that can perfectly predict content quality, our setting reduces to the full information setting we present earlier on. The best strategy is to only display top-K contribution. However, if the accuracy of the ML algorithm is not perfect and is upper bounded by a constant (not a function of time horizon T), we still suffer from the "curse of exploration". The intuition is, for each arriving arm, the ML algorithm essentially provides some number of "free explorations" (based on the ML prediction). However, to safely drop a low-quality arm, the number of explorations we need stays in the order of $\Omega(\log T)$.

In this section, we discuss an interesting case when there exists a ML algorithm whose error decreases in T and approaches 0 when T goes to infinity. For example, if each arm comes in with a feature vector, via collecting users' votes, we can train the ML to predict arms' quality. If the error rate gets to 0 as T gets large, we might use this to replace the random dropping mechanic. However, there are many other challenges, e.g., the training data comes from users' strategic choices and is not i.i.d. drawn, to design such an algorithm.

We discuss a simple linear model to demonstrate this idea (for details please refer to supplementary material): Suppose the contributed content at time t comes with a feature vector $x_t \in \mathbb{R}^D$. Consider the following linear model $q_t = \theta^\top x_t$, where $\theta \in \mathbb{R}^D$ is the unknown parameter. Therefore we know as soon as we can learn the θ correctly, we will be able to safely predict the quality of a newly arrived content. To estimate θ, we can collect a set of x_t along with its estimated qualities $\tilde{q}_t$ (through displays) at certain randomly selected time points, and perform linear regression.

Future directions

We outline a few interesting and promising future directions.

Effort sensitive model Effort sensitive models have been thoroughly studied in (Ghosh and McAfee 2011; Ghosh and McAfee 2012; Witkowski et al. 2013; Ho et al. 2015) for modeling the quality of user-generated content. These works consider the case that the content quality q_t is endogenously decided by an effort variable e_t. Agents' strategic decisions will not only be deciding whether to contribute but also be deciding which effort level to choose before contribution. While we think our proposed solution framework can be extended to this effort sensitive case, the analysis will be much more complicated; as now when agents reason about their utilities, they also need to reason about the effort exertion actions from all future agents. This challenge is also noted by Liu and Chen (2016).

A Dueling bandit approach There is also an interesting interleave between the incentive design and interface design problems. In our current set of results, when the mechanism designer displays a content, each content will receive a feedback based on its true quality. We can imagine another way (i.e., a different interface) of collecting quality information would be to ask each user to select her preferred content within two of them (or within a set of content). From the learning perspective, this falls into the scope of the newly arising study of dueling bandit (Yue et al. 2012). From the incentive perspective, how should we choose the set of content to display to users at each time step, while ensuring users with high quality arms are incentivized to contribute? It would be interesting to study how different incentives can be provided in different interfaces.

Conclusion

We propose a bandit algorithm `Rand_UCB` for solving the problem of incentivizing high quality contributions from sequentially arriving users with unknown quality. The algorithm builds on the classical UCB1 algorithm, with an additional layer of "random dropping" to tune the amount of explorations over time. We show that `Rand_UCB` helps eliminate the incentives for low quality contributions, provides incentives for high quality contributions (due to bounded number of explorations for the low quality ones), and achieves sub-linear regrets. We also offer discussions on possible extensions, including replacing random dropping with existing machine learning tools for reducing the amount of explorations. We hope this work will open up the discussion of designing incentive compatible sequential learning methods to collect high quality contributions for online platforms.

Acknowledgments We thank the anonymous reviewers for their helpful comments. This research was partially supported by NSF grant CCF-1512964 and CCF-1301976, ONR grant N00014-15-1-2335, and Harvard Data Science Initiative.

References

[Auer, Cesa-Bianchi, and Fischer 2002] Auer, P.; Cesa-Bianchi, N.; and Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47(2-3):235–256.

[Bubeck, Cesa-Bianchi, and others 2012] Bubeck, S.; Cesa-Bianchi, N.; et al. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5(1):1–122.

[Chakrabarti et al. 2009] Chakrabarti, D.; Kumar, R.; Radlinski, F.; and Upfal, E. 2009. Mortal multi-armed bandits. In *Advances in Neural Information Processing Systems 21*.

[Frazier et al. 2014] Frazier, P.; Kempe, D.; Kleinberg, J.; and Kleinberg, R. 2014. Incentivizing exploration. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*.

[Ghosh and Hummel 2011] Ghosh, A., and Hummel, P. 2011. A game-theoretic analysis of rank-order mechanisms for user-generated content. In *Proceedings of the 12th ACM Conference on Electronic Commerce*.

[Ghosh and Hummel 2012] Ghosh, A., and Hummel, P. 2012. Implementing optimal outcomes in social computing: A game-theoretic approach. In *Proceedings of the 21st International Conference on World Wide Web*.

[Ghosh and Hummel 2013] Ghosh, A., and Hummel, P. 2013. Learning and incentives in user-generated content: Multi-armed bandits with endogenous arms. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*.

[Ghosh and McAfee 2011] Ghosh, A., and McAfee, P. 2011. Incentivizing high-quality user-generated content. In *Proceedings of the 20th International Conference on World Wide Web*.

[Ghosh and McAfee 2012] Ghosh, A., and McAfee, P. 2012. Crowdsourcing with endogenous entry. In *Proceedings of the 21st International Conference on World Wide Web*.

[Gonen and Pavlov 2007] Gonen, R., and Pavlov, E. 2007. An incentive-compatible multi-armed bandit mechanism. In *Proceedings of the twenty-sixth annual ACM symposium on Principles of distributed computing*, 362–363. ACM.

[Ho et al. 2015] Ho, C.-J.; Slivkins, A.; Suri, S.; and Vaughan, J. W. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*.

[Jain and Parkes 2013] Jain, S., and Parkes, D. C. 2013. A game-theoretic analysis of the esp game. *ACM Transactions on Economics and Computation*.

[Jain, Chen, and Parkes 2009] Jain, S.; Chen, Y.; and Parkes, D. C. 2009. Designing incentives for online question and answer forums. In *Proceedings of the 10th ACM Conference on Electronic Commerce*.

[Lai and Robbins 1985] Lai, T. L., and Robbins, H. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

[Liu and Chen 2016] Liu, Y., and Chen, Y. 2016. A bandit framework for strategic regression. In *Advances in Neural Information Processing Systems*, 1813–1821.

[Mansour et al. 2016] Mansour, Y.; Slivkins, A.; Syrgkanis, V.; and Wu, Z. S. 2016. Bayesian exploration: Incentivizing exploration in bayesian games. *arXiv preprint arXiv:1602.07570*.

[Mansour, Slivkins, and Syrgkanis 2015] Mansour, Y.; Slivkins, A.; and Syrgkanis, V. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 565–582. ACM.

[Witkowski et al. 2013] Witkowski, J.; Bachrach, Y.; Key, P.; and Parkes, D. C. 2013. Dwelling on the Negative: Incentivizing Effort in Peer Prediction. In *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing (HCOMP'13)*, 190–197.

[Yue et al. 2012] Yue, Y.; Broder, J.; Kleinberg, R.; and Joachims, T. 2012. The k-armed dueling bandits problem. *Journal of Computer and System Sciences* 78(5):1538–1556.

Supplementary materials

Proof of Lemma 1

Proof. If user t contributes, her utility can be written as $U_t = \mathbb{E}[\sum_{t'=t+1}^{\infty} \frac{1}{|A(t')|}] - c_t$. The expectation is over the randomness of the qualities of future arriving users. It is easy to show that if user t chooses to contribute when $|A(t)| = k$, user t will also choose to contribute when $|A(t)| < k$ (simply by comparing each term $\mathbb{E}[1/|A(t')|]$ for $t' > t$). Similarly, if user t chooses not to contribute when $|A(t)| = k$, she won't choose to contribute when $|A(t)| > k$. Therefore, there exists a threshold $k(\cdot)$ such that user t will contribute if only if $|A(t)| < k(c_t)$. □

Proof of Lemma 2

Proof. Let $U_{t,j}(q_t, c_t)$ be the payoff of user t with quality rank j if she contributes and $V_{t,j}(q_t)$ be the value she receives from her contribution being displayed. We can write $U_{t,j}(q_t, c_t) = V_{t,j}(q_t) - c_t$. For simplicity, below we use $U_{t,j}$ and $V_{t,j}$ to denote $U_{t,j}(q_t, c_t)$ and $V_{t,j}(q_t)$.

If $j > K$, user t will not contribute since her contributions won't be displayed at all. So we need only consider the case $j \leq K$.

Let's first consider the case $j = K$. By definition, we have $V_{t,K} = \mathbb{E}\sum_{t'=t+1}^{\infty} \mathbb{1}(\omega_t(t'))$. Define $F_k = \Pr(q < q_t; c \geq c_K)$, represents the probability that a random user having quality $q < q_t$ and cost $c \geq c_K$, where c_K is a constant represents the minimum value for user to contribute if their rank is K. We can write that

$$V_{t,K} = \sum_{t'=1}^{\infty} F_k^{t'} = \frac{F_k}{1 - F_k}.$$

Now consider any $0 < j < K$, we have

$$V_{t,j} = F_j(1 + V_{t,j}) + (1 - F_j)(1 + V_{t,j+1})) = 1 + F_j V_{t,j} + (1 - F_j)V_{t,j+1}$$
$$\Rightarrow V_{t,j} = \frac{1 + (1 - F_j)V_{t,j+1}}{1 - F_j} = \frac{1}{1 - F_j} + V_{t,j+1} = \sum_{i=j}^{K-1} \frac{1}{1 - F_i} + \frac{F_k}{1 - F_k}$$

User t will choose to contribute if and only if $j \leq K$ and $U_{t,j} = V_{t,j} - c_t \geq 0$. Observe from the definition that F_i is monotonically increasing in q_t, so we know that $V_{t,j}$ is increasing in q_t as well. Observe the definition of $V_{t,j}$, we can write $V_{t,j}(q_t) = F_j(q_t)$. Therefore, user t chooses to contribute if and only if $F_j(q_t) \geq c_t$. □

Proof of Theorem 5

Proof. Consider step t, and shorthand its quality as q. Agent t observes that $k_t = k$, that is the quality of his arm is higher than the k-th highest one. Denote the following random variable:

$$Z_q(t) := \mathbb{1}(q_t > q) \tag{2}$$

Then $\mathbb{E}[Z_q(t)] = p_t(1 - F(q))$. The first condition that we will need is

$$\sum_{t'=t+1}^{T} \mathbb{E}[Z_q(t')] = \sum_{t'=t+1}^{T} p_{t'}(1 - F(q)) = S_{[t:T]} \cdot (1 - F(q)) < K - k + 1. \tag{3}$$

The above condition is saying that the expected number of better arms that arrive in the future in less than $K - k + 1$. This condition is easy to understand that if the expected better arms to be arriving is larger than $K - k + 1$, agent t may drop out of the top K set. Denote by

$$S_{[t:T]}(1 - F(q)) := (1 - \epsilon)(K - k + 1), 0 < \epsilon < 1. \tag{4}$$

For instance, we can set ϵ as $\epsilon = \frac{1}{2}$ and we enforce our first condition that

$$S_{[t:T]}(1 - F(q)) = \frac{K - k + 1}{2}$$

Using Bernstein's inequality we have

$$\Pr[\sum_{t'=t+1}^{T} Z_q(t) - \sum_{t'=t+1}^{T} \mathbb{E}[Z_q(t)] > \epsilon(K - k + 1)]$$
$$\leq \exp\left(-\frac{0.5 \cdot (\epsilon(K - k + 1))^2}{\sum_{t'=t+1}^{T} \mathrm{Var}(Z_q(t)) + \frac{1}{3}\max_{t'=t+1}^{T} |Z_q(t;\omega) - \mathbb{E}[Z_q(t)]| \epsilon(K - k + 1)}\right) \tag{5}$$

Since $\mathrm{Var}(Z_q(t)) \le \mathbb{E}[Z_q^2(t)] \le p_t(1 - F(q))$ we know

$$\sum_{t'=t+1}^{T} \mathrm{Var}(Z_q(t)) \le S_{[t:T]}(1 - F(q)) = (1 - \epsilon)(K - k) \tag{6}$$

and $|Z_q(t;\omega) - \mathbb{E}[Z_q(t)]| \le 1$, we know

$$\text{LHS of Eqn. (5)} \le \exp\left(-\frac{0.5 \cdot (\epsilon(K - k + 1))^2}{(1 - \epsilon)(K - k) + \frac{1}{3}\epsilon(K - k + 1)}\right) = \exp\left(-\frac{0.5 \cdot \epsilon^2(K - k + 1)}{(1 - \epsilon) + \frac{1}{3}\epsilon}\right) \tag{7}$$

Denote this event as

$$\mathcal{W}_t := \{\sum_{t'=t+1}^{T} Z_q(t) - \sum_{t'=t+1}^{T} \mathbb{E}[Z_q(t)] > \epsilon(K - k + 1)\}$$

Then under $\overline{\mathcal{W}_t}$, with probability

$$\Pr[\overline{\mathcal{W}_t}] \ge 1 - \exp\left(-\frac{0.5 \cdot \epsilon^2(K - k + 1)}{(1 - \epsilon) + \frac{1}{3}\epsilon}\right) \tag{8}$$

q will stay in top K quantile from time t to T. Particularly when ϵ is set to be 0.5, we have the probability become $1 - e^{-\frac{3}{16}(K - k_t + 1)}$. The rest to prove is the following two aspects:

- Bound on selection of t, if contributed.
- Bound on the number of contributed arms.

We first prove that the number of contributed arms is bounded. At time t, there are already $|A(t)|$ arms. Suppose from $t' = t$ to T, there is an arm arriving with quality $q_{t'} < q_K$, that is the quality is lower than the K-th highest. Then the expected number of selection of t' can be fairly straightforwardly bounded as follows:

$$\mathbb{E}[n_{t'}(T)] \le \frac{8}{(q_K - q_{t'})^2} \log(T - t') + \text{const.} \tag{9}$$

This follows by standard three-way argument for proving UCB1's regret ((Auer, Cesa-Bianchi, and Fischer 2002), details omitted) in that when

$$n_{t'}(T) \ge \frac{8}{(q_K - q_{t'})^2} \log(T - t'),$$

the probability of selecting t' over top-K options is bounded. Therefore the expected utility of such arms of contributing is at most $O(\frac{\log(T - t')}{t}) - c_{t'}$. When $t = \Omega(\log T)$, and $c_{t'} > 0$, there is no incentive for such low quality arms to contribute as $O(\frac{\log(T - t')}{t}) \to 0$. Then the total number of arm up to time T is bounded by

$$|A(t)| + \sum_{t'=t+1}^{T} Z_{q_K}(t')$$

Again $\mathbb{E}[\sum_{t'=t+1}^{T} Z_{q_K}(t')] = S_{[t:T]}(1 - F(q_K))$. Using Chernoff bound we know

$$\Pr\left[\sum_{t'=t+1}^{T} Z_{q_K}(t') \ge (1 + \delta)S_{[t:T]}(1 - F(q_K))\right] \le e^{-\frac{\delta S_{[t:T]}(1 - F(q_K))}{3}} \tag{10}$$

Take δ as $\delta := \frac{3 \log T}{S_{[t:T]}(1 - F(q_K))}$ we know above probability is bounded by $1/T$. So w.h.p.,

$$|A(t)| + \sum_{t'=t+1}^{T} Z_{q_K}(t') \le |A(t)| + S_{[t:T]} + 3 \log T$$

Then follow standard three-way UCB proof (Eqn. (9)) we know

$$\mathbb{E}[n_t(T)] \ge T - t - (|A(t)| + S_{[t:T]} + 3 \log T) \frac{8 \log(T - t')}{(q - q_K)^2} - \text{const.} \tag{11}$$

Next we will prove that with probability at least $1 - 2/t^2$ that

$$|a(t)| + S_{[t:T]} = O(\log T + S_{[1:T]}).$$

To prove this, consider $t = \Omega(\log T)$: since $\mathbb{E}[|A(t)|] \leq S_{[1:t]}$, using Chernoff bound we know that w.h.p. that $|A(t)| \leq O(S_{[1:t]})$. When set $p_t = C/t$, the above bound becomes $O(\log T + S_{[1:T]}) = O(\log T)$. Then agent's utility becomes

$$(T-t)(1 - e^{-\frac{3}{16}(K - k_t + 1)}) \cdot p_t - O(\log^2 T \cdot p_t) - c_t.$$

When $t = o(T)$, suppose $t = O(T^\theta), 0 < \theta < 1$. Then the above quantity becomes at the order of

$$\Theta(\frac{T - T^\theta - \log^2 T}{T^\theta}) - c_t > 0$$

On the other hand, when $t = o(\log T)$, for a low quality arm, it is guaranteed that $\mathbb{E}[n_t(T)] \geq \Omega(\log T)$. Therefore his utility writes as follows by contributing: $p_t \Omega(\log T) - c_t > 0$, when T is large. So contributing will be a better action. □

Proof for Lemma 6

Proof. First of all, the expected total number of arms up to any time t can be bounded as follows: $\sum_{t'=1}^{t} p_{t'} \leq \int_{t'=t}^{t} C/t' dt' = C \log t$. Denote event that there is a new arm being added at time t as new_t. Use Chernoff bound, we know that

$$\Pr\Big[\sum_{t'=1}^{t} \mathsf{new}_t \geq 8 \sum_{t'=1}^{t} p_{t'}\Big] \leq e^{-\sum_{t'=1}^{t} p_{t'}} = O(1/t^2).$$

The last equality is due to the fact that $\sum_{t'=1}^{t} = C \log t + O(1)$. Then with high probability $1 - O(1/t^2)$ that the number of arms is at most $8C \log t$. Supposing that the best arm is $q_{t'}, t' \leq t$. Define ϵ-close arm to $q_{t'}$ as any arm such that $|q - q_{t'}| \leq \epsilon$. Consider the following four cases:

(1) If $t - t' < \sqrt{t}$, that is the best arm is a recently added arm. Regret incurred in this phase is at most $\sqrt{t}$.

(2) When $t - t' > \sqrt{t}$. We can prove the following that for any arm k that is contributed before $t - \sqrt{t}$, and is not ϵ close (denoting the gap as $\epsilon_k > \epsilon$), the number of selection is bounded as

$$\mathbb{E}[n_k(t)] \leq \frac{8 \log t}{\epsilon_k^2} + \text{const.}$$

This follows from standard UCB1 three way argument with noting that

$$\Pr\Big[|I_k(t) - q_k| \geq \sqrt{\frac{2\log(t - t_k)}{n_k(t)}}\Big] \leq 2e^{-2 \cdot 4 \frac{2\log(t-t_k)}{n_k(t)} \cdot n_k(t)} \leq 2/t^2$$

where t_k denotes the arrival time of arm k. The first inequality is due to Chernoff bound, and the last inequality is due to the fact that $t - t_k \geq \sqrt{t}$. When set $2\sqrt{\frac{2\log(t-t_k)}{n_k(t)}} = \epsilon_k$, we bound $\mathbb{E}[n_k(t)]$. Further the regret is bounded by

$$\frac{8 \log t}{\epsilon_k} + \text{const.} \leq \frac{8 \log t}{\epsilon} + \text{const.}$$

(3) For the sub-optimal arm k that is contributed between $[t - \sqrt{t}, t]$, the regret can be similarly argued as the first case (1), which is on the order of $\sqrt{t}$.

(4) For ϵ-close arms, the regret is at most ϵt.

Adding up, the regret is bounded as

$$\sqrt{t} + \epsilon t + 8C \log t \cdot (\frac{8 \log t}{\epsilon} + \text{const.}) = \epsilon t + \frac{64 C \log^2 t}{\epsilon} + O(\sqrt{t}) \tag{12}$$

Set $\epsilon := 8\sqrt{C} t^{-1/2} \cdot \log t$ we have

$$\epsilon t + \frac{64 \log^2 t}{\epsilon} = 16\sqrt{C}\sqrt{t} \cdot \log t$$

□

Proof for Lemma 7

Proof. Denote by $c_k^*(t)$ the threshold for contributing a new arm at time t, when the new arm's quality rank is $k \leq K$, for a sub-linear time $t = o(T)$. That is $c_k^*(t)$ is the smallest q that satisfies that $S_{[t:T]} \cdot (1 - F(q)) < K - k + 1$. When $k = 1$ we have that

$$c_1^*(t) = \begin{cases} F^{-1}(1 - \frac{K}{S_{[t:T]}}), \text{ if } 1 - \frac{K}{S_{[t:T]}} > 0 \\ 0, \text{ o.w.} \end{cases}$$

Again consider a time t that is sub-linear in T, denoting as $t = O(T^\theta)$ for some $0 < \theta < 1$. Consider the following two cases. When $\max_{i \in a(t)} q_i \geq c_1^*(t)$, we have proved a lower bound (i.e., $c_1^*(t)$). When $\max_{i \in A(t)} q_i < c_1^*(t)$, we know the following two facts:

- For $o(T) = t' > t$, if $\max_{i \in A(t')} q_i < c_1^*(t)$, and $q_{t'} \geq c_1^*(t)$, and arm t' is selected to add, then agent t' is willing to contribute.
- $\mathbb{E}[\sum_{t'=t}^{o(T)} Z_{c_1^*(t)}(t')] > 1$.

The first proof can be easily adapted from our proof for Theorem 5 that if an agent has incentive to contributed an arm with $q \geq c_1^*(t)$ at time t (according to our definition of $c_1^*(t)$), he will also have incentives to do so at $t' > t$. As we can show that with $t' > t$, the probability of the agent staying in the top-K set will increase.

The second argument can also be adapted from the proof for Theorem 5. First we know from the proof of Theorem 5 that

$$\Pr\left[\sum_{t'=t}^{T} Z_{c_1^*(t)}(t') > K - 2\right] \to 1,$$

at the order of $1 - O(1/T)$. This is because for the threshold case, is it allowed to have $K - 1$ better arms to arrive in the future. Next we need to prove that $\mathbb{E}[\sum_{t'=t}^{o(T)} Z_{c_1^*(t)}(t')]/\mathbb{E}[\sum_{t'=t}^{T} Z_{c_1^*(t)}(t')]$ is non-negligible. Note the following fact that

$$\sum_{t'=T^\theta}^{T} 1/t' = T^{-\theta} \sum_{t=1}^{T^{1-\theta}} \leq T^{-\theta} \int_{t=1}^{T^{1-\theta}} 1/tdt \leq O((1-\theta)\log T \cdot T^{-\theta})$$

and $\sum_{t'=T^\theta}^{T^\theta + \log T} 1/t' = O(\log T \cdot T^{-\theta})$. Then we assert $\mathbb{E}[\sum_{t'=t}^{o(T)} Z_{c_1^*(t)}(t')]$ takes a non-negligible fraction of $\mathbb{E}[\sum_{t'=t}^{T} Z_{c_1^*(t)}(t')]$ if $t' = T^\theta$, and we appropriately select such a θ. Further from $t' = T^\theta$ to another sub-linear time, there will be at least one better arm with $q_{t'} \geq c_1^*(t)$ will be added. Also we can find a $t = o(T)$ such that $\frac{K}{S_{[t:T]}} \gg K$ – this is due to the selection of p_t, i.e., setting $p_t = C/t$, we have $0 < S_{[t:T]} \leq O(\log T \cdot T^{-\theta})$. Therefore there will be such a time $c_1^*(t) \to 1$. □

A linear model for predicting content quality using ML

Suppose the contributed content at time t comes with a feature vector $x_t \in \mathbb{R}^d$. For simplicity of analysis, we assume that x_t is drawn from a unit ball such that $||x_t|| \leq 1$. Consider the following linear model $q_t = \theta^\top x_t$, where $\theta \in \mathbb{R}^d$ is the unknown parameter. Therefore we know as soon as we can learn the θ correctly, we will be able to safely predict the quality of a newly arrived content. To learn such θ we use the following regression procedure 2.

Algorithm 2: A linear regression procedure for predicting qualities of newly arrived contents

- Randomly sample a certain number of time points (to ensure i.i.d. samples). Denote such time points up to time t as $E(t)$, such that $|E(t)| = C_E \cdot \log t$, for some constant $C_E > 0$.
- Promise a constant number $T_E > 0$ of displays for contents contributed at time points in $E(t)$, to ensure those contents will be contributed.
- Use the information collected from the constant number of displayed to estimate noisy quality $\tilde{q}_n,\ n \in E(t)$. Denote this collected set of training data as $\{(x_n, \tilde{q}_n)\}_{n \in E(t)}$.
- Linear regress over above set of data: $\tilde{\theta}_t := \text{argmax}_{\theta:||\theta|| \leq 1} \sum_{n \in E(t)} (\theta^\top x_n - \tilde{q}_n)^2$.
- Use the estimated $\tilde{\theta}_t$ to predict the quality of newly arrived contents $\tilde{q}_t := \tilde{\theta}_t^\top x_t$.
- Pre-set a threshold $\epsilon > 0$. When $t = \Omega(T^\theta),\ 0 < \theta < 1$, only add newly contributed contents to the active exploration set if $\tilde{q}_t \geq q_K + 3\epsilon$.

Without repeating all the analysis, we provide intuition on why the above system works: first it is very easy to verify that

$$\tilde{q}_t = \theta^\top x_t + \epsilon_t$$

where ϵ_t is zero mean and has bounded support. Then applying standard linear regression results (Liu and Chen 2016) to obtain that with high probability $1 - O(1/t)$ that $||\tilde{\theta}_t - \theta|| \leq O(\frac{d}{\log t})$. When t and T are large enough such that $O(\frac{d}{\log t}) \leq \epsilon$, with high probability, only the content with quality $q_t \geq q_K + \epsilon$ would be willing to contribute. Further since there will be at most $1/\epsilon$ arms being contributed before q_K reaches the maximum 1 (bounded number of arms), there will be incentives for the high quality arms to contribute.

It is worth to note that, when t is large that the error in estimating θ is sufficiently small, for arriving agents with low quality contents, they know with high probability that their contributed contents will be accurately predicted and thus not displayed. So they will have no incentives to contribute. Furthermore, we would like to note that as time t goes large that $||\tilde{\theta}_t - \theta||$ becomes negligible, this setting reduces to the full information setting we detailed earlier on.