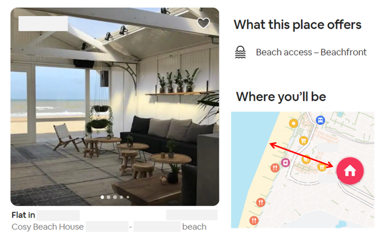
Bandits Meet Mechanism Design to Combat Clickbait in Online Recommendation

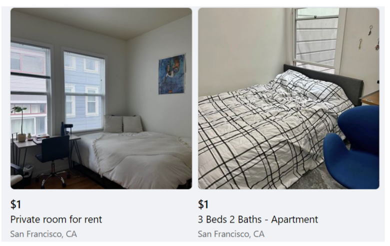
Thomas Kleine Buening [1] Aadirupa Saha [2] Christos Dimitrakakis [3] Haifeng Xu [4]

[1]University of Oslo [2]Apple ML Research [3]University of Neuchatel [4]University of Chicago

Motivation

- Recommendation platforms serve as intermediates between **vendors** and **users** so as to recommend **items** from the former to the latter.
- Vendor chosen **item descriptions** are an essential aspect of the problem that is often ignored. These invite vendors to strategically exaggerate their true value in the description to increase their **click-rate**.



We combine **bandit learning** with **mechanism design** to incentivize desirable vendor strategies under uncertainty while minimizing regret.

The Strategic Click-Bandit Problem

Every (strategic) arm $i \in [K]$ is associated with

1) a **reward distribution** with mean μ_i, and
2) a **click-rate** s_i which is **strategically** chosen by arm i.

Interaction Protocol.

```
Learner commits to an algorithm M, which is shared with all arms
Arms choose strategies (s_1, ..., s_K) ∈ [0,1]^K, unknown to the learner
for t = 1, ..., T do
    Algorithm M selects arm i_t ∈ [K]
    Arm i_t is clicked with probability s_{i_t}, i.e., c_{t,i_t} ~ Bern(s_{i_t})
    if i_t was clicked (c_{t,i_t} = 1) then
        Arm i_t receives utility 1 from the click
        M observes noisy post-click reward r_{t,i_t} ∈ [0,1] with mean μ_{i_t}
```

We must learn both the **strategically chosen click-rates** $s_1, \dots, s_K$ and the **post-click rewards** $\mu_1, \dots, \mu_K$ through **repeated interaction**.

Learner's Utility. The learner's utility of selecting an arm i with click-rate s_i and post-click value μ_i is denoted by $u(s_i, \mu_i)$. As an example, consider

$$u(s, \mu) = s\mu - \lambda(s - \mu)^2.$$

However, we derive our results for a **broad class of utility functions** $u : [0,1] \times [0,1] \to \mathbb{R}$ satisfying basic regularity assumptions (Lipschitzness ...).

Arms' Utility. Each arm i aims to maximize its **total number of clicks** given algorithm M and strategies (s_i, s_{-i}):

$$v_i(M, s_i, s_{-i}) := \mathbb{E}_M\left[\sum_{t=1}^{T} \mathbb{I}\{i_t = i\}\, c_{t,i}\right].$$

We can also express this as $v_i(M, s_i, s_{-i}) = \mathbb{E}_M[n_T(i)] \cdot s_i$ where $n_T(i)$ is the number of times i has been selected by the algorithm.

Nash Equilibrium and Strategic Regret

We study the situation where the arms respond to the learner's algorithm by playing a (possibly mixed) **Nash Equilibrium** of the general-sum game induced by the utilities $v_1, \dots, v_K$.

Note that the arms' strategy space is given by $[0,1]$. Let $\boldsymbol{\sigma} \in \Sigma^K$ denote a mixed strategy profile, i.e., a distribution over pure strategies $\boldsymbol{s} \in [0,1]^K$. Let

$$\mathrm{NE}(M) := \{\boldsymbol{\sigma} \in \Sigma^K : \boldsymbol{\sigma} \text{ is NE under } M\}$$

denote the set of all Nash equilibria for the K arms under algorithm M.

The **Strategic Regret** of M under a **pure-strategy NE** $\boldsymbol{s} \in \mathrm{NE}(M)$ is:

$$R_T(M, \boldsymbol{s}) := \mathbb{E}\left[\sum_{t=1}^{T} u(s^*, \mu^*) - u(s_{i_t}, \mu_{i_t})\right].$$

Accordingly, for a **mixed-strategy NE** $\boldsymbol{\sigma} \in \mathrm{NE}(M)$:

$$R_T(M, \boldsymbol{\sigma}) := \mathbb{E}_{\boldsymbol{s} \sim \boldsymbol{\sigma}}[R_T(M, \boldsymbol{s})].$$

Strong Strategic Regret is defined under the **worst-case NE** in $\mathrm{NE}(M)$:

$$R_T^+(M) := \max_{\boldsymbol{\sigma} \in \mathrm{NE}(M)} R_T(M, \boldsymbol{\sigma}).$$

Weak Strategic Regret is defined under the **best-case NE** in $\mathrm{NE}(M)$:

$$R_T^-(M) := \min_{\boldsymbol{\sigma} \in \mathrm{NE}(M)} R_T(M, \boldsymbol{\sigma}).$$

Naturally, $R_T^-(M) \le R_T^+(M)$.

Limitations of Incentive-Unaware Learning

Proposition (simplified). The algorithm with **oracle knowledge** of both the post-click rewards $\mu_1, \dots, \mu_K$ and arm strategies $s_1, \dots, s_K$, and every round $t \in [T]$ plays the utility maximizing arm

$$i_t = \operatorname*{argmax}_{i \in [K]} u(s_i, \mu_i)$$

suffers **linear regret** $\Omega(T)$ in **every** Nash equilibrium of the arms.

The above suggests that any **incentive-unaware** algorithm that is oblivious to the strategic nature of the arms will fail to achieve low regret.

No-Regret Incentive-Aware Learning

From past observations, we construct **lower** and **upper confidences** on the arms' **strategies** (i.e., click-rates) and the **mean post-click rewards** denoted $\underline{s}_i^t$ and $\overline{s}_i^t$ and $\underline{\mu}_i^t$ and $\overline{\mu}_i^t$, respectively

While playing optimistically w.r.t. $\mu_1, \dots, \mu_K$, we **threaten arms with elimination** if we **detect** them deviating from the desired strategies, i.e., the strategies maximizing the learner's utility.

If we can show that the threat of elimination is **credible** and **justified** it will incentivize arms to play approximately the desired strategies.

Mechanism 0: UCB with Screening (UCB-S)

```
A_0 = [K]
for t = 1, ..., T do
    if A_{t-1} ≠ ∅ then
        Select i_t ∈ argmax_{i ∈ A_{t-1}} μ̄_i^{t-1}
    else
        Select i_t uniformly at random from [K]
    Arm i_t is clicked with probability s_{i_t}, i.e., c_{t,i_t} ~ Bern(s_{i_t})
    if i_t was clicked (c_{t,i_t} = 1) then
        Observe post-click reward r_{t,i_t}
    if s̄_{i_t}^t < min_{μ ∈ [μ̲_{i_t}^t, μ̄_{i_t}^t]} s*(μ) or s̲_{i_t}^t > max_{μ ∈ [μ̲_{i_t}^t, μ̄_{i_t}^t]} s*(μ) then
        Ignore arm i_t in future rounds: A_t ← A_{t-1} \ {i_t}
```

Characterizing the Nash Equilibria under UCB-S

Let $\Delta_i := \mu^* - \mu_i$ with $\mu^* := \max_{j \in [K]} \mu_j$. Let $s^*(\mu) := \operatorname{argmax}_{s \in [0,1]} u(s, \mu)$ denote the strategy maximizing the learner's utility u given post-click reward μ. Hence, $s^*(\mu_i)$ is the **desired strategy** for arm i.

Theorem (simplified): For every pure-strategy profile in the support of a Nash equilibrium, i.e., $\boldsymbol{s} \in \mathrm{supp}(\boldsymbol{\sigma})$ with $\boldsymbol{\sigma} \in \mathrm{NE}(\text{UCB-S})$, we find that

$$s_i = s^*(\mu_i) + O\left(\sqrt{\frac{K \log(T)}{T}} \vee \Delta_i\right).$$

Due to our **uncertainty** about the arms' **strategies** and **rewards**, we can only **approximately** incentivize the desired strategies $s^*(\mu_1), \dots, s^*(\mu_K)$.

In particular, under the **UCB-S Mechanism** every arm i's strategy is $\tilde{O}(\sqrt{K/T} \vee \Delta_i)$ close to the **desired strategy**.

Strong Strategic Regret of UCB-S

Theorem (simplified): The **strong strategic regret** of **UCB-S** is bounded as

$$R_T(\text{UCB-S}) = O\left(\sqrt{KT \log(T)}\right)$$

That is, the **upper bound** holds for **every** equilibrium $\boldsymbol{\sigma} \in \mathrm{NE}(\text{UCB-S})$.

A more detailed bound (see paper) can be derived consisting of a first term due to the arms exploiting UCB-S' uncertainty about their strategies, and a second term due to the standard MAB regret.

Lower Bound on Weak Strategic Regret

Theorem (simplified): For any algorithm M there exists a problem instance such that the algorithm M suffers **weak strategic regret** $R_T^-(M) = \Omega(\sqrt{KT})$.

That is, any algorithm M suffers at least regret $R_T(M, \boldsymbol{\sigma}) = \Omega(\sqrt{KT})$ in **every** of its incentivized equilibria $\boldsymbol{\sigma} \in \mathrm{NE}(M)$.