Incentivizing Exploration in Linear Contextual Bandits under Information Gap

Huazheng Wang*
huazheng.wang@oregonstate.edu
Oregon State University
Corvallis, Oregon, USA

Haifeng Xu
haifengxu@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Chuanhao Li
cl5ev@virginia.edu
University of Virginia
Charlottesville, Virginia, USA

Zhiyuan Liu
zhiyuan.liu@colorado.edu
University of Colorado, Boulder
Boulder, Colorado, USA

Hongning Wang
hw5x@virginia.edu
University of Virginia
Charlottesville, Virginia, USA

ABSTRACT

Contextual bandit algorithms have been popularly used to address interactive recommendation, where the users are assumed to be cooperative to explore all recommendations from a system. In this paper, we relax this strong assumption and study the problem of incentivized exploration with myopic users, where the users are only interested in recommendations with their currently highest estimated reward. As a result, in order to obtain long-term optimality, the system needs to offer compensation to incentivize the users to take the exploratory recommendations. We consider a new and practically motivated setting where the context features employed by the user are more *informative* than those used by the system: for example, features based on users' private information are not accessible by the system. We develop an effective solution for incentivized exploration under such an information gap, and prove that the method achieves a sublinear rate in both regret and compensation. We theoretically and empirically analyze the added compensation due to the information gap, compared with the case where the system has access to the same context features as the user does, i.e., without information gap. Moreover, we also provide a compensation lower bound of this problem.

CCS CONCEPTS

• **Theory of computation** → **Online learning algorithms**; • **Information systems** → **Recommender systems**.

KEYWORDS

Incentivized exploration, linear bandits, information gap

ACM Reference Format:
Huazheng Wang, Haifeng Xu, Chuanhao Li, Zhiyuan Liu, and Hongning Wang. 2023. Incentivizing Exploration in Linear Contextual Bandits under Information Gap. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3604915.3608794

*Work was done while the first author was a PhD student at the University of Virginia.

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09…$15.00
https://doi.org/10.1145/3604915.3608794

1 INTRODUCTION

Contextual bandit algorithms have been popularly used to address the interactive recommendation problems [24, 31, 34], where the system learns the best recommendation policy by interacting with users. Classical bandit research studies the single-party setting, where the system has a full control over which arm to pull. In interactive recommendation, this means all the system's recommendations will be taken by the users for feedback, which enables the system to trade off between exploitation and exploration for long-term optimality. However, in real-world recommender systems, one often faces a *two-party* game between the system and its short-term users, who have *different* interests and roles in this game. Specifically, the system aims at maximizing long-term cumulative reward, which requires exploration in the entire problem space. However, the decision about which recommendation to take is made by the users, and the system can only observe the reward feedback associated with the users' decisions. To make things even worse, the users often act as *myopic* agents, who only seek to maximize their short-term utilities, i.e., exploit the item with the currently best estimated reward. This division leads to the problems of under-exploration and selection bias: the best choice may remain unexplored forever if it appears sub-optimal initially. To align the two parties' interest, the system has to offer compensation to users so that they are motivated to take the exploratory recommendations, which in turn helps system maximize long-term cumulative reward. This is known as the *incentivized exploration* problem [14, 20, 27].

We take restaurant recommendation as an example to illustrate the problem. Myopic agents (customers) tend to visit the restaurant with historically high ratings on the platform. To incentivize exploration, the platform can provide compensations, such as coupons and discounts, to encourage users to visit restaurants currently with lower ratings but also fewer reviews (and hence the estimation might not be accurate there). Then by collecting more feedback on such under-explored restaurants, both the system and the users can better figure out which restaurant turns out to be the best choice. Thus the system can improve its recommendation to other users, while the myopic users may choose better restaurants in future.

Besides recommender systems, incentivized exploration can be applied in a wide range of domains such as e-commerce platforms, crowdsourced information discovery and citizen science (see Frazier et al. [14] for more examples).

The system's goal in incentivized exploration is to minimize total compensation while maximizing cumulative rewards [14, 16, 32]. Existing solutions assume both parties maintain the same reward estimation. This assumption is necessary for the system to compute the compensation based on the difference of users' estimated reward between the currently best choice and the exploratory choice. Under a context-free setting (aka. the Multi-armed Bandit (MAB) [6, 21] in literature), this assumption naturally holds because both parties maintain the same estimated mean reward on each candidate arm. And most existing incentivized exploration solutions work under this setting. However, under the contextual bandit setting [1, 5, 24], the two parties may associate the same observed rewards with *different context features*. For example, in restaurant recommendation the users may access restaurant features related to their dine-in experience such as difficulty of parking or waiting time, which are not accessible by the system. To obtain the same quality of reward estimation, the system has to resort to other observations to infer such user-specific features [7, 33]. This situation can be easily understood by an extreme setting with a finite number of recommendation candidates: the system only observes the ID of each candidate item, while the users employ informative features about the items. As a result, the system suffers from a much slower convergence rate in reward estimation than the users. We refer to this representation asymmetry as the *information gap* between the two parties, which brings in new challenges to incentivized exploration. For example, the system no longer knows which candidate item has the best estimated reward on the user side.

In this paper, we study the problem of incentivized exploration in linear contextual bandits under information gap. We propose an algorithm that effectively incentivizes the users to explore under the information gap so that the system can maintain a sublinear regret in collecting cumulative reward in recommendation. Our key idea is that although the system suffers from information disadvantage and cannot compute the minimum compensation precisely, offering a larger amount of compensation guarantees sufficiency for users to explore. And this added compensation should shrink fast enough such that the total compensation is still sublinear. We prove that in T rounds of interaction our algorithm achieves compensation and regret both in the order of $O(d_v\sqrt{T}\log T)$ with information gap and $O(d_x\sqrt{T}\log T)$ without information gap, where d_x and d_v are the dimensions of context features used by the users and the system, respectively. The results suggest that incentivized exploration is still possible under information gap, and the added cost is realized by the extra compensation that is dominated by d_v. We also prove the compensation lower bound of incentivized exploration in linear contextual bandits, which generalizes the result of compensation lower bound in MAB settings reported in [32]. Our empirical studies in both synthetic data and real-world datasets also validate the effectiveness and cost-efficiency of the proposed algorithm.

2 RELATED WORK

Contextual bandits are emerging solutions to recommender system in both online [15, 24, 34] and offline settings [11, 12, 18, 26] settings. We focused on on-policy bandit learning in this paper. The incentivized exploration problem in multi-armed bandits has been studied since [14, 20]. See Slivkins [30] for an overview. One line of the studies assumes the system has information advantage on observing the full interaction history while users do not [17, 20, 27, 28]. The system leverages the information asymmetry to recommend exploratory arms as long as the users do not have a better choice from their perspective. Simchowitz and Slivkins [29] proposed the first study of incentivizing exploration in reinforcement learning in this line. Another line considers the setting where the interaction history is publicly available to both system and users and the system need to offer compensation to an arm for incentivized exploration [10, 14, 32]. Our setting follows the second line of research.

Incentivized learning with compensation was first studied in [14] in a Bayesian setting with discounted regret and compensation. Chen et al. [10] studied a heterogeneous users setting, where user diversity led to their solution with constant compensation. Agrawal and Tulabandhula [3] considered heterogeneous contexts in a contextual bandit setting. In [32], the authors analyzed the non-Bayesian and non-discounted reward case and showed $O(\log T)$ regret and compensation in a stochastic MAB setting. Liu et al. [25] considered the reward feedback is biased because of the compensation. Kannan et al. [19] considered incentivized exploration for fair recommendation. Our setting is mostly similar to [32], i.e., non-Bayesian and non-discounted reward, but is studied under the linear contextual bandit setting. We should note all the aforementioned studies assume the system and the users share the same information such as arm pulls, rewards and contexts, and the system calculates the compensation based on the shared information. Our setting is strictly more challenging, where the information gap is caused by information asymmetry: the system cannot access the feature vectors employed by the users. As a result, users' reward estimation will be different from the system's and the precise amount of compensation is harder to compute.

3 PROBLEM DEFINITION

Notations and assumptions. We study the problem under a linear contextual bandit setting, where the system interacts with myopic users for T rounds. At each round t, a user u arrives at the system, observes the system-provided recommendations $\mathcal{A}_t$ together with the associated compensation, and pulls an arm a_t (i.e., takes a recommended item). Both the system and the user observe the resulting reward $r_{a_t,t}$ and update their estimations accordingly. In reality, a recommender system interacts with thousands of users whereas each user only occasionally interacts with the system to meet their short-term information need. Therefore, their behavior is naturally myopic. We thus refer to our users as *short-term* users. In a contextual bandit setting, each arm a represents a recommendation candidate and is associated with a context feature vector. In our problem, for arm $a \in \mathcal{A}_t$, the system observes a feature vector $\mathbf{v}_a$ from a d_v-dimensional subspace and the user u observes a feature vector $\mathbf{x}_a$ from a d_x-dimensional subspace. Without loss of generality, we assume $\{\mathbf{x}_a\}$ spans $\mathbb{R}^{d_x}$ and $\{\mathbf{v}_a\}$ spans $\mathbb{R}^{d_v}$ — if not,

the standard PCA technique can be used to reduce the dimensions of raw features to d_x and d_v [22]. Essentially we consider the features span the whole vector space respectively, which means there is no feature without support on both sides and the dimensionality cannot be further reduced.

Assumption 1 (Information Gap). *There exists a linear transformation $P \in \mathbb{R}^{d_x \times d_v}$ (where $d_v \geq d_x$) such that for any arm a,*

$$\mathbf{x}_a = P\mathbf{v}_a \tag{1}$$

Examples of information gap. We now describe a few real-world examples where the gap exists and is inevitable in order to motivate the above assumption on $d_v \geq d_x$, i.e., features used on the user side belong to a lower dimensional space. A notable special case of linear bandits with information gap is a K-armed contextual bandit problem, where the system knows nothing beyond the indices of arms. In this case, the context vectors used by the system are the K-dimensional one-hot vectors e_a, while the users may employ d_x-dimensional feature representations of the same arms. The information gap ($K > d_x$) is encoded in the transformation matrix P. Another example we have discussed is the restaurant recommendation scenario where users may use features related to their dine-in experience to represent the candidate choices. The users can employ these informative features and enjoy faster reward estimation convergence; but the system suffers when it cannot access users' features. In this example, the transformation matrix P hides the user-side information from the system.

Note that having a larger number of features (longer feature vector) is *not* equivalent to having a more informative representation. Another practical example is that the context vectors used by the system may include many useless or redundant features, which should not play any role in reward estimation, i.e., a sparse regression setting. In this example, the system's features are clearly less informative, because of the useless features; but the system does not know which features are useless. This unfortunately leads to a slower convergence of parameter estimation and a wider confidence interval of reward estimation on the system side, which is the key challenge solved in our paper for incentivized exploration.

The information gap between the two parties is characterized by matrix P. The linear transformation assumption is to guarantee both parties face a linear reward mapping, which we state below.

Reward mapping. Following a linear contextual bandit setting, the expected reward of arm a is determined by the inner product between the context features and unknown bandit model parameter. From the user side, we have $\mathbf{E}[r_a] = \mathbf{x}_a^\top \boldsymbol{\theta}_x^*$, where $\boldsymbol{\theta}_x^*$ is the unknown model parameter to be estimated by the user. From Assumption 1, we have $\mathbf{x}_a^\top \boldsymbol{\theta}_x^* = \mathbf{v}_a^\top P^\top \boldsymbol{\theta}_x^*$, which suggests there always exists a parameter $\boldsymbol{\theta}_v^* = P^\top \boldsymbol{\theta}_x^*$ on the system side satisfying the same linear reward mapping. We summarize the reward mapping on the two sides as

$$\mathbf{E}[r_a] = \mathbf{x}_a^\top \boldsymbol{\theta}_x^* = \mathbf{v}_a^\top \boldsymbol{\theta}_v^*$$

After the user pulls arm a_t, the reward $r_{a_t,t}$ is observed by both sides as

$$r_{a_t,t} = \mathbf{E}[r_{a_t}] + \eta_t \tag{2}$$

where η_t is R-sub-Gaussian noise. Without loss of generality, we assume that the norm of the features and parameters are bounded as $\|\mathbf{x}_a\|_2 \leq \|\mathbf{v}_a\|_2 \leq 1, \|\boldsymbol{\theta}_x^*\|_2 \leq 1, \|\boldsymbol{\theta}_v^*\|_2 \leq 1$, which naturally bound the expected reward in the range of $[-1, 1]$ and simplify the analysis later. Note that the assumption of $\|\mathbf{x}_a\|_2 \leq \|\mathbf{v}_a\|_2$ is equivalent as assuming the largest singular value of P is upper bounded by 1. Intuitively, this means the linear transformation does not amplify the magnitude of the features. One can always find the satisfying $\mathbf{x}_a$ by re-scaling $\boldsymbol{\theta}_x^*$ accordingly.

The system and the users estimate their own model parameters using ridge regression separately, denoted as $\hat{\boldsymbol{\theta}}_{v,t}$ and $\hat{\boldsymbol{\theta}}_{x,t}$, by the same observed rewards $\{r_{a_t,t}\}$ but different context features. As a result, the two parties would predict different rewards for the same arm a, denoted as $\hat{r}_{x,a,t} = \mathbf{x}_a^\top \hat{\boldsymbol{\theta}}_{x,t}$ and $\hat{r}_{v,a,t} = \mathbf{v}_a^\top \hat{\boldsymbol{\theta}}_{v,t}$.

Objective. The users and the system have different objectives in this sequential decision making problem: a short-term user aims to maximize his/her instantaneous reward, while the system aims to maximize the long-term cumulative reward. At each round t, without any incentive, a short-term user u will exploit the arm with the highest estimated reward, i.e., $a = \arg\max_{i \in \mathcal{A}_t} \hat{r}_{x,i,t}$. It is well known that such exploitation-only decisions will lead to sub-optimal cumulative reward in the long term. In order to balance exploitation and exploration, the system has to provide compensations to encourage the short-term user to explore. Specifically, the system offers compensation $c_{a,t}$ for pulling arm a. Given the incentives, the user maximizes the instantaneous utility by pulling arm $a_t = \arg\max_{i \in \mathcal{A}_t} \hat{r}_{x,i,t} + c_{i,t}$.

The system seeks to maximize the cumulative reward, or equivalently, minimize the *cumulative regret* while also minimizing the *total compensation* in expectation. The system's regret is defined as

$$R(T) = \sum_{t=1}^{T} \left(\mathbf{E}[r_{a_t^*,t}] - \mathbf{E}[r_{a_t,t}] \right) \tag{3}$$

where a_t^* is the optimal arm with the highest expected reward at time t. The total compensation is defined as

$$C(T) = \sum_{t=1}^{T} \mathbf{E}[c_{a_t,t}] \tag{4}$$

An effective incentivized exploration method should balance the trade-off among exploration, exploitation and compensation to obtain *sublinear* cumulative regret and *sublinear* total compensation.

4 METHOD

We present our solution on incentivized exploration under information gap when the system explores according to the Linear Upper Confidence Bound (LinUCB) strategy [1, 13, 24]. Then we show that the solution can be easily adapted to the simpler problem setting of incentivized exploration without the information gap. We leave the study of incentivizing other exploration strategies such as Thompson Sampling [2, 4, 9] as future work.

4.1 Incentivized exploration under information gap

We present Algorithm 1 to show how the system incentivizes myopic users to follow the desired exploration strategy under information gap. At each round t, the system and the user u_t observe context features $\{\mathbf{v}_a\}_{a \in \mathcal{A}_t}$ and $\{\mathbf{x}_a\}_{a \in \mathcal{A}_t}$ respectively for the same arm set $\mathcal{A}_t$. Both parties estimate their parameters using ridge

Algorithm 1 Incentivized LinUCB under Information Gap

Inputs: λ, δ
Initialize: $\mathbf{A}_v = \lambda \mathbf{I}_{d_v}, \mathbf{b}_v = 0$
for $t = 1$ to T **do**
  System and user u_t observe context vectors $\{\mathbf{v}_a\}_{a\in\mathcal{A}_t}$ and $\{\mathbf{x}_a\}_{a\in\mathcal{A}_t}$ respectively
  System calculates compensation $c_{a,t} = 4CB_{v,t}(\mathbf{v}_a)$ for arm a (Eq (6))
  // Ridge regression on the user side:
  $\mathbf{A}_{x,t} = \sum_{i=1}^{t-1} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^\mathsf{T} + \lambda \mathbf{I}_{d_x}, \mathbf{b}_{x,t} = \sum_{i=1}^{t-1} \mathbf{x}_{a_i} r_{a_i}$
  $\hat{\boldsymbol{\theta}}_{x,t} = \mathbf{A}_{x,t}^{-1}\mathbf{b}_{x,t}$
  User pulls arm $a_t = \arg\max_{a\in\mathcal{A}_t} \hat{r}_{x,a,t} + c_{a,t}$
  Reward r_{a_t} is revealed
  // Ridge regression on the system side:
  $\mathbf{A}_{v,t+1} = \mathbf{A}_{v,t} + \mathbf{v}_{a_t}\mathbf{v}_{a_t}^\mathsf{T}, \mathbf{b}_{v,t+1} = \mathbf{b}_{v,t} + \mathbf{v}_{a_t} r_{a_t}$
  $\hat{\boldsymbol{\theta}}_{v,t+1} = {\mathbf{A}_{v,t+1}}^{-1}\mathbf{b}_{v,t+1}$
end for

regression with same reward observations and their own features. The system needs to motivate the user to explore arm a_t according to the LinUCB strategy based on its current parameter estimation $\hat{\boldsymbol{\theta}}_{v,t}$. To achieve so, the system offers compensation $c_{a_t,t}$ to arm a_t according to Eq (6). Note that the system does not offer incentives to the other arms and sets $c_{i,t} = 0, \forall i \neq a_t$. The myopic user pulls the arm that maximizes the sum of his/her estimated reward $\hat{r}_{x,a,t}$ and the compensation $c_{a,t}$. We will see in Lemma 2 that the user is guaranteed to pull the system desired arm a_t.

Denote $CB_{x,t}(\mathbf{x}_a)$ as the width of the user's estimation confidence interval of arm a at time t, which is computed as $CB_{x,t}(\mathbf{x}_a) = \alpha_{x,t}\|\mathbf{x}_a\|_{A_{x,t}^{-1}}$, where $\alpha_{x,t} = R\sqrt{d_x \log\frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$. $\alpha_{x,t}$ is the upper bound of the width of confidence ellipsoid and is set according to Theorem 2 of [1]. Similar to $CB_{x,t}(\mathbf{x}_a)$, we denote the width of confidence interval on the system side as $CB_{v,t}(\mathbf{v}_a) = \alpha_{v,t}\|\mathbf{v}_a\|_{A_{v,t}^{-1}}$, where $\alpha_{v,t} = R\sqrt{d_v \log\frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$.

The key challenge in incentivized exploration under information gap is that the system does not maintain the same reward estimation as the user's, because the two sides use different features to estimate rewards. This prevents us from computing the minimum required compensation and makes the problem non-trivial. We have to carefully determine the compensation: a larger amount of incentive is required to guarantee that user will explore while we also need to keep the incentives small to maintain a sublinear total compensation. We first use the following lemma to show that on the same arm, the confidence interval of the system's reward estimation is no smaller than the confidence interval of the user's estimate. This lemma guarantees in Algorithm 1 the system provides sufficient incentive for the user to pull its desired arms for exploration.

Lemma 1. *Consider two ridge regression estimators that estimate the model parameters with the same reward observations but different features satisfying Assumption 1. For any $t \geq 0$ and arm $a \in \mathcal{A}_t$, we have*

$$CB_{v,t}(\mathbf{v}_a) \geq CB_{x,t}(\mathbf{x}_a), \tag{5}$$

i.e., the confidence interval maintained on the system side is no smaller than that on the user side.

Proof. Note that $CB_{v,t}(\mathbf{v}_a) = \alpha_{v,t}\|\mathbf{v}_a\|_{\mathbf{A}_{v,t}^{-1}}$ and $CB_{x,t}(\mathbf{x}_a) = \alpha_{x,t}\|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}}$. In the following, we separately prove $\|\mathbf{v}_a\|_{\mathbf{A}_{v,t}^{-1}} \geq \|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}}$ and $\alpha v, t \geq \alpha x, t$. By Eq (1), we have $\mathbf{A}_{x,t} - \lambda\mathbf{I} = \sum_{i=1}^{t}\mathbf{x}_{a_i}\mathbf{x}_{a_i}^\mathsf{T} = \sum_{i=1}^{t} P\mathbf{v}_{a_i}\mathbf{v}_{a_i}^\mathsf{T}P^\mathsf{T} = P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}$ and $\|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}} = \sqrt{\mathbf{x}_a^\mathsf{T}\mathbf{A}_{x,t}^{-1}\mathbf{x}_a} = \sqrt{\mathbf{v}_a^\mathsf{T}P^\mathsf{T}\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right)^{-1}P\mathbf{v}_a}$. In order to prove

$$\mathbf{v}_a^\mathsf{T}\mathbf{A}_{v,t}^{-1}\mathbf{v}_a \geq \mathbf{x}_a^\mathsf{T}\mathbf{A}_{x,t}^{-1}\mathbf{x}_a = \mathbf{v}_a^\mathsf{T}P^\mathsf{T}\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right)^{-1}P\mathbf{v}_a,$$

we show that $\mathbf{A}_{v,t}^{-1} - P^\mathsf{T}\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right)^{-1}P$ is a positive semi-definite matrix using the property of Schur complement. Specifically, denote

$$M = \begin{bmatrix} \mathbf{A}_{v,t}^{-1} & P^\mathsf{T} \\ P & \left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I} \end{bmatrix}.$$

We have

$$\begin{aligned} M/\mathbf{A}_{v,t}^{-1} &= \left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I} - \left(P^\mathsf{T}\right)^\mathsf{T}\mathbf{A}_{v,t}P^\mathsf{T} \\ &= P\mathbf{A}_{v,t}P^\mathsf{T} - \lambda PP^\mathsf{T} + \lambda\mathbf{I} - P\mathbf{A}_{v,t}P^\mathsf{T} \\ &= \lambda\left(\mathbf{I} - PP^\mathsf{T}\right) \geq 0 \end{aligned}$$

where the last inequality is because P's largest singular value is smaller than 1, the eigenvalues of PP^T are smaller than 1 and thus $\mathbf{I} - PP^\mathsf{T} \geq 0$. As $\mathbf{A}_{v,t}^{-1} \succ 0$ and $M/\mathbf{A}_{v,t}^{-1} \geq 0$, according to the property of Schur complement we have $M \succeq 0$. Then as $\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I} = \mathbf{A}_{x,t} \succ 0$ and $M \succeq 0$, applying the property again we have $M/\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right) \geq 0$, which gives us $\mathbf{A}_{v,t}^{-1} - P^\mathsf{T}\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right)^{-1}P \succeq 0$. Therefore, we have $\mathbf{v}_a^\mathsf{T}\mathbf{A}_{v,t}^{-1}\mathbf{v}_a - \mathbf{v}_a^\mathsf{T}P^\mathsf{T}\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right)^{-1}P\mathbf{v}_a \geq 0$ for any $\mathbf{v}_a \in \mathbb{R}^{d_v}$. Moreover, note that $\alpha v, t = R\sqrt{d_v \log\frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$, $\alpha x, t = R\sqrt{d_x \log\frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$ (see Lemma 3), and $d_v \geq d_x$, so $\alpha v, t \geq \alpha x, t$. Combining the two inequalities, we have $CB_{v,t}(\mathbf{v}_a) \geq CB_{x,t}(\mathbf{x}_a), \forall t \geq 0, a \in \mathcal{A}_t$. □

Based on Lemma 1, we can prove the sufficient compensation to incentivize user to pull the desired arm under information gap.

Lemma 2. *For all $t \geq 0$, with probability at least $1 - 2\delta$, the users are incentivized to pull the desired arm with compensation*

$$c_{a_t,t} = 4CB_{v,t}(\mathbf{v}_{a_t}) \tag{6}$$

to arm

$$a_t = \arg\max_a \left(\mathbf{v}_a^\mathsf{T}\hat{\boldsymbol{\theta}}_{v,t} + 2CB_{v,t}(\mathbf{v}_a)\right), \tag{7}$$

i.e., the arm with the highest (relaxed) upper confidence bound according to the system's estimate. [1]

[1]While we considered myopic users in the paper, this lemma and the following result can be generalized to other user models. For example, explorative user who makes decisions by maximizing upper confidence bound $g = \arg\max_i \hat{r}_{x,i,t} + CB_{x,t}(\mathbf{x}_i)$ can

Proof. In order to incentivize the user to pull arm a_t, the *minimum required compensation* is $\max_i \hat{r}_{x,i,t} - \hat{r}_{x,a_t,t}$. However, since the system cannot access the context features the user uses and thus maintains different reward estimates, it has to provide compensation larger than the minimum required amount.

Denote the user's greedy choice as $g = \arg\max_i \hat{r}_{x,i,t}$. To show that $c_{a_t,t}$ is sufficient, we need to prove that the user prefers the exploratory arm a_t with compensation over his/her greedy choice, i.e., $\hat{r}_{x,g,t} \le \hat{r}_{x,a_t,t} + c_{a_t,t}$.

Based on Lemma 3, for all $t \ge 0$ with probability at least $1-\delta$, we have $|\hat{r}_{x,a,t} - \mathbf{E}[r_a]| \le CB_{x,t}(\mathbf{x}_a)$ and $|\hat{r}_{v,a,t} - \mathbf{E}[r_a]| \le CB_{v,t}(\mathbf{v}_a)$ hold for any arm a. Using the union bound, with probability at least $1-2\delta$ we have

$$|\hat{r}_{x,a,t} - \hat{r}_{v,a,t}| \le |\hat{r}_{x,a,t} - \mathbf{E}[r_a]| + |\mathbf{E}[r_a] - \hat{r}_{v,a,t}| \le CB_{x,t}(\mathbf{x}_a) + CB_{v,t}(\mathbf{v}_a) \tag{8}$$

Then we can bound the user's reward estimate from the system side as follows,

$$\begin{aligned}\hat{r}_{x,g,t} &\le \hat{r}_{v,g,t} + CB_{x,t}(\mathbf{x}_g) + CB_{v,t}(\mathbf{v}_g)\\ &\le \hat{r}_{v,g,t} + 2CB_{v,t}(\mathbf{v}_g) \le \hat{r}_{v,a_t,t} + 2CB_{v,t}(\mathbf{v}_{a_t})\\ &\le \hat{r}_{x,a_t,t} + CB_{x,t}(\mathbf{v}_{a_t}) + CB_{v,t}(\mathbf{v}_{a_t}) + 2CB_{v,t}(\mathbf{v}_{a_t})\\ &\le \hat{r}_{x,a_t,t} + 4CB_{v,t}(\mathbf{v}_{a_t})\end{aligned} \tag{9}$$

where the first and fourth steps are based on Eq (8), the second and last steps are based on Lemma 1, and the third inequality is based on the UCB strategy in Eq (7). □

It is worth noting that the system follows a more optimistic arm selection strategy in Eq (7) using a confidence interval twice larger than the classical LinUCB algorithm's. We follow this relaxed upper confidence bound because we need to consider the uncertainty on both parties as the first step of the derivation in Eq (9) suggested (details can be found in the appendix). It is unclear whether we can incentivize the user to follow the classical LinUCB algorithm. Intuitively, our exploration strategy results in a twice larger regret than the classical LinUCB's, which is still in the same order for T. We provide the detailed regret and compensation upper bound of Algorithm 1 in the Analysis Section.

4.2 Incentivized exploration without information gap

Our solution can be easily adopted to solve the incentivized exploration for linear bandits *without* information gap, where the system and the users observe same context features, i.e., $P = \mathbf{I}$. It is a simple derivation from our results in information gap setting, but it is with independent interest to the community: this setting can also be viewed as a contextual version of incentivized exploration for MAB in Wang and Huang [32] where we generalize from $\mathbf{x}_a = e_a \in \mathbb{R}^K$ to a real vector, and has not been reported in existing literature. In Algorithm 2, we show how the system incentivizes the myopic users to follow the desired exploration strategy without information gap.

Without information gap, the system and the users maintain the same parameter and reward estimations, and the *minimum required compensation* to incentivize the user according to LinUCB equals to the difference of the estimated rewards between the currently best arm and the exploratory arm. The system thus only needs to offer compensation by,

$$c_{a_t,t} = \max_i \hat{r}_{x,i,t} - \hat{r}_{x,a_t,t} \tag{10}$$

to arm $a_t = \arg\max_a \left(\mathbf{x}_a^\top \hat{\boldsymbol{\theta}}_{x,t} + CB_{x,t}(\mathbf{x}_a)\right)$. The user will pull the exploratory arm, because $a_t = \arg\max_i \hat{r}_{x,i,t} + c_{i,t}$, i.e., arm a_t can maximize user's instantaneous utility. Since Algorithm 2 guarantees the users are incentivized to pull arms according to LinUCB, its regret follows LinUCB's in the order of $O(d_x\sqrt{T}\log T)$ (see Theorem 3 of [1]). Its compensation upper bound is stated below.

Algorithm 2 Incentivized LinUCB without Information Gap

Inputs: λ, δ
Initialize: $\mathbf{A}_x = \lambda\mathbf{I}, \mathbf{b}_x = 0$
for $t = 1$ to T **do**
  System and user u_t observe context vectors $\{\mathbf{x}_a\}_{a\in\mathcal{A}_t}$
  // Ridge regression:
  $\mathbf{A}_{x,t} = \sum_{i=1}^{t-1} \mathbf{x}_{a_i}\mathbf{x}_{a_i}^\top + \lambda\mathbf{I}_{d_x}, \mathbf{b}_{x,t} = \sum_{i=1}^{t-1} \mathbf{x}_{a_i} r_{a_i}, \hat{\boldsymbol{\theta}}_{x,t} = \mathbf{A}_{x,t}^{-1}\mathbf{b}_{x,t}$
  System calculate compensation $c_{a,t} = \max_i \hat{r}_{x,i,t} - \hat{r}_{x,a,t}$ for arm a (Eq (10))
  User pulls arm $a_t = \arg\max_{a\in\mathcal{A}} \hat{r}_{x,a,t} + c_{a,t}$
  Reward r_{a_t} is revealed
end for

Theorem 1 (Compensation upper bound without information gap). *With probability at least $1-\delta$, the total compensation provided in Algorithm 2 is upper bounded by*

$$C(T) \le \left(R\sqrt{d_x \log\frac{1+T/\lambda}{\delta}} + \sqrt{\lambda}\right)\sqrt{Td_x\log(\lambda + \frac{T}{d_x})}$$

Proof Sketch. First, with a high probability the compensation at round t is upper bounded by the confidence interval, i.e., $c_{a_t,t} \le CB_{x,t}(\mathbf{x}_{a_t})$. The total compensation can then be upper bounded by $\sum_t CB_{x,t}(\mathbf{x}_{a_t})$, which can be bounded using Lemma 11 of [1]. □

Note that without information gap, both the regret and compensation upper bounds are in the order of $O(d_x\sqrt{T}\log T)$, with a linear dependency on the feature dimension d_x.

Discussion. Without information gap, i.e., the two parties have access to the same features and maintain the same reward estimations, the system can offer the minimum required compensation as shown in Eq (10) to incentivize exploration. With information gap, compensate by Eq (6) can still successfully incentivize exploration in a high probability manner, but it is inevitably larger than the minimum amount. More specifically, without information gap the required compensation can be computed deterministically in Eq (10); otherwise, the system can only estimate the reward difference with a high probability (as shown in Lemma 2). We also notice without information gap the system does not compensate if the greedy choice also has the largest upper confidence bound, which happens more often in the later rounds when the reward estimation converges. But with information gap, our algorithm always compensates, because $CB_{v,t}(\mathbf{v}_{a_t}) > 0$, i.e., the system does

be incentivized by $c_{a_t,t} = 5CB_{v,t}(\mathbf{v}_{a_t})$. This can be proved by adding a $CB_{x,t}(\mathbf{v}_{a_t})$ to the LHS of Eq (9) and $CB_{v,t}(\mathbf{v}_{a_t})$ to the RHS.

not know if the user's greedy choice is also preferred in terms of its UCB. We will show in the next section that the total compensation is still sublinear under information gap.

What if $d_v < d_x$. We have discussed the setting where the users have information advantage, i.e., $d_v \geq d_x$, in Section 4.1. We now discuss the setting where the system has information advantage, i.e., $d_v < d_x$, as a complement. In this setting, although the system can learn faster than the users, such information advantage cannot accelerate users' reward estimation, i.e., confidence interval on user side still depends on d_x. Thus the amount of compensation required to incentivize users to explore cannot be reduced to depend on d_v even if the system knows more information.

Computation complexity. The two presented algorithms have similar time complexity as LinUCB algorithm. If applying Sherman–Morrison formula to accelerate matrix inverse with rank-one update, the time complexities of Algorithm 1 and Algorithm 2 are $O(d_v^2 T)$ and $O(d_x^2 T)$, respectively.

5 ANALYSIS

We first analyze the regret and compensation upper bound of Algorithm 1, and then discuss the compensation lower bound of the problem.

5.1 Regret and compensation upper bound

Theorem 2. *With probability at least $1-3\delta$, the cumulative regret of Algorithm 1 is bounded by*

$$R(T) \leq \left(2R\sqrt{d_v \log \frac{1+T/\lambda}{\delta}} + \sqrt{\lambda}\right)\sqrt{T d_v \log(\lambda + \frac{T}{d_v})}$$

Theorem 2 shows that the cumulative regret of Algorithm 1 is in the order of $O(d_v\sqrt{T \log T})$. The proof mostly follows the regret analysis of LinUCB, though we have to use a wider confidence interval for exploration. Note that the resulting probability is $1-3\delta$, because the users will follow the system's exploration strategy with probability at least $1-2\delta$ as shown in Lemma 2 and the confidence bound holds with probability at least $1-\delta$.

Theorem 3. *With probability at least $1-2\delta$, the total compensation provided in Algorithm 1 is upper bounded by*

$$C(T) \leq \left(4R\sqrt{d_v \log \frac{1+T/\lambda}{\delta}} + \sqrt{\lambda}\right)\sqrt{T d_v \log(\lambda + \frac{T}{d_v})}$$

Theorem 3 shows that the total compensation of Algorithm 1 is in the order of $O(d_v\sqrt{T \log T})$. Combining Theorem 2 and 3, we show that our proposed algorithm can incentivize exploration under information gap and achieve both sublinear regret and compensation. We notice that the two upper bounds linearly depend on the system's feature dimension d_v. Comparing to the no information gap setting where we showed both the regret and compensation is in the order of $O(d_x\sqrt{T \log T})$, the added regret and compensation are $O\big((d_v - d_x)\sqrt{T \log T}\big)$. And the corresponding high probability guarantee drops a little. These results suggest that the complexity/difficulty of the problem is characterized by the dimensionality of the observed context features, which is exactly where the information gap comes from.

Remark. Our results can be generalized to the setting where users observe different features for the same arm, i.e., $\mathbf{x}_{a,u}$ is associated with arm a for user u. In this setting, Algorithm 1 can still incentivize users to explore and the theorems still hold, as long as Assumption 1 holds for every user, i.e., there exists a P_u for any user u.

5.2 Compensation lower bound

We now prove a gap-dependent asymptotic compensation lower bound of incentivized exploration in linear bandits with finite arms, and show that our result recovers the lower bound in non-contextual bandits in [32].

Let $G_{x,T} = \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{x}_{a_t}\mathbf{x}_{a_t}^\intercal\right]$. Without loss of generality, assume arm 1 is the best arm and $\Delta_a = \mathbf{E}[r_1] - \mathbf{E}[r_a] = (\mathbf{x}_1 - \mathbf{x}_a)^\intercal \boldsymbol{\theta}^*$ is the reward gap between arm a and the best arm.

Theorem 4 (Compensation lower bound without information gap). *Consider any consistent algorithm observing context features $\{\mathbf{x}_a\}_{a\in\mathcal{A}}$ that guarantees an $o(T^p)$ regret upper bound for any $T > 0$ and $0 < p \leq 1$. In order to incentivize a user with a least square estimator of rewards to follow the algorithm's choice, the total compensation $C(T)$ for sufficiently large T is*

$$\Omega\left(c_x(\mathcal{A}, \boldsymbol{\theta}^*)\log(T)\right),$$

where $c_x(\mathcal{A}, \boldsymbol{\theta}^*)$ is the optimal value of the following optimization problem

$$\begin{aligned} c_x(\mathcal{A}, \boldsymbol{\theta}^*) &= \inf_{\alpha \geq 0} \sum_{a\in\mathcal{A}} \alpha_{\mathbf{x}_a} \frac{\Delta_a}{3} \\ s.t.\ \|\mathbf{x}_a\|^2_{H^{-1}_{x,T}} &\leq \frac{\Delta_a^2}{2}, \forall \mathbf{x}_a \text{ with } \Delta_a > 0 \end{aligned} \tag{11}$$

where $H_{x,T} = \sum_{a\in\mathcal{A}} \alpha_{\mathbf{x}_a}\mathbf{x}_a\mathbf{x}_a^\intercal$.

While we cannot further simplify the expression of $c_x(\mathcal{A}, \boldsymbol{\theta}^*)$ since this is an instance-dependent lower bound, we construct an example to further illustrate the lower bound analysis.

Example. When $\{\mathbf{x}_a = e_a \in \mathbb{R}^{d_x}\}_{a\in\mathcal{A}}$ are the basis vectors, the problem reduces to a non-contextual K-armed bandit with $K = d_x$. By setting $\|\mathbf{x}_a\|^2_{H^{-1}_{x,T}} = \Delta_a^2/2$, we have $\alpha_{\mathbf{x}_a} = 2/\Delta_a^2$ and $c_x(\mathcal{A}, \boldsymbol{\theta}^*) = \sum_{a\in\mathcal{A},\Delta_a>0} \frac{2}{3\Delta_a}$. This gives us the compensation lower bound as follows,

$$C(T) = \Omega\left(\sum_{a\in\mathcal{A},\Delta_a>0} \frac{\log(T)}{\Delta_a}\right)$$

This result recovers the lower bound of incentivized exploration in non-contextual bandits in [32]. We also notice that the result can be further bounded as

$$C(T) = \Omega\left(\frac{d_x \log(T)}{\max_{a\in\mathcal{A}} \Delta_a}\right),$$

where we observe a linear dependency on dimension d_x.

Note that our compensation lower bound has an order $\Omega(\log(T))$, because it is gap-dependent. We leave the question of whether one can obtain an $\Omega(\sqrt{T})$ gap-independent compensation lower bound for general infinite arm setting, which will match our upper bound in Theorem 3, as an open problem.

Corollary 1 (Compensation lower bound under information gap). *Consider any consistent algorithm observing context features $\{\mathbf{v}_a\}_{a\in\mathcal{A}}$ that guarantees an $o(T^p)$ regret upper bound for any $T > 0$ and $0 < p \leq 1$. To incentivize the user who observes context features $\{\mathbf{x}_a\}_{a\in\mathcal{A}}$ satisfying Assumption 1 with a least square estimator, the total compensation $C(T)$ for sufficiently large T is*

$$\Omega\left(c_v(\mathcal{A},\boldsymbol{\theta}^*)\log(T)\right),$$

where $c_v(\mathcal{A},\boldsymbol{\theta}^*)$ is the optimal value of the following optimisation problem

$$c_v(\mathcal{A},\boldsymbol{\theta}^*) = \inf_{\alpha\geq 0}\sum_{a\in\mathcal{A}}\alpha_{\mathbf{v}_a}\frac{\Delta_a}{3}$$

$$s.t.\ \|\mathbf{v}_a\|^2_{H^{-1}_{v,T}} \leq \frac{\Delta_a^2}{2}, \forall \mathbf{v}_a \text{ with } \Delta_a > 0$$

where $H_{v,T} = \sum_{a\in\mathcal{A}}\alpha_{\mathbf{v}_a}\mathbf{v}_a\mathbf{v}_a^\intercal$.

Considering a similar example of K-armed bandit setting where $d_v = K$, we can obtain

$$C(T) = \Omega\left(\frac{d_v\log(T)}{\max_{a\in\mathcal{A}}\Delta_a}\right)$$

where we observe a linear dependency on dimension d_v.

6 EXPERIMENTS

We evaluate the effectiveness of our proposed incentivized exploration solution on both synthetic data and real-world datasets to confirm our theoretical analysis about the proposed solutions.

6.1 Synthetic data

• **Setup**. In our simulations, we generate a size-K recommendation candidate pool $\mathcal{A}$, in which each candidate item a is associated with a d_v-dimension vector $\mathbf{v}_a$ as the system observed features and a d_x-dimension vector $\mathbf{x}_a$ as the user observed features. Each dimension of $\mathbf{v}_a$ is drawn from a set of zero-mean Gaussian distributions with variances sampled from a uniform distribution $U(0,1)$. Each $\mathbf{v}_a$ is then normalized to $\|\mathbf{v}_a\|_2 = 1$. We then sample the elements of the $d_x \times d_v$ transformation matrix P from $N(0,1)$ and normalize each row i by $\|P_i\|_2 = 1$. Following Assumption 1, the user observed features $\mathbf{x}_a$ are generated as $\mathbf{x}_a = P\mathbf{v}_a$. P guarantees that $\|\mathbf{x}_a\|_2 \leq \|\mathbf{v}_a\|_2 = 1$. User's model parameter $\boldsymbol{\theta}^*_x$ is sampled from $N(0,1)$ and normalized to $\|\boldsymbol{\theta}^*_x\|_2 = 1$. System's model parameter is set to $\boldsymbol{\theta}^*_v = P\boldsymbol{\theta}^*_x$. At each round t, the same set of recommendation candidate pool were presented to all the algorithms, but the system and the user observe their different features respectively. After the user takes an item a_t, both the user and the system observe its reward following Eq (2). We set d_x to 5, d_v to 100, the standard deviation of Gaussian noise η_t to 0.1, and the arm pool size K to 100 in our simulations.

We compare the following algorithms: 1) ILinUCB-InfoGap: our Algorithm 1 where $\{\mathbf{v}_a\}_{a\in\mathcal{A}_t}$ is observed by the system; 2) ILinUCB-NoGap: our Algorithm 2 where both the system and the user observe $\{\mathbf{x}_a\}_{a\in\mathcal{A}}$; 3) NoCompensation: a baseline system that does not offer any compensation to the user. The myopic user estimates the reward with ridge regression and always take the current best item considering the incentives. We set the probability $\delta = 0.01$ and regularization coefficient $\lambda = 0.1$ for all the algorithms.

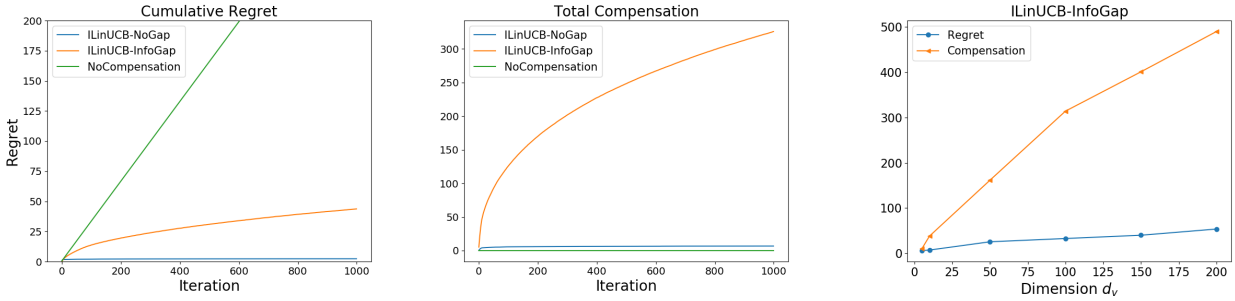
• **Results and analysis.** We report the averaged results of 10 runs where in each run we sample a random model parameter $\boldsymbol{\theta}^*_x$. In Figure 1(a), we observe that without providing any compensation, the myopic user suffers a linear regret, which emphasizes the importance of incentivized exploration in interactive recommendation. Both ILinUCB-InfoGap and ILinUCB-NoGap enjoy sublinear regret and compensation. The added regret of ILinUCB-InfoGap shows the algorithm explores slower in the large $\mathbb{R}^{d_v}$ space because of the information gap. We notice that the total compensation of ILinUCB-InfoGap in Figure 1(b) is sublinear and keeps increasing. The algorithm has to always compensate due to the information gap as we discussed before. ILinUCB-NoGap, however, rarely compensates in the later stage. This is because when system explored sufficiently, greedy choice on the user side agrees with the UCB strategy on the system side, and thus no compensation is needed. In Figure 1(c), we vary the dimension of system's feature d_v from 5 to 200 while fixing $d_x = 5$. We observe that both regret and compensation increase linearly with d_v, which confirms our theoretical upper bounds.

In Figure 2(a) and Figure 2(b), we simulate a K-armed bandit setting where only the indices of the items are available to the system. The system sets $\mathbf{v}_a = e_a \in \mathbb{R}^K$. The rest of the settings are the same as described above. In this setting, our ILinUCB-InfoGap explores almost equivalently to UCB1 [6] and can be viewed as a more optimistic version of the Incentivized UCB algorithm in [32] with a wider confidence interval due to the information gap. The system observes the least information in this setting. We notice that its regret and compensation are much larger than the results in Figure 1 where $\{\mathbf{v}_a\}_{a\in\mathcal{A}}$ is more informative about the rewards. This again confirms that the system inevitably suffers higher regret and compensation when the features are less informative.

6.2 Real-world datasets

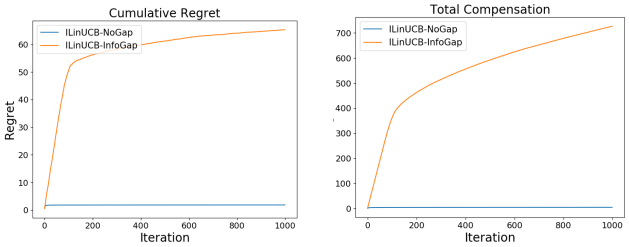
• **Setup.** We now evaluate our solution on two real-world datasets, LastFM and Delicious. The LastFM dataset is extracted from the music streaming service Last.fm, and the Delicious dataset is extracted from the social bookmark sharing service Delicious. The two datasets are created by the HetRec 2011 workshop with the goal of investigating the usage of heterogeneous information in recommender systems [2]. The LastFM dataset contains 1,892 users and 17,632 items (artists). The Delicious dataset contains 1,861 users and 69,226 items (URLs). We pre-process the datasets following [8, 34]. Specifically, we build recommendation candidate pool with size $K = 25$ by first selecting one item from those non-zero reward items based on the observations in the dataset, and then randomly selecting the other 24 from those zero-reward items. The reward is defined as follows: on LastFM dataset, if a user listened to the recommended artist at least once, the reward is 1, otherwise 0; on Delicious dataset, the reward of recommending a bookmarked URL is 1, otherwise 0. At each round, the system and the user observe the same candidate pool but with different features. To construct context features for the two parties, we first extract the TF-IDF feature vector of an item using all tags associated with the item, which uniquely represents the content of that item. To create information

[2]Datasets and their full description is available at http://grouplens.org/datasets/hetrec-2011



(a) Regret (b) Compensation (c) Varying dimension d_v

Figure 1: Simulation result on randomly sampled features with $d_x = 5$ and $d_v = 100$;



(a) Regret (b) Compensation

Figure 2: MAB setting where the system only observes the indices of the arms.

gap between the system and the users, we apply PCA to map the TF-IDF feature vectors to different dimensions. Users observe the 25-dimension context features, which are the same as those used in the single-party setting [8, 34], i.e., $d_x = 25$. In ILinUCB-NoGap, the system observes the same features as the users with $d_v = 25$. In ILinUCB-InfoGap, we test two different levels of information gap by setting d_v to 100 and 250. Note that the features with different dimensions produced by PCA naturally satisfy our Assumption 1. Since the real-world datasets did not include users' response to incentives, we simulate users' myopic decision with ridge regression. We set the hyperparameters of all algorithms the same as what we used for synthetic data if not specified.

• **Result and analysis.** In Figure 3 (a) & (c), we report the reward ratio normalized by the reward collected from a random policy, following the setting from [8, 34]; and the resulting performance curve is thus the higher the better. We observe that ILinUCB-NoGap achieves larger reward ratio than ILinUCB-InfoGap because of more efficient exploration with more informative context features. ILinUCB-InfoGap with $d_v = 250$ obtains a smaller reward ratio than the algorithm with $d_v = 100$. This follows our theoretical analysis that with larger information gap, it is slower to explore the $\mathbb{R}^{d_v}$ space.

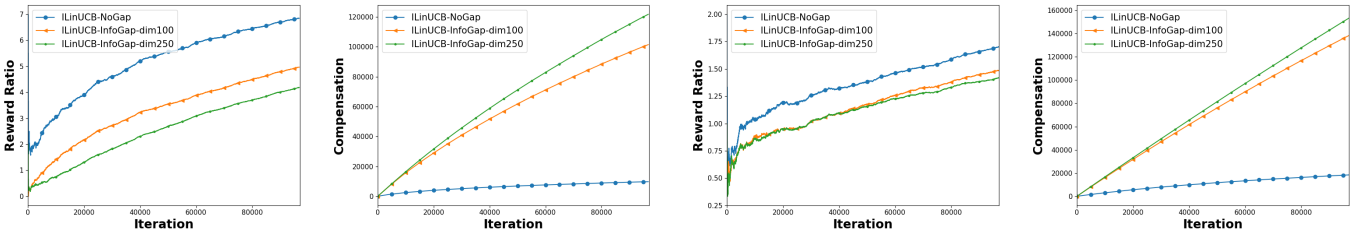
We show the total compensation in Figure 3 (b) & (d). We notice that ILinUCB-InfoGap requires significantly more compensation than ILinUCB-NoGap to incentivize the exploration, and the difference in total compensation between ILinUCB-InfoGap with $d_v = 250$ and $d_v = 100$ is much smaller than the difference between ILinUCB-InfoGap with $d_v = 100$ and ILinUCB-NoGap. This suggests that the larger total compensation is not only because of the slow exploration in the $\mathbb{R}^{d_v}$ space. The main reason is in ILinUCB-NoGap, the system does not compensate if the user's greedy choice is the same as the system's decision. ILinUCB-InfoGap, on the other hand, has to always compensate due to information gap. This result suggests an interesting practical direction for future research: the total compensation could be reduced if the system can determine which item is most preferred by the user with information gap. Overall, the results validate our theoretical understanding that the system suffers higher compensation and lower reward from observing less informative features.

7 CONCLUSIONS AND FUTURE WORK

In this paper, we introduced a new and practically-motivated problem of incentivized exploration under information gap in linear contextual bandits. The key challenge is the information asymmetry in the observed context features between a system and a myopic user. We proposed an algorithm that offers sufficient compensation to guarantee users would follow LinUCB's exploration strategy. We proved the regret and compensation upper bound of our algorithm are in the order of $O(d_v\sqrt{T}\log T)$ under information gap and $O(d_x\sqrt{T}\log T)$ without information gap. We also analyzed the compensation lower bound of the problem. In our future work, we plan to study how to incentivize the users following other types of exploration strategies such as Thompson Sampling [2, 4, 9]. Our empirical study also suggests that even under the information gap, the algorithm could still have a chance to stop the incentives earlier to reduce the total cost. How to theoretically analyze this is an interesting future direction. It is also important to investigate whether we can obtain a gap-independent $\Omega(\sqrt{T})$ compensation lower bound to match with the upper bound. Another interesting direction is to consider non-linear feature transformation between the system and the users.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This work was supported by NSF IIS 2007492, IIS-2128019, IIS-1838615, CCF-2303372, ARO W911NF-23-1-0030 and Bloomberg Data Science Ph.D. Fellowship.



(a) LastFM reward (b) LastFM compensation (c) Delicious reward (c) Delicious compensation

Figure 3: Results on LastFM and Delicious datasets.

A MISSING PROOFS

Lemma 3 (Theorem 2 of [1]). *With probability at least $1-\delta$, the parameter $\boldsymbol{\theta}_x^*$ lies in the confidence ellipsoid of $\hat{\boldsymbol{\theta}}_{x,t}$ satisfying*

$$\|\hat{\boldsymbol{\theta}}_{x,t} - \boldsymbol{\theta}_x^*\|_{A_{x,t}} \leq \alpha_{x,t}, \forall t \geq 0$$

where $\alpha_{x,t} = R\sqrt{d_x \log \frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$.

A.1 Proof of Theorem 1

Proof. Following the definition of total compensation, we have

$$\begin{aligned}
\mathrm{C}(T) &= \sum_{t=1}^{T} \mathbf{E}[c_{a_t,t}] = \sum_{t=1}^{T} \left(\max_i \hat{r}_{x,i,t} - \hat{r}_{x,a_t,t} \right) \\
&\leq \sum_{t=1}^{T} \left(\max_i \left(\hat{r}_{x,i,t} + CB_{x,t}(\mathbf{x}_i) \right) - \hat{r}_{x,a_t,t} \right) \\
&= \sum_{t=1}^{T} \left(\hat{r}_{x,a_t,t} + CB_{x,t}(\mathbf{x}_{a_t}) - \hat{r}_{x,a_t,t} \right) = \sum_{t=1}^{T} CB_{x,t}(\mathbf{x}_{a_t})
\end{aligned}$$

where the third step holds with probability at least $1-\delta$ and the fourth step is based on the UCB arm selection strategy.

So with probability at least $1-\delta$, we bound the total compensation as follows,

$$\begin{aligned}
\mathrm{C}(T) &\leq \sum_{t=1}^{T} CB_{x,t}(\mathbf{x}_{a_t}) \leq \sqrt{T \sum_{t=1}^{T} CB_{x,t}^2(\mathbf{x}_{a_t})} \\
&= \sqrt{T \sum_{t=1}^{T} \alpha_{x,t}^2 \|\mathbf{x}_a\|^2_{\mathbf{A}_{x,t}^{-1}}} \leq \sqrt{T \alpha_{x,T}^2 \sum_{t=1}^{T} \|\mathbf{x}_a\|^2_{\mathbf{A}_{x,t}^{-1}}} \\
&\leq \alpha_{x,T} \sqrt{T \sum_{t=1}^{T} \|\mathbf{x}_a\|^2_{\mathbf{A}_{x,t}^{-1}}}
\end{aligned}$$

According to Lemma 11 of [1], $\sum_{t=1}^{T} \|\mathbf{x}_a\|^2_{\mathbf{A}_{x,t}^{-1}} \leq d_x \log(\lambda + T/d_v)$.

Combining with $\alpha_{x,t} = R\sqrt{d_x \log \frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$, we can complete the proof. □

A.2 Proof of Theorem 2

Proof. We bound cumulative regret by

$$\begin{aligned}
\mathrm{R}(T) &= \sum_{t=1}^{T} \left(\mathbf{E}[r_{a_t^*}] - \mathbf{E}[r_{a_t}] \right) = \sum_{t=1}^{T} \left(\mathbf{v}_{a_t^*}^{\mathsf{T}} \boldsymbol{\theta}_v^* - \mathbf{v}_{a_t}^{\mathsf{T}} \boldsymbol{\theta}_v^* \right) \\
&\leq \sum_{t=1}^{T} \left(\mathbf{v}_{a_t^*}^{\mathsf{T}} \hat{\boldsymbol{\theta}}_{v,t} + 2CB_{v,t}(\mathbf{v}_{a_t^*}) - \mathbf{v}_{a_t}^{\mathsf{T}} \boldsymbol{\theta}_v^* \right) \\
&\leq \sum_{t=1}^{T} \left(\mathbf{v}_{a_t}^{\mathsf{T}} \hat{\boldsymbol{\theta}}_{v,t} + 2CB_{v,t}(\mathbf{v}_{a_t}) - \mathbf{v}_{a_t}^{\mathsf{T}} \boldsymbol{\theta}_v^* \right) \leq \sum_{t=1}^{T} 2CB_{v,t}(\mathbf{v}_{a_t})
\end{aligned}$$

The third step holds with probability at least $1-\delta$ according to the definition of confidence interval. The fourth step holds with probability at least $1-2\delta$ according to Lemma 2, where the users are incentivized to explore according to UCB strategy as shown in Eq (7). Taking a union bound, the above inequality holds with probability at least $1-3\delta$.

We continue bounding the cumulative regret with probability at least $1-3\delta$ as follows,

$$\begin{aligned}
\mathrm{R}(T) &\leq 2\sqrt{T \sum_{t=1}^{T} CB_{v,t}^2(\mathbf{v}_{a_t})} = 2\sqrt{T \sum_{t=1}^{T} \alpha_{v,t}^2 \|\mathbf{v}_a\|^2_{\mathbf{A}_{v,t}^{-1}}} \\
&\leq 2\alpha_{v,T} \sqrt{T \sum_{t=1}^{T} \|\mathbf{v}_a\|^2_{\mathbf{A}_{v,t}^{-1}}} \\
&\leq \left(2R\sqrt{d_v \log \frac{1+T/\lambda}{\delta}} + \sqrt{\lambda} \right) \sqrt{T d_v \log(\lambda + \frac{T}{d_v})}
\end{aligned}$$

where we finish the proof by combining $\sum_{t=1}^{T} \|\mathbf{v}_a\|^2_{\mathbf{A}_{v,t}^{-1}} \leq d_v \log(\lambda + T/d_v)$ and $\alpha_{v,t} = R\sqrt{d_v \log \frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$. □

A.3 Proof of Theorem 3

Proof. With probability at least $1-2\delta$, we have

$$\begin{aligned}
\mathrm{C}(T) &\leq \sum_{t=1}^{T} 4CB_{v,t}(\mathbf{v}_{a_t}) \leq 4\sqrt{T \sum_{t=1}^{T} CB_{v,t}^2(\mathbf{v}_{a_t})} \\
&= 4\sqrt{T \sum_{t=1}^{T} \alpha_{v,t}^2 \|\mathbf{v}_a\|^2_{\mathbf{A}_{v,t}^{-1}}} \leq 4\alpha_{v,T} \sqrt{T \sum_{t=1}^{T} \|\mathbf{v}_a\|^2_{\mathbf{A}_{v,t}^{-1}}} \\
&\leq \left(4R\sqrt{d_v \log \frac{1+T/\lambda}{\delta}} + \sqrt{\lambda} \right) \sqrt{T d_v \log(\lambda + \frac{T}{d_v})}
\end{aligned}$$

□

A.4 Proof of Theorem 4

Proof. Our proof relies on the following lemmas: □

Lemma 4 (Theorem 1 in Lattimore and Szepesvari [23]). *Assume $G_{x,T}$ is invertible for sufficiently large T. For all suboptimal $a \in \mathcal{A}$ it holds that*

$$\limsup_{T\to\infty} \log T \|\mathbf{x}_a - \mathbf{x}_1\|^2_{G^{-1}_{x,T}} \leq \frac{\Delta_a^2}{2}$$

Lemma 5 (Theorem 8 in Lattimore and Szepesvari [23]). *For any $\delta \in [1/T, 1)$, T sufficiently large and t_0 such that G_{x,t_0} is almost surely non-singular,*

$$\mathbb{P}\left(\exists t \geq 0, \mathbf{x}_a : |\hat{r}_{x,a,t} - \mathbf{E}[r_a]| \geq \sqrt{\|\mathbf{x}_a\|^2_{G^{-1}_{x,t}} f_{T,\delta}}\right) \leq \delta$$

where for some $c > 0$ universal constant $f_{T,\delta} = 2\left(1 + \frac{1}{\log(T)}\right)\log(1/\delta) + cd_x \log(d_x \log(T))$.

We first prove that after a fixed time point, with high probability pulling arm a once requires compensation at least $\Delta_a/3$. The proof idea is similar to the proof of Theorem 1 in [32]. We then derive the asymptotic compensation lower bound.

Based on Lemma 4, we can obtain the following inequality for all sub-optimal arms:

$$\limsup_{T\to\infty} \log(T) \|\mathbf{x}_a\|^2_{G^{-1}_{x,T}} \leq \frac{\Delta_a^2}{2} \tag{12}$$

which is also stated in the Corollary 2 in [23].

Let $N_a(T)$ be the number of times arm a is pulled in T rounds. Since the algorithm has $o(T)$ regret, we can find $T_1'(\delta)$ such that the best arm is pulled at least $T/2$ times with probability $1 - \delta/2$. Using the concentration bound we know there exists $T_1''(\delta)$ such that for $t > T_1''(\delta)$ with probability $1 - \delta/2$ the confidence interval of the best arm's reward estimation is smaller than $\Delta_2/3$ where Δ_2 is the reward gap between the best arm and second best arm. Let $T_1(\delta) = \max(T_1'(\delta), T_1''(\delta))$ and for all $t > T_1(\delta)$, with probability $1 - \delta$ we have $\hat{r}_{x,1,t} \geq \mathbf{E}[r_1] - \Delta_2/3$.

We argue a similar result for any suboptimal arm a. Based on Eq (12), there exists a $T_a(\delta)$ such that for any $t > T_a(\delta)$, with probability $1 - \delta$

$$\|\mathbf{x}_a\|^2_{G^{-1}_{x,t}} \leq \frac{\Delta_a^2}{2\log(T)} \leq \frac{\Delta_a^2}{9 f_{T,\delta}}$$

Combining with the concentration bound in Lemma 5, we have for any $t > T_a(\delta)$ with probability $1 - \delta$, $\hat{r}_{x,a,t} - \mathbf{E}[r_a] \leq \Delta_a/3$.

Let $T(\delta) = \max_i T_i(\delta)$ and we know that for any $t > T(\delta)$, the minimum required compensation to incentivize the user to pull arm a is

$$\max_i \hat{r}_{x,i,t} - \hat{r}_{x,a,t} \geq \hat{r}_{x,1,t} - \hat{r}_{x,a,t} \geq \mathbf{E}[r_1] - \frac{\Delta_2}{3} - \mathbf{E}[r_a] - \frac{\Delta_a}{3} \geq \frac{\Delta_a}{3} \tag{13}$$

with probability at least $1 - \delta$.

We then use the optimization problem in Eq (11) to obtain the compensation lower bound, where the optimization minimizes the total compensation and satisfies the consistent constraints that the gaps of all suboptimal arms are identified with high confidence. With probability at least $1 - \delta$, for sufficiently large T the total compensation is

$$C(T) \geq \sum_{a\in\mathcal{A}} \mathbf{E}[N_a(T)] \frac{\Delta_a}{3}$$

$\alpha_{\mathbf{x}_a} = \mathbf{E}[N_a(T)]/\log(T)$ is asymptotically feasible for large T because it satisfies

$$\limsup_{T\to\infty} \|\mathbf{x}_a\|^2_{H^{-1}_{x,T}} = \limsup_{T\to\infty} \log(T) \|\mathbf{x}_a\|^2_{G^{-1}_{x,T}} \leq \frac{\Delta_a^2}{2}$$

where $G_{x,T} = \log(T) H_{x,T}$. Thus for any $\epsilon > 0$, $\|\mathbf{x}_a\|^2_{H^{-1}_{x,T}} \leq \Delta_a^2/2 + \epsilon$ and

$$C(T) \geq \sum_{a\in\mathcal{A}} \mathbf{E}[N_a(T)] \frac{\Delta_a}{3} \geq c_{x,\epsilon}(\mathcal{A}, \boldsymbol{\theta}^*) \log(T) \tag{14}$$

where $c_{x,\epsilon}(\mathcal{A}, \boldsymbol{\theta}^*)$ is the the optimal value of the optimization problem in Eq (11) by replacing $\Delta_a^2/2$ with $\Delta_a^2/2 + \epsilon$. Since $\inf_{\epsilon>0} c_{x,\epsilon}(\mathcal{A}, \boldsymbol{\theta}^*) = c_x(\mathcal{A}, \boldsymbol{\theta}^*)$ and $T \to \infty$ we have the total compensation as

$$\Omega\left(c_x(\mathcal{A}, \boldsymbol{\theta}^*) \log(T)\right)$$

and then built Theorem 4 based on this result and the known lower regret bound of linear bandits in [23]. Theorem 4 can also recover the compensation lower bound in non-contextual setting in [32].

REFERENCES

[1] Yasin Abbasi-yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved Algorithms for Linear Stochastic Bandits. In *NIPS*. 2312–2320.

[2] Marc Abeille and Alessandro Lazaric. 2017. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*. PMLR, 176–184.

[3] Priyank Agrawal and Theja Tulabandhula. 2020. Incentivising Exploration and Recommendations for Contextual Bandits with Payments. In *Multi-Agent Systems and Agreement Technologies*. Springer, 159–170.

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

[5] Peter Auer. 2002. Using Confidence Bounds for Exploitation-Exploration Tradeoffs. *Journal of Machine Learning Research* 3 (2002), 397–422.

[6] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.

[7] Bin Bi, Milad Shokouhi, Michal Kosinski, and Thore Graepel. 2013. Inferring the demographics of search users: Social data meets search queries. In *Proceedings of the 22nd international conference on World Wide Web*. 131–140.

[8] Nicolo Cesa-Bianchi, Claudio Gentile, and Giovanni Zappella. 2013. A gang of bandits. In *Advances in Neural Information Processing Systems*. 737–745.

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

[10] Bangrui Chen, Peter Frazier, and David Kempe. 2018. Incentivizing exploration by heterogeneous users. In *Conference On Learning Theory*. PMLR, 798–818.

[11] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 456–464.

[12] Minmin Chen, Can Xu, Vince Gatto, Devanshu Jain, Aviral Kumar, and Ed Chi. 2022. Off-policy actor-critic for recommender systems. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 338–349.

[13] Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. 2011. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. 208–214.

[14] Peter Frazier, David Kempe, Jon Kleinberg, and Robert Kleinberg. 2014. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*. ACM, 5–22.

[15] Dalin Guo, Sofia Ira Ktena, Pranay Kumar Myana, Ferenc Huszar, Wenzhe Shi, Alykhan Tejani, Michael Kneier, and Sourav Das. 2020. Deep bayesian bandits: Exploring in online personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*. 456–461.

[16] Christoph Hirnschall, Adish Singla, Sebastian Tschiatschek, and Andreas Krause. 2018. Learning user preferences to incentivize exploration in the sharing economy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[17] Nicole Immorlica, Jieming Mao, Aleksandrs Slivkins, and Zhiwei Steven Wu. 2018. Incentivizing Exploration with Selective Data Disclosure. *arXiv preprint arXiv:1811.06026* (2018).

[18] Olivier Jeunen and Bart Goethals. 2021. Pessimistic reward models for off-policy learning in recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 63–74.

[19] Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh Pai, Aaron Roth, Rakesh Vohra, and Zhiwei Steven Wu. 2017. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. 369–386.

[20] Ilan Kremer, Yishay Mansour, and Motty Perry. 2014. Implementing the "wisdom of the crowd". *Journal of Political Economy* 122, 5 (2014), 988–1012.

[21] Tze Leung Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6, 1 (1985), 4–22.

[22] Sahin Lale, Kamyar Azizzadenesheli, Anima Anandkumar, and Babak Hassibi. 2019. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490* (2019).

[23] Tor Lattimore and Csaba Szepesvari. 2017. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*. PMLR, 728–737.

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

[25] Zhiyuan Liu, Huazheng Wang, Fan Shen, Kai Liu, and Lijun Chen. 2020. Incentivized Exploration for Multi-Armed Bandits under Reward Drift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4981–4988.

[26] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Ji Yang, Minmin Chen, Jiaxi Tang, Lichan Hong, and Ed H Chi. 2020. Off-policy learning in two-stage recommender systems. In *Proceedings of The Web Conference 2020*. 463–473.

[27] Yishay Mansour, Aleksandrs Slivkins, and Vasilis Syrgkanis. 2015. Bayesian incentive-compatible bandit exploration. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. ACM, 565–582.

[28] Mark Sellke and Aleksandrs Slivkins. 2020. Sample complexity of incentivized exploration. *arXiv preprint arXiv:2002.00558* (2020).

[29] Max Simchowitz and Aleksandrs Slivkins. 2021. Exploration and incentives in reinforcement learning. *arXiv preprint arXiv:2103.00360* (2021).

[30] Aleksandrs Slivkins. 2017. Incentivizing exploration via information asymmetry. *XRDS: Crossroads, The ACM Magazine for Students* 24, 1 (2017), 38–41.

[31] Huazheng Wang, Qingyun Wu, and Hongning Wang. 2017. Factorization bandits for interactive recommendation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[32] Siwei Wang and Longbo Huang. 2018. Multi-armed Bandits with Compensation. In *NeurIPS*.

[33] Ingmar Weber and Carlos Castillo. 2010. The demographics of web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 523–530.

[34] Qingyun Wu, Huazheng Wang, Quanquan Gu, and Hongning Wang. 2016. Contextual bandits in a collaborative environment. In *SIGIR 2016*. ACM, 529–538.