Incentivizing Exploration in Linear Bandits under Information Gap

Huazheng Wang [1] **Haifeng Xu** [1] **Chuanhao Li** [1] **Zhiyuan Liu** [2] **Hongning Wang** [1]

Abstract

We study the problem of incentivizing exploration for myopic users in linear bandits, where the users tend to exploit arm with the highest predicted reward instead of exploring. In order to maximize the long-term reward, the system offers compensation to incentivize the users to pull the exploratory arms, with the goal of balancing the trade-off among exploitation, exploration and compensation. We consider a new and practically motivated setting where the context features observed by the user are more *informative* than those used by the system, e.g., features based on users' private information are not accessible by the system. We propose a new method to incentivize exploration under such information gap, and prove that the method achieves both sublinear regret and sublinear compensation. We theoretical and empirically analyze the added compensation due to the information gap, compared with the case that the system has access to the same context features as the user, i.e., without information gap. We also provide a compensation lower bound of our problem.

1. Introduction

The traditional multi-armed bandit (MAB) (Lai & Robbins, 1985) research studies the single-party setting, where the system has a full control over which arm to pull and can trade off exploitation and exploration for long-term optimality. However, in many real-world applications, such as recommender systems and e-commerce platforms, one often faces a *two-party* game between the system and its users, and the two parties have *different* interests. The system aims at maximizing the long-term reward by recommending exploratory arms; but it cannot directly pull the arm to receive the reward. On the other hand, the arm can only be pulled by the *myopic* users, who seek to maximize their short-term utilities. This leads to the problem of under-exploration and selection bias: the best arm may remain unexplored forever if it appears sub-optimal initially. To align the two parties' interest, the system should offer compensations to the users so that the users are motivated to pull the exploratory arm and maximize the long-term reward. This problem is known as *incentivizied exploration* (Kremer et al., 2014; Frazier et al., 2014; Mansour et al., 2015).

Incentivized exploration has been studied in the MAB setting, where the system's goal is to balance the trade-off among *exploration, exploitation and compensation*, i.e., minimizing total payments while maximizing cumulative rewards (Frazier et al., 2014; Hirnschall et al., 2018; Wang & Huang, 2018). Previous solutions assume both the users and the system have access to the same information and both parties maintain the same reward estimation. This assumption is necessary for the system to compute the compensation based on the users' estimated reward difference between the currently best arm and the exploratory arm. Under MAB setting, this assumption naturally holds because both parties observe the same reward feedback and estimate with averaged reward. However, under the contextual bandit setting (Auer, 2002; Li et al., 2010; Abbasi-yadkori et al., 2011), both parties observe the same rewards but may access different context features. This would lead to different reward estimation and convergence. For example, the users could access the features related to their own private information, which are not accessible by the system. An extreme case in a finite arm setting is that the system may only observe the indices of the arm (which degenerates to the non-contextual MAB), while the users employ informative feature representations of the arms. This representation asymmetry is what we call the *information gap* between the two parties. This gap leads to different reward estimation between the two parties and brings in the new challenges to incentivized exploration. For example, it is even unclear which arm is currently the best on the user side.

In this paper, we study the problem of incentivized exploration in linear contextual bandits under information gap. We proposed an algorithm that incentivizes the user to explore according the Linear UCB strategy (Li et al., 2010; Abbasi-yadkori et al., 2011). The key idea to conquer information gap is that although the system suffers from an information disadvantage and cannot compute the minimum compensation precisely, offering a larger amount of compen-

[1]University of Virginia [2]University of Colorado, Boulder. Correspondence to: Huazheng Wang <hw7ww@virginia.edu>.

Preprint.

sation 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 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 with information gap, and the cost of the information gap is realized by the extra compensation that dominated by d_v. We also proved the compensation lower bound of incentivized exploration in linear bandits, which recovers the result of compensation lower bound in non-contextual bandits reported in Wang & Huang (2018). Our simulation-based empirical studies also validate the effectiveness and cost-efficiency of the proposed algorithm.

2. Problem Definition

Notations and assumptions. We study the problem under a linear bandit setting, where a myopic user sequentially interacts with the system for T rounds. At each round t, the user observes compensation offered by the system, and pulls an arm a_t from a given arm set $\mathcal{A}_t$. Both the system and the user observe the resulting reward $r_{a_t,t}$ and update their estimations accordingly.

In a contextual bandit setting, each arm a is associated with a context feature vector. In our problem, for arm $a \in \mathcal{A}_t$, the system observes a feature $\mathbf{v}_a$ from a d_v-dimensional subspace and the users observes a feature $\mathbf{x}_a$ from a d_x-dimensional subspace. Without loss of generality, we will assume $\mathbf{x}_a \in \mathbb{R}^{d_x}$ and $\mathbf{v}_a \in \mathbb{R}^{d_v}$ — if not, the standard PCA technique can be used to reduce the feature dimensions to d_x, d_v (Lale et al., 2019). Essentially we consider the features span the whole vector space respectively, which means there is no redundant feature on both sides and the dimensionality cannot be further reduced.

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

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

where $d_v \geq d_x$.

The assumption on $d_v \geq d_x$, i.e., features used by the user belong to a lower dimension space is motivated by many real-world scenarios: for example, users can construct features related to their private information (e.g., age, gender, income or health). 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 system has no choice but to set the context features as K-dimension basis vectors, whereas the user can observe low-dimensional informative feature representations of the arms with $d_x \ll K$.

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

Examples of information gap. We discussed an extreme case in the introduction where the system is not allowed to access any arm feature except the indices of arms. In this case, the context vectors used by the system are the K-dimension one-hot vectors, while the user may observe and employ d_x-dimension feature representations of the same arms. The information gap ($K > d_x$) is encoded in the transformation matrix P. Now let us consider a less extreme example. Some features could be the combination of both the user's information and item's property, e.g., joint of user's income and the item's price, or joint of user's gender and the item's category. This is a typical way to construct features in the practical recommender systems. The users can employ these informative features and enjoy faster convergence. The system will suffer if it cannot access users' private information. In this example, the transformation matrix P contains the private information hidden from the system.

Note that having access to more features is not equivalent to have more informative representations. Another practical example is that the context vectors used by the system may include many useless or redundant features, where the corresponding weights in the model parameter $\boldsymbol{\theta}_v^*$ are zeros, i.e., a sparse regression setting. The information gap is captured by the transformation matrix P where the corresponding columns are zero vectors. In this example, the system's features are clearly less informative, i.e., $d_v > d_x$, because of the useless features.

Reward mapping. Following a linear bandit setting, the expected reward of arm a is determined by the inner product between the context features and an unknown bandit parameter. From the user's perspective, we have

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

where $\boldsymbol{\theta}_x^*$ is the unknown model parameter on the user side.

Based on 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 follow:

$$\mathbf{E}[r_a] = \mathbf{x}_a^\top \boldsymbol{\theta}_x^* = \mathbf{v}_a^\top \boldsymbol{\theta}_v^* \tag{2}$$

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

$$r_{a_t,t} = \mathbf{E}[r_{a_t}] + \eta_t \tag{3}$$

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 bounds the expected reward in the range of $[-1, 1]$ and simplifies the analysis. 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 user 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 $\hat{r}_{x,a,t} = \mathbf{x}_a^\mathsf{T}\hat{\boldsymbol{\theta}}_{x,t}$ and $\hat{r}_{v,a,t} = \mathbf{v}_a^\mathsf{T}\hat{\boldsymbol{\theta}}_{v,t}$. Note that since both feature sets $\{\mathbf{x}_a\}$ and $\{\mathbf{v}_a\}$ can generate the same rewards, $d_v \geq d_x$ suggests that features in $\{\mathbf{x}_a\}$ can better characterize the reward mapping, thus more *informative*. The less informative features lead to a slower convergence of the parameter estimation and a wider confidence interval of the reward estimation. Such an information gap brings in new challenges of incentivized exploration.

Objective. The users and the system have different objectives in this sequential decision making problem: the user aims to maximize his/her short-term instantaneous reward, while the system aims to maximize the long-term cumulative reward. At each round t, without any incentive, the myopic user 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 the 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 user to explore. Specifically, the system offers compensation $c_{a,t}$ for pulling arm a. Given the incentives, the users maximize 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 regret is defined as

$$R(T) = \sum_{t=1}^{T} \big(\mathbf{E}[r_{a_t^*}] - \mathbf{E}[r_{a_t}]\big) \tag{4}$$

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

$$C(T) = \sum_{t=1}^{T} \mathbf{E}[c_{a_t,t}] \tag{5}$$

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

Algorithm 1 Incentivized LinUCB under Information Gap

Inputs: λ, δ
Initialize: $\mathbf{A}_x = \lambda\mathbf{I}_{d_x}, \mathbf{A}_v = \lambda\mathbf{I}_{d_v}, \mathbf{b}_x = 0, \mathbf{b}_v = 0$
for $t = 1$ to T **do**
  System and user observe context vectors $\{\mathbf{x}_a\}_{a\in\mathcal{A}_t}$ and $\{\mathbf{v}_a\}_{a\in\mathcal{A}_t}$ respectively
  System calculates compensation $c_{a,t}$ for arm a according to Eq (7)
  User pulls arm $a_t = \arg\max_{a\in\mathcal{A}} \hat{r}_{x,a,t} + c_{a,t}$
  System and user observe reward r_{a_t}
  // Update on the system side:
  $\mathbf{A}_{v,t+1} \leftarrow \mathbf{A}_{v,t} + \mathbf{v}_{a_t}\mathbf{v}_{a_t}^\mathsf{T}$, $\mathbf{b}_{v,t+1} \leftarrow \mathbf{b}_{v,t} + \mathbf{v}_{a_t} r_{a_t}$
  $\hat{\boldsymbol{\theta}}_{v,t+1} \leftarrow {\mathbf{A}_{v,t+1}}^{-1}\mathbf{b}_{v,t+1}$
  // Update on the user side:
  $\mathbf{A}_{x,t+1} \leftarrow \mathbf{A}_{x,t} + \mathbf{x}_{a_t}\mathbf{x}_{a_t}^\mathsf{T}$, $\mathbf{b}_{x,t+1} \leftarrow \mathbf{b}_{x,t} + \mathbf{x}_{a_t} r_{a_t}$
  $\hat{\boldsymbol{\theta}}_{x,t+1} \leftarrow \mathbf{A}_{x,t+1}^{-1}\mathbf{b}_{x,t+1}$
end for

3. Method

We present our solution on incentivized exploration under information gap when the system explores according to the Linear UCB strategy (Li et al., 2010; Chu et al., 2011; Abbasi-yadkori et al., 2011). Then we show that the solution can be easily adopted to the simpler problem setting of incentivized exploration without the information gap.

3.1. Incentivized exploration under information gap

We present Algorithm 1 to show how the system incentivizes the myopic user to follow the desired exploration strategy under information gap. At each round, the system and the user observe context features $\{\mathbf{x}_a\}$ and $\{\mathbf{v}_a\}$ respectively for the same arm set $\mathcal{A}_t$. The system needs to motivate the user to explore arm a_t according to LinUCB strategy based on its current parameter estimation $\hat{\boldsymbol{\theta}}_{v,t}$. To incentivize the user to pull arm a_t, the system offers compensation $c_{a_t,t}$ according to Eq (7). 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 $c_{a,t}$ In Lemma 3 we guarantee that the user will pull the system desired arm a_t. Both the system and the user then observe reward feedback r_{a_t}, and update their parameters using ridge regression accordingly.

Denote $CB_{x,t}(\mathbf{x}_a)$ as the width of the user's 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

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

The value of $\alpha_{x,t}$ is the upper bound of the width of confidence ellipsoid and is set according to the following lemma.

Lemma 1 (Theorem 2 of (Abbasi-yadkori et al., 2011)). *With probability at least $1-\delta$, the parameter $\boldsymbol{\theta}_x^*$ lies in the confidence ellipsoid of $\hat{\boldsymbol{\theta}}_{x,t}$ satisfying*

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

for all $t \geq 0$.

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 learn and predict rewards. This prevents us from computing minimum required compensation and makes the problem non-trivial. We have to carefully determine the compensation: a larger amount of incentives 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 by the system's reward estimation is no smaller than the confidence interval by the user's estimate. This lemma guarantees in Algorithm 1 the system provides sufficient incentives to the user to pull the arms according to an upper confidence bound type exploration strategy.

Lemma 2. *Consider two least square estimators (ridge regression) that estimate the model parameters with the same reward observations but different features satisfying Assumption 1. For all $t \geq 0$ and all arm $a \in \mathcal{A}_t$, we have*

$$CB_{v,t}(\mathbf{v}_a) \geq CB_{x,t}(\mathbf{x}_a), \tag{6}$$

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

Proof Sketch. Since $CB_{v,t}(\mathbf{v}_a) = \alpha_{v,t}\|\mathbf{v}_a\|_{\mathbf{A}_{v,t}^{-1}}$ and $CB_{x,t}(\mathbf{x}_a) = \alpha_{x,t}\|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}}$, we can prove $\|\mathbf{v}_a\|_{\mathbf{A}_{v,t}^{-1}} \geq \|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}}$ and $\alpha_t^v \geq \alpha_t^x$ separately. It is obvious that $\alpha_t^v \geq \alpha_t^x$ because $d_v \geq d_x$. Substitute $\mathbf{x}_a = P\mathbf{v}_a$ and we can prove that $\mathbf{A}_{v,t}^{-1} - P^{\mathsf{T}}\left(P\mathbf{A}_{v,t}P^{\mathsf{T}}\right)^{-1}P$ is a positive semi-definite matrix, which leads to $\|\mathbf{v}_a\|_{\mathbf{A}_{v,t}^{-1}} \geq \|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}}$. □

The intuition behind this lemma is straightforward. The confidence interval characterizes the uncertainty of reward prediction. Since the estimator on the users side uses more informative features, its parameter estimation converges faster and its confidence interval is smaller than that maintained on the system side.

Based on Lemma 2, we have the following lemma,

Lemma 3. *For all $t \geq 0$, with probability at least $1-2\delta$, the users are incentivized to pull the desired arm with compensation*

$$c_{a_t,t} = 4CB_{v,t}(\mathbf{v}_{a_t}) \tag{7}$$

to arm

$$a_t = \arg\max_a\left(\mathbf{v}_a^{\mathsf{T}}\hat{\boldsymbol{\theta}}_{v,t} + 2CB_{v,t}(\mathbf{v}_a)\right), \tag{8}$$

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

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., $\hat{r}_{x,g,t} \leq \hat{r}_{x,a_t,t} + c_{a_t,t}$.

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

$$\begin{aligned}|\hat{r}_{x,a,t} - \hat{r}_{v,a,t}| &\leq |\hat{r}_{x,a,t} - \mathbf{E}[r_a]| + |\mathbf{E}[r_a] - \hat{r}_{v,a,t}| \\ &\leq CB_{x,t}(\mathbf{x}_a) + CB_{v,t}(\mathbf{v}_a)\end{aligned} \tag{9}$$

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

$$\begin{aligned}\hat{r}_{x,g,t} &\leq \hat{r}_{v,g,t} + CB_{x,t}(\mathbf{x}_g) + CB_{v,t}(\mathbf{v}_g) \\ &\leq \hat{r}_{v,g,t} + 2CB_{v,t}(\mathbf{v}_g) \\ &\leq \hat{r}_{v,a_t,t} + 2CB_{v,t}(\mathbf{v}_{a_t}) \\ &\leq \hat{r}_{x,a_t,t} + CB_{x,t}(\mathbf{v}_{a_t}) + CB_{v,t}(\mathbf{v}_{a_t}) + 2CB_{v,t}(\mathbf{v}_{a_t}) \\ &\leq \hat{r}_{x,a_t,t} + 4CB_{v,t}(\mathbf{v}_{a_t})\end{aligned} \tag{10}$$

The first and fourth steps are based on Eq (9). The second and last steps are based on Lemma 2. The third inequality is based on the UCB strategy in Eq (8). □

It is worth noting that the system follows a more optimistic arm selection strategy in Eq (8) using a confidence interval

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 observe context vectors $\{\mathbf{x}_a\}_{a \in \mathcal{A}_t}$
  System calculate compensation $c_{a,t}$ for arm a according to Eq (11)
  User pulls arm $a_t = \arg\max_{a \in \mathcal{A}} \hat{r}_{x,a,t} + c_{a,t}$
  System and user observe reward r_{a_t}
  $\mathbf{A}_{x,t+1} \leftarrow \mathbf{A}_{x,t} + \mathbf{x}_{a_t}\mathbf{x}_{a_t}^{\mathsf{T}}$, $\mathbf{b}_{x,t+1} \leftarrow \mathbf{b}_{x,t} + \mathbf{x}_{a_t} r_{a_t}$
  $\hat{\boldsymbol{\theta}}_{x,t+1} \leftarrow \mathbf{A}_{x,t+1}^{-1}\mathbf{b}_{x,t+1}$
end for

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 (10) suggested. 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 regret and compensation upper bound of Algorithm 1 in Section 4.

3.2. Incentivized exploration without information gap

Our solution can be easily adopted to solve the incentivized exploration problem of without information gap. In Algorithm 2, we show how the system incentivizes the myopic user to follow the desired exploration strategy in this simpler setting.

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

$$c_{a_t,t} = \max_i \hat{r}_{x,i,t} - \hat{r}_{x,a_t,t} \tag{11}$$

to arm $a_t = \arg\max_a \left(\mathbf{x}_a^{\mathsf{T}}\hat{\boldsymbol{\theta}}_{x,t} + CB_{x,t}(\mathbf{x}_a)\right)$. 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 that the user is incentivized to pull arms according to LinUCB, its regret is the same as LinUCB's in the order of $O(d_x\sqrt{T}\log T)$ (see Theorem 3 of (Abbasi-yadkori et al., 2011)). Its compensation upper bound is stated below.

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

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

Proof Sketch. We can first show that with a high probability the compensation at round t is upper bounded by the confidence interval, i.e., $c_{a_t,t} \leq CB_{x,t}(\mathbf{x}_{a_t})$. Then the total compensation can be upper bounded by $\sum_t CB_{x,t}(\mathbf{x}_{a_t})$, which can be bounded using Lemma 11 of Abbasi-yadkori et al. (2011). □

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 predictions, the system can offer the minimum required compensation as shown in Eq (11) to incentivize exploration. With information gap, compensate by Eq (7) 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 (11); otherwise, the system can only estimate the reward difference with a high probability (as shown in Lemma 3). 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 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.

4. Analysis

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

4.1. Regret and compensation upper bound

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

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

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 3 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*

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

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 showed that our proposed algorithm can incentivize exploration under information gap and achieve 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 $O((d_v-d_x)\sqrt{T\log T})$. 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, exactly where the information gap comes from.

4.2. 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 of incentivized exploration reported in non-contextual bandits in (Wang & Huang, 2018).

Let $G_{x,T} = \mathbb{E}\left[\sum_{t=1}^{T} \mathbf{x}_{a_t}\mathbf{x}_{a_t}^\mathsf{T}\right]$. Without loss of generality assume arm 1 is the best arm and $\Delta_a = \mathbf{E}[r_1] - \mathbf{E}[r_a] = (\mathbf{x}_1 - \mathbf{x}_a)^\mathsf{T}\boldsymbol{\theta}^*$ 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 $0<p\le 1$. 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

$$\begin{aligned} c_x(\mathcal{A},\boldsymbol{\theta}^*) &= \inf_{\alpha\ge 0}\sum_{\mathbf{x}_a}\alpha_{\mathbf{x}_a}\frac{\Delta_a}{3} \\ \text{s.t. } &\|\mathbf{x}_a\|^2_{H^{-1}_{x,T}} \le \frac{\Delta_a^2}{2}, \forall \mathbf{x}_a \text{ with } \Delta_a > 0 \end{aligned} \tag{12}$$

where $H_{x,T} = \sum_{\mathbf{x}_a}\alpha_{\mathbf{x}_a}\mathbf{x}_{a_t}\mathbf{x}_{a_t}^\mathsf{T}$.

Our proof relies on the following lemmas:

Lemma 4 (Theorem 1 in Lattimore & Szepesvari (2017))**.** *Assume $G_{x,T}$ is invertible for sufficiently large T. For all suboptimal $a\in\mathcal{A}$ it holds that*

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

Lemma 5 (Theorem 8 in Lattimore & Szepesvari (2017))**.** *For any $\delta\in[1/T,1)$, T sufficiently large and t_0 such that G_{t_0} is almost surely non-singular,*

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

where for some $c>0$ universal constant

$$f_{T,\delta} = 2\left(1 + \frac{1}{\log(T)}\right)\log(1/\delta) + cd_x\log(d_x\log(T))$$

Proof Sketch. Suppose an algorithm is consistent with regret $o(T^p)$, Lemma 4 suggests that the algorithm must collect a sufficient number of samples such that the width of the confidence interval is small enough to identify the suboptimal arms. Since the algorithm has $o(T)$ regret, we can find t_1 such that the best arm is pulled at least $T/2$ times; and because of the concentration result in Lemma 5, its confidence interval is smaller than $\Delta_2/3$ where Δ_2 is the reward gap between the best arm and second best arm. This means for $t>t_1$ we have $\hat{r}_{x,1,t} \ge \mathbf{E}[r_1] - \Delta_2/3$ with a high probability.

For any other arm a, from Lemma 4 and the concentration bound we can show that it will also be pulled enough times such that its confidence interval is smaller than $\Delta_a/3$ with a high probability after a fixed round t_a. Therefore, for $t>t_a$ we have $\hat{r}_{x,a,t} \le \mathbf{E}[r_a] + \Delta_a/3$. Combining the two inequalities we know that after a fixed time point, the minimum required compensation to incentivize the user to pull arm a is $\hat{r}_{x,1,t} - \hat{r}_{x,a,t} \ge \Delta_a/3$. We then use the optimization problem in Eq (12) 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. □

Next, we construct an example to illustrate our 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 (Wang & Huang, 2018). 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 is in the order of $\Omega(\log(T))$, because it is a gap-dependent bound. 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

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

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

where $H_{v,T} = \sum_{\mathbf{v}_a} \alpha_{\mathbf{v}_a} \mathbf{v}_{a_t} \mathbf{v}_{a_t}^{\mathsf{T}}$.

The proof of compensation lower bound under information gap mostly follows Theorem 4 by simply replacing the user's feature $\mathbf{x}_a$ with the system's feature $\mathbf{v}_a$. The main difference is that when applying the concentration bound in Lemma 5 to derive the minimum required compensation, we still use $\mathbf{x}_a$ because the minimum amount is based on the *user's* estimated reward difference between the currently best arm and the exploratory arm. However, we notice that $\mathbf{x}_a$ or d_x does not directly appear in this lower bound. The impact of $\mathbf{x}_a$ being in a lower-dimensional space is that we have a faster concentration bound to have the confidence interval smaller than $\Delta_a/3$ at an earlier time point. Since we consider $T \to \infty$, this does not change the order of the bound and the final result is dominated by $\mathbf{v}_a$.

Considering a similar example of K-armed bandit setting where $K = d_v$, 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.

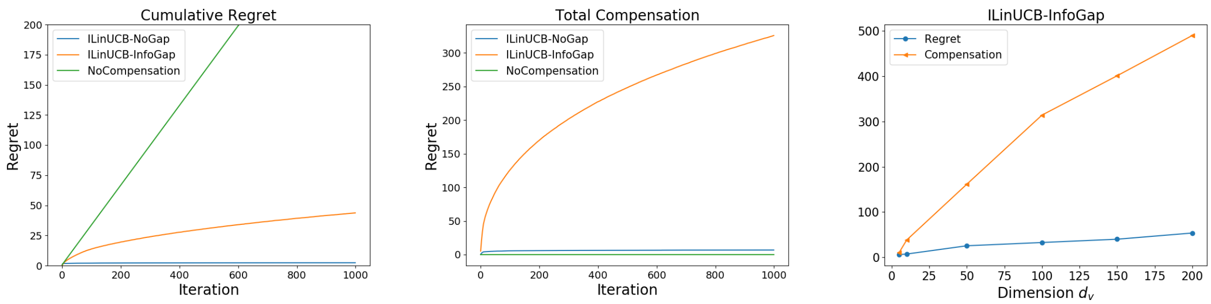
5. Experiments

We use simulation-based experiments to verify the effectiveness of our proposed incentivized exploration solution. In our simulations, we generate a size-K arm pool $\mathcal{A}$, in which each arm 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 arms were presented to all the algorithms, but the system and the user observe different features respectively. After the user pulls an arm a_t, both the user and the system observe its reward following Eq (3). We set d_x to 5, d_v to 100, the standard derivation σ 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 always pulls the current best arm. We set the probability $\delta = 0.01$ and regularization coefficient $\lambda = 0.1$ for all the algorithms.

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. 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 R^{d_v} space because of the information gap.

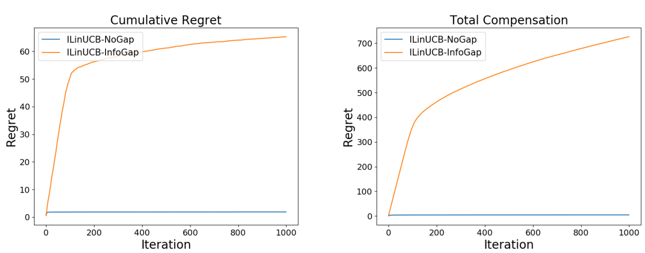
We notice that the total compensation of ILinUCB-InfoGap in Figure 1(b) is sublinear and keeps increasing. The algorithm has to always compensate due to the information gap as we discussed before. ILinUCB-NoGap, however, rarely compensates in the later stage. This is because when



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

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

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 increases linearly with d_v, which confirms our theoretical upper bound.



(a) Regret (b) Compensation

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

In Figure 2, we simulate a K-armed bandit setting where only the indices of the arms 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 (Auer et al., 2002) and can be viewed as a more optimism version of the Incentivized UCB algorithm in (Wang & Huang, 2018) with a wider confidence interval in consideration of 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. Related Work

The incentivized exploration in multi-armed bandits has been studied since (Kremer et al., 2014; Frazier et al., 2014). See Slivkins (2017) for an overview. One line of the studies (Kremer et al., 2014; Mansour et al., 2015; Immorlica et al., 2018; Sellke & Slivkins, 2020) assume the system has information advantage on observing the full arm-pulling history while users do not. The system leverages the information asymmetry to recommend exploratory arms as long as the users do not have a better choice from their perspective. Another line (Frazier et al., 2014; Chen et al., 2018; Wang & Huang, 2018) considered the setting where the arm-pulling history is publicly available to both system and users and the system offers compensations to an arm for incentivized exploration. Our setting follows this line of research.

Incentivized learning with monetary payments was first studied in (Frazier et al., 2014) in a Bayesian setting with discounted regret and compensation. Chen et al. (2018) studied a heterogeneous users setting, where user diversity led to their solution with constant compensation. Agrawal & Tulabandhula (2020) considered heterogeneous contexts in a contextual bandit setting. In (Wang & Huang, 2018), 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. (2020) considered the reward feedback is biased because of the compensation.Kannan et al. (2017) considered incentivized exploration for fair recommendation. Our setting is mostly similar to Wang & Huang (2018), 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. 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' and the precise amount of payment is harder to compute.

There are several recent works study low-rank bandits, which however are intrinsically different from ours. For example, Lale et al. (2019) consider the contexts are sampled from a low-dimensional subspace and propose a PCA-based solution to reduce the dimension. Yang et al. (2021) study multi-task linear bandits with a shared low-rank structure. These methods assume the learning problem is generated from a low-rank structure but presented in a high-dimensional space. But in our setting, the system's observed contexts are already sampled from a high-dimensional compact space, whose dimension cannot be further reduced. The information gap in representation asymmetry is a unique problem in this two-party game setting.

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 to 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. As our future work, we plan to study how to incentivize the users following other types of exploration strategy such as Thompson Sampling (Chapelle & Li, 2011; Agrawal & Goyal, 2013; Abeille & Lazaric, 2017). 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.

References

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *NIPS*, pp. 2312–2320. 2011.

Abeille, M. and Lazaric, A. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pp. 176–184. PMLR, 2017.

Agrawal, P. and Tulabandhula, T. Incentivising exploration and recommendations for contextual bandits with payments. In *Multi-Agent Systems and Agreement Technologies*, pp. 159–170. Springer, 2020.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pp. 127–135. PMLR, 2013.

Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

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

Chapelle, O. and Li, L. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pp. 2249–2257, 2011.

Chen, B., Frazier, P., and Kempe, D. Incentivizing exploration by heterogeneous users. In *Conference On Learning Theory*, pp. 798–818. PMLR, 2018.

Chu, W., Li, L., Reyzin, L., and Schapire, R. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

Frazier, P., Kempe, D., Kleinberg, J., and Kleinberg, R. Incentivizing exploration. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 5–22. ACM, 2014.

Hirnschall, C., Singla, A., Tschiatschek, S., and Krause, A. Learning user preferences to incentivize exploration in the sharing economy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Immorlica, N., Mao, J., Slivkins, A., and Wu, Z. S. Incentivizing exploration with selective data disclosure. *arXiv preprint arXiv:1811.06026*, 2018.

Kannan, S., Kearns, M., Morgenstern, J., Pai, M., Roth, A., Vohra, R., and Wu, Z. S. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pp. 369–386, 2017.

Kremer, I., Mansour, Y., and Perry, M. Implementing the "wisdom of the crowd". *Journal of Political Economy*, 122(5):988–1012, 2014.

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

Lale, S., Azizzadenesheli, K., Anandkumar, A., and Hassibi, B. Stochastic linear bandits with hidden low rank structure. *arXiv preprint arXiv:1901.09490*, 2019.

Lattimore, T. and Szepesvari, C. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pp. 728–737. PMLR, 2017.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670. ACM, 2010.

Liu, Z., Wang, H., Shen, F., Liu, K., and Chen, L. Incentivized exploration for multi-armed bandits under reward drift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4981–4988, 2020.

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

Sellke, M. and Slivkins, A. Sample complexity of incentivized exploration. *arXiv preprint arXiv:2002.00558*, 2020.

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

Wang, S. and Huang, L. Multi-armed bandits with compensation. In *NeurIPS*, 2018.

Yang, J., Hu, W., Lee, J. D., and Du, S. S. Impact of representation learning in linear bandits. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=edJ_HipawCa.

A. Proof Details

Proof of Lemma 2. According to the definition of confidence interval, $CB_{v,t}(\mathbf{v}_a) = \alpha_{v,t}\|\mathbf{v}_a\|_{\mathbf{A}_{v,t}^{-1}}$ and $CB_{x,t}(\mathbf{x}_a) = \alpha_{x,t}\|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}}$. We first prove that $\|\mathbf{v}_a\|_{\mathbf{A}_{v,t}^{-1}} \geq \|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}}$. 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

$$
\begin{aligned}
\|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}} &= \sqrt{\mathbf{x}_a^\mathsf{T}\mathbf{A}_{x,t}^{-1}\mathbf{x}_a} \\
&= \sqrt{\mathbf{v}_a^\mathsf{T}P^\mathsf{T}\left((P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}) + \lambda\mathbf{I}\right)^{-1}P\mathbf{v}_a}
\end{aligned}
$$

We prove

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

by showing $\mathbf{A}_{v,t}^{-1} - P^\mathsf{T}\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right)^{-1}P$ is a positive semi-definite matrix based on the property of Schur complement.

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

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

The last step holds because P's largest singular value is smaller than 1, the eigenvalues of PP^T are smaller than 1 and $\mathbf{I} - PP^\mathsf{T} \succeq 0$. Because $\mathbf{A}_{v,t}^{-1} \succ 0$ and $M/\mathbf{A}_{v,t}^{-1} \succeq 0$, according to the property of Schur complement we have $M \succeq 0$. Because $\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I} = \mathbf{A}_{x,t} \succ 0$ and $M \succeq 0$, applying the property again we have $M/\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right) \succeq 0$, which gives us $\mathbf{A}_{v,t}^{-1} - P^\mathsf{T}\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right)^{-1}P \succeq 0$. By the definition of positive semi-definite matrix, we have $\mathbf{v}_a^\mathsf{T}\mathbf{A}_{v,t}^{-1}\mathbf{v}_a - \mathbf{v}_a^\mathsf{T}P^\mathsf{T}\left(\left(P(\mathbf{A}_{v,t} - \lambda\mathbf{I})P^\mathsf{T}\right) + \lambda\mathbf{I}\right)^{-1}P\mathbf{v}_a \geq 0$, which means $\|\mathbf{v}_a\|_{\mathbf{A}_{v,t}^{-1}} \geq \|\mathbf{x}_a\|_{\mathbf{A}_{x,t}^{-1}}$.

According to Lemma 1, $\alpha_t^v = R\sqrt{d_v \log\frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$ and $\alpha_t^x = R\sqrt{d_x \log\frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$. Since $d_v \geq d_x$, we have $\alpha_t^v \geq \alpha_t^x$. Combining the two results and we finished the proof of $CB_{v,t}(\mathbf{v}_a) \geq CB_{x,t}(\mathbf{x}_a)$ holds for any arm a at any time t. □

Proof of Theorem 1. 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,

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

According to Lemma 11 of (Abbasi-yadkori et al., 2011), $\sum_{t=1}^{T}\|\mathbf{x}_a\|^2_{\mathbf{A}_{x,t}^{-1}} \leq d_x \log(\lambda + T/d_v)$. Combining with $\alpha_{x,t} = R\sqrt{d_x \log \frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$ and we finished the proof. □

Proof of Theorem 2. We bound cumulative regret by

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

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 3, where the users are incentivized to pull arms according to UCB exploration strategy as shown in Eq (8). Taking a union bound and 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,

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

where we finished 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 $\alpha_{v,t} = R\sqrt{d_v \log\frac{1+t/\lambda}{\delta}} + \sqrt{\lambda}$ together. □

Proof of Theorem 3. With probability at least $1 - 2\delta$, we have

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

□

Proof of Theorem 4. 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 (Wang & Huang, 2018). We then derive the asymptotic compensation lower bound.

Based on Lemma 4, we can obtain

$$
\limsup_{T\to\infty} \log(T)\|\mathbf{x}_a\|^2_{G_{x,T}^{-1}} \leq \frac{\Delta_a^2}{2} \tag{13}
$$

which is also stated in the Corollary 2 in (Lattimore & Szepesvari, 2017).

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 (13), there exists a $T_a(\delta)$ such that for any $t > T_a(\delta)$, with probability $1-\delta$

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

Combining with the concentration bound in Lemma 5 and 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

$$\max_i \hat{r}_{x,i,t} - \hat{r}_{x,a,t} \geq \hat{r}_{x,1,t} - \hat{r}_{x,a,t} \geq \mathbf{E}[r_1] - \frac{\Delta_2}{3} - \mathbf{E}[r_a] - \frac{\Delta_a}{3} \geq \frac{\Delta_a}{3} \tag{14}$$

with probability at least $1-\delta$.

We then use the optimization problem in Eq (12) 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

$$C(T) \geq \sum_{a\in\mathcal{A}} \mathbf{E}[N_a(T)]\frac{\Delta_a}{3} \geq c_{x,\epsilon}(\mathcal{A}, \boldsymbol{\theta}^*)\log(T) \tag{15}$$

where $c_{x,\epsilon}(\mathcal{A}, \boldsymbol{\theta}^*)$ is the the optimal value of the optimization problem in Eq (12) 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

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

$\square$