Uplifting Bandits

Yu-Guan Hsieh*
University of Grenoble Alpes
yu-guan.hsieh@univ-grenoble-alpes.fr

Shiva Prasad Kasiviswanathan
Amazon
kasivisw@gmail.com

Branislav Kveton
Amazon
bkveton@amazon.com

Abstract

We introduce a multi-armed bandit model where the reward is a sum of multiple random variables, and each action only alters the distributions of some of them. After each action, the agent observes the realizations of all the variables. This model is motivated by marketing campaigns and recommender systems, where the variables represent outcomes on individual customers, such as clicks. We propose UCB-style algorithms that estimate the *uplifts* of the actions over a baseline. We study multiple variants of the problem, including when the baseline and affected variables are unknown, and prove sublinear regret bounds for all of these. We also provide lower bounds that justify the necessity of our modeling assumptions. Experiments on synthetic and real-world datasets show the benefit of methods that estimate the uplifts over policies that do not use this structure.

1 Introduction

Multi-armed bandit (MAB) is an important framework for sequential decision making under uncertainty [8, 24, 26]. In this problem, a learner repeatedly takes action and receives their rewards, while the outcomes of the other actions are unobserved. The goal of the learner is to maximize their cumulative reward over time by balancing exploration (select actions with uncertain reward estimates) and exploitation (select actions with high reward estimates). MAB has applications in online advertisement, recommender systems, portfolio management, and dynamic channel selection in wireless networks [7].

One prominent question in the MAB literature is how the dependencies between the actions can be exploited to improve statistical efficiency. Popular examples include *linear bandits* [13] and *combinatorial bandits* [10]. In this work, we study a different structured bandit problem with the following three features: (*i*) the reward is the sum of the payoffs of a fixed set of variables; (*ii*) these payoffs are observed; and (*iii*) each action only affects a small subset of the variables. This structure arises in many applications, as discussed below.

(a) Online Marketing. An ecommerce platform can opt among several marketing strategies (actions) to encourage customers to make more purchases on their website. As different customers can be sensitive to a different marketing strategy, regarding each of them as a variable, it is natural to expect that each action would likely affect only a subset of the variables. The payoff associated to a customer can be for example the revenue generated by that customer; then the reward is just the total revenue received by the platform.

(b) Product Discount. Consider a company that uses discount strategies to increase their sales. It is common to design discount strategies that only apply to a small subset of products. In this case, we

*Work done during an internship at Amazon.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Algorithm	UCB	UPUCB (b)	UPUCB	UPUCB-nAff	UPUCB-iLift
Affected variables known	No	Yes	Yes	No	No
Baseline payoffs known	No	Yes	No	No	No
Regret Bound	$\frac{Km^2}{\Delta}$	$\frac{KL^2}{\Delta}$		$\frac{KL^2}{\Delta}$	$\frac{K\,\mathrm{clip}(\Delta/\Delta_{\text{up}}, L, m)^2}{\Delta}$

Table 1: Summary of our regret bounds for uplifting bandits. Constant and logarithmic factors are ignored throughout. For simplicity, we assume here all actions affect exactly L of the m variables and all the suboptimal actions have the same suboptimality gap Δ. K and Δ_{up} are respectively the number of actions and a lower bound on individual uplift (Δ_{up} is formally introduced in Appendix E). The operator clip restricts the value of its first variable to the range defined by its second and third variables, $\mathrm{clip}(x, \alpha, \beta) = \max(\alpha, \min(\beta, x))$.

can view the sales of each product as a variable and each discount strategy as an action, and assume that each action only has a significant impact on the sales of those products discounted by this action.

(c) Movie Recommendation. Consider a bandit model for movie recommendation where actions correspond to different recommendation algorithms, and the variables are all the movies in the catalog. For each user, define a set of binary payoffs that indicate whether the user watches a movie in the catalog, and the reward is the number of movies from the catalog that this user watches. Since a recommendation (action) for this user contains (promotes) only a subset of movies, it is reasonable to assume that only their associated payoffs are affected by that action.

Our Contributions. To begin with, we formalize an *uplifting bandit* model that captures the aforementioned structure, with the term *uplift* borrowed from the field of *uplift modeling* [20, 33] to indicate that the actions' effects are incremental over a certain baseline. The model is stochastic, that is, the payoffs of the variables follow an unknown distribution that depends on the chosen action. We study this problem under various assumptions on the learners' prior knowledge about **(1)** the baseline payoff of each variable and **(2)** the set of affected variables of each action. Our first result (Section 3) shows that when both **(1)** and **(2)** are known, a simple modification of the upper confidence bound (UCB) algorithm [3] (Algorithm UPUCB (b)) for estimating the uplifts has a much lower regret than its standard implementation. Roughly speaking, when m is the number of variables and L is the number of variables affected by each action, we get a $O(L^2)$ regret bound instead of $O(m^2)$. This results in a major reduction for $L \ll m$, and is a distinguishing feature of all our results.

Going one step further, we design algorithms that have minimax optimal regret bounds without assuming that either **(1)** or **(2)** is known. When **(1)** the baseline payoffs are unknown and **(2)** the affected variables are known, we compute differences of UCBs to estimate the uplifts (Algorithm UPUCB). In contrast to standard UCB methods, these differences are *not* optimistic, in the sense they are not high-probability upper bounds on the estimated quantity. This construction reflects the fact that the feedback for any single action also provides information about the baseline. When **(2)** the affected variables are also unknown, we identify them on the fly to maintain suitable estimates of the uplifts. We study two approaches, which differ in what they know about the affected variables. Algorithm UPUCB-nAff (Section 6) knows an upper bound on number of affected variables, whereas Algorithm UPUCB-iLift (Appendix E) knows a lower bound on individual uplift. Our regret bounds are summarized in Table 1. These results are further complemented with lower bounds that justify the need for our modeling assumptions (Section 4). To demonstrate the generality of our setup and how our algorithmic ideas extend beyond vanilla multi-armed bandits, we discuss contextual extensions of our model in Section 7. The experiments in Section 8 confirm the benefit of our approach.

Related Work. The goals of both uplifting modeling and MAB is to help selecting the optimal action. The former achieves it by modelling the incremental effect of an action on an individual's behavior. Despite this apparent connection between the two concepts, few papers explicitly link them together. We believe that this is because uplifting can be solved by classic bandit algorithms with a redefined reward. This approach was taken in [6, 34].

Instead, our paper focuses on bandit problems in which estimating the uplifts improves the statistical efficiency of the algorithms, and this is made possible thanks to the 'sparsity' of the actions' effects. Prior to our work, sparsity assumptions in bandits primarily concerned the sparsity of the parameter vectors in linear bandits [2, 5, 21]. A notable exception is Kwon et al. [23] who studied a variant of the MAB where the sparsity is reflected by the fact that the number of arms with positive reward is small. Our work is orthogonal to all of these in that we look at a different form of sparsity. As

we will see in Section 7, while sparsity in parameter can be a cause of sparsity in action effect, the improvement of regret is established with a different mechanism.

The additive structure of the reward and observability of individual payoffs also suggest some similarity between our model and that of combinatorial semi-bandits [11, 22, 31]. There, the learner selects at each round a subset of ground items and the reward is generally defined as the sum of the payoffs of the selected items. However, this seeming similarity comes with an important conceptual difference: the set of items, or variables, for which we observe the outcomes are selected in semi-bandits, while for us they are fixed and inherent to the reward generation mechanism. As an example, in Application **(c)**, combinatorial bandit models optimize the number of movies that are watched among the recommended ones, while we also consider movies that are not recommended. In fact, as argued by Wang et al. [37], a recommendation is only effective if the user actually watches the movie and would not watch it without the recommendation. We refer the readers to Appendix B for more technical details and also a comparison with the causal bandit model.

Organization. We introduce our uplifting bandit model along with its various variations in Section 2. Over Sections 3, 5, 6, we provide regret bounds for these variations, with a lower bound presented in Section 4. We discuss contextual extensions in Section 7 and experimental results in Section 8.

2 Problem Description

We start by formally introducing our uplifting bandit model. We illustrate it in Fig. 1 and summarize our notation in Appendix A. Contextual extensions of this model are discussed in Section 7.

A (K, m)-uplifting bandit is a stochastic bandit with K actions and m underlying variables. Each action $a \in \mathcal{A} := \{1, ..., K\}$ is associated with a distribution $\mathcal{P}^a$ on $\mathbb{R}^m$. At each round t, the learner chooses an action $a_t \in \mathcal{A}$ and receives reward $r_t = \sum_{i \in \mathcal{V}} y_t(i)$ where $\mathcal{V} := \{1, ..., m\}$ is the set of all variables and $y_t = (y_t(i))_{i \in \mathcal{V}} \sim \mathcal{P}^{a_t}$ is the payoff vector.[2] Our model is distinguished by two assumptions that we describe below.

(I) Limited Number of Affected Variables. Let $\mathcal{V}^a \subseteq \mathcal{V}$ be the subset of variables affected by action a and $\mathcal{P}^0$ be the baseline distribution that the variables' payoffs follow when no action is taken. By definition $\mathcal{P}^a$ and $\mathcal{P}^0$ have the same marginal distribution on $\overline{\mathcal{V}^a} := \mathcal{V} \setminus \mathcal{V}^a$, the variables unaffected by action a.[3] While the above condition is always satisfied with $\mathcal{V}^a = \mathcal{V}$, we are interested in the case of $L^a := \mathrm{card}(\mathcal{V}^a) \ll m$, meaning only a few variables are affected by a. We define L as a uniform bound on L^a, so that $\max_{a \in \mathcal{A}} L^a \leq L$. For convenience of notation, we write $[T] := \{1, ..., T\}$ and $\mathcal{A}_0 = \mathcal{A} \cup \{0\}$, where 0 is used for all the quantities related to the baseline.

(II) Observability of Individual Payoff. In addition to the reward r_t, we assume that the learner observes all the payoffs $(y_t(i))_{i \in \mathcal{V}}$ after an action is taken in round t.

Uplift and Noise. Let $y^a = (y^a(i))_{i \in \mathcal{V}}$ be a random variable with distribution $\mathcal{P}^a$. We define $\mu^a(i) = \mathbb{E}[y^a(i)]$ and $\xi^a(i) = y^a(i) - \mu^a(i)$ respectively as the expected value and the noise component of $y^a(i)$. We use μ^a (resp. μ^0) to denote the vector of $(\mu^a(i))_{i \in \mathcal{V}}$ (resp. $(\mu^0(i))_{i \in \mathcal{V}}$), and refer to μ^0 as the baseline payoffs. The *individual* uplift associated to a pair $(a, i) \in \mathcal{A} \times \mathcal{V}$ is defined as $\mu^a_{\text{up}}(i) = \mu^a(i) - \mu^0(i)$. An individual uplift can be *positive* or *negative*. We obtain the (total) uplift of an action by summing its individual uplifts over all the variables affected by that action

$$r^a_{\text{up}} = \sum_{i \in \mathcal{V}^a} \mu^a_{\text{up}}(i) = \sum_{i \in \mathcal{V}^a} (\mu^a(i) - \mu^0(i)). \tag{1}$$

Let $r^a = \sum_{i \in \mathcal{V}} \mu^a(i)$ be the expected reward of an action or of pure observation. We also have $r^a_{\text{up}} = r^a - r^0$ since $\mu^a(i) = \mu^0(i)$ as long as $i \notin \mathcal{V}^a$.

A real-value random variable X is said to be σ-sub-Gaussian if for all $\gamma \in \mathbb{R}$, it holds $\mathbb{E}[\exp(\gamma X)] \leq \exp(\sigma^2 \gamma^2 / 2)$. Throughout the paper, we assume that $\xi^a(i)$ is 1-sub-Gaussian for all $a \in \mathcal{A}_0$ and $i \in \mathcal{V}$. Note that we do not assume that $(y^a(i))_{i \in \mathcal{V}}$ are independent, i.e., the elements in the noise vector $(\xi^a(i))_{i \in \mathcal{V}}$ may be correlated, for the following two reasons:

[2] The terms *reward* and *payoff* distinguish r_t and $(y_t(i))_{i \in \mathcal{V}}$.

[3] If the actions in fact have small impact on the variables in $\overline{\mathcal{V}^a}$, our model is misspecified and incurs additional linear regret whose size is proportional to the impact of the actions on $\overline{\mathcal{V}^a}$. An interesting question is how we can design algorithms that self-adapt to the degree of misspecification.

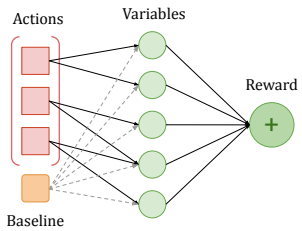
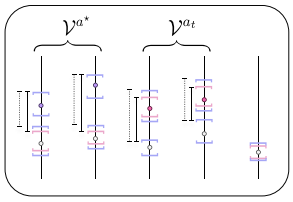
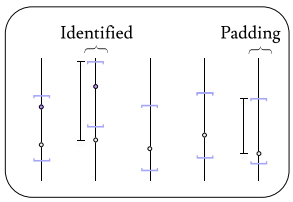


Figure 1: Illustration of the uplifting bandit model. This example has $K = 3$ actions, $m = 5$ variables and each action affects $L^a = 2$ variables. Dash lines indicate the variables' payoffs follow the baseline distribution $\mathcal{P}^0$ by default.



(a) UPUCB: After suboptimal action a_t is taken, the confidence intervals of $(\mu^a(i))_{i\in\mathcal{V}^{a_t}}$ and $(\mu^0(i))_{i\in\mathcal{V}^{a^\star}\setminus\mathcal{V}^{a_t}}$ shrink (from purple to pink). Hence the uplifting index of a_t decreases while that of $a^\star$ increases (from dash to solid).



(b) UPUCB-nAff (b): To compute the uplifting index of an action, we identify a set of variables whose associated confidence intervals do not contain the baseline payoff and pad it with the variables with the largest UCBs on uplifts.

Figure 2: Explanation about UPUCB and UPUCB-nAff (b) using model from Fig. 1. The five vertical bars correspond to the five actions while the y axis corresponds to the value of the payoff.

- The independence assumption is not always realistic. In our first example, it excludes any potential correlation between two customers' purchases.
- While a learner can exploit their knowledge on the noise covariance matrix to reduce the regret, as for example shown by Degenne and Perchet [14], incorporating such knowledge complicates the algorithm design and analysis. However, we believe this is an interesting future direction to pursue.

Regret. The learner's performance is characterized by their regret. To define it, we denote by $a^\star \in \arg\max_{a\in\mathcal{A}} r^a$ an action with the largest expected reward and by $r^\star = r^{a^\star} = \max_{a\in\mathcal{A}} r^a$ the largest expected reward. The regret of the learner after T rounds is then given by

$$\mathrm{Reg}_T = r^\star T - \sum_{t=1}^{T} r^{a_t} = \sum_{a\in\mathcal{A}} \sum_{t=1}^{T} \mathbb{1}\{a_t = a\}\Delta^a, \tag{2}$$

where $\Delta^a = r^\star - r^a$ is the suboptimality gap of action a. In words, we compare the learner's cumulative (expected) reward against the best we can achieve by taking an optimal action at each round. In the following, we also write $\Delta = \min_{a\in\mathcal{A},\Delta^a>0} \Delta^a$ for the minimum non-zero suboptimality gap and refer to it as the suboptimality gap of the problem.

UCB for Uplifting Bandits. The UCB algorithm [4], at each round, constructs an *upper confidence bound* (UCB) on the expected reward of each action and chooses the action with the highest UCB. When applied to our model, we get a regret of $\mathcal{O}(Km^2 \log T/\Delta)$, as the noise in the reward is m-sub-Gaussian (recall that we do not assume independence of the payoffs). However, this approach completely ignores the structure of our problem. As we show in Section 3, we can achieve much smaller regret by focusing on the uplifts.

Problem Variations. In the following sections, we study several variants of the above basic problem that differ in the prior knowledge about $(\mathcal{P}^a)_{a\in\mathcal{A}_0}$ that the learner possesses.

(a) Knowledge of Baseline Payoffs. We consider two scenarios based on whether the learner knows the baseline payoffs $\mu^0 := (\mu^0(i))_{i\in\mathcal{V}}$ or not.

(b) Knowledge of Affected Variables. We again consider two scenarios based on whether the learner knows the affected variables associated with each action $(\mathcal{V}^a)_{a\in\mathcal{A}}$ or not.

3 Case of Known Baseline and Known Affected Variables

We start with the simplest setting where both the affected variables and the baseline payoffs are known. To address this problem, we make the crucial observation that for any two actions, the difference in their rewards equals to that of their uplifts. As an important consequence, uplift maximization has the same optimal action as total reward maximization, and we can replace rewards by uplifts in the definition of the regret (2). Formally, we write $r^\star_{\mathrm{up}} = r^{a^\star}_{\mathrm{up}} = \max_{a\in\mathcal{A}} r^a_{\mathrm{up}}$ for the largest uplift; then $\mathrm{Reg}_T = r^\star_{\mathrm{up}} T - \sum_{t=1}^{T} r^{a_t}_{\mathrm{up}}$. By making this transformation, we gain in statistical efficiency because

$r^a_{\text{up}} = \mathbb{E}[\sum_{i\in\mathcal{V}^a}(y^a(i) - \mu^0(i))]$ can now be estimated much more efficiently under our notion of sparsity. Since both μ^0 and $(\mathcal{V}^a)_{a\in\mathcal{A}}$ are known, we can directly construct a UCB on r^a_{up}. For this, we define for all rounds $t \in [T]$, actions $a \in \mathcal{A}$, and variables $i \in \mathcal{V}$ the quantities

$$N^a_t = \sum_{s=1}^{t} \mathbb{1}\{a_s = a\}, \quad c^a_t = \sqrt{\frac{2\log(1/\delta')}{N^a_t}}, \quad \hat{\mu}^a_t(i) = \sum_{s=1}^{t} \frac{y_s(i)\,\mathbb{1}\{a_s = a\}}{\max(1, N^a_t)}, \tag{3}$$

where $\delta' > 0$ is a tunable parameter. In words, N^a_t, $\hat{\mu}^a_t(i)$, and c^a_t represent respectively the number of times that action a is taken, the empirical estimate of $\mu^a(i)$, and the associated radius of confidence interval calculated at the end of round t. The UCB for action $a \in \mathcal{A}$ and variable $i \in \mathcal{V}^a$ at round t is $U^a_t(i) = \hat{\mu}^a_{t-1}(i) + c^a_{t-1}$. We further define $\tau^a_t = \sum_{i\in\mathcal{V}^a}(U^a_t(i) - \mu^0(i))$ as the *uplifting index*, and refer to UPUCB (b) as the algorithm that takes an action with the largest uplifting index τ^a_t at each round t (Algorithm 3, Appendix A). Here, the suffix (b) indicates that the method operates with the knowledge of the baseline payoffs.

Since UPUCB (b) is nothing but a standard UCB with transformed rewards $r'_t = \sum_{i\in\mathcal{V}^{a_t}}(y_t(i) - \mu^0(i))$, and $\mathbb{E}[r'_t] = r^{a_t}_{\text{up}}$ defines exactly the same regret as the original reward, a standard analysis for UCB yields the following result.

Proposition 1. *Let $\delta' = \delta/(2KT)$. Then the regret of* UPUCB (b) *(Algorithm 3, Appendix A), with probability at least $1-\delta$, satisfies*

$$\text{Reg}_T \leq \sum_{a\in\mathcal{A}:\Delta^a>0} \left(\frac{8(L^a)^2 \log(2KT/\delta)}{\Delta^a} + \Delta^a \right). \tag{4}$$

As expected, the regret bound does not depend on m and scales with L^2. This is because the transformed reward of action a is only L^a-sub-Gaussian. The improvement is significant when $L \ll m$. However, this also comes at a price, as both the baseline payoffs and affected variables need to be known. We address these shortcomings in Sections 5 and 6. On a side note, we remark that observing the aggregated payoff $\sum_{i\in\mathcal{V}^{a_t}} y_t(i)$ is sufficient for UPUCB (b), but this will not be the case for the other algorithms presented in our paper.

Estimating Baseline Payoffs from Observational Data. In practice, the baseline payoffs μ^0 can be estimated from observational data, which gives confidence intervals that $\mu^0(i)$ lie in with high probability. The uplifting indices can then be constructed by subtracting the lower confidence bounds of $\mu^0(i)$ from $U^a_t(i)$. If the number of samples is n, the widths of the confidence intervals are in $\mathcal{O}(1/\sqrt{n})$. Identification of a^* is possible only when the sum of these widths over L variables is at most Δ, which requires $n = \Omega(L^2/\Delta^2)$. Otherwise, these errors persist at each iteration and n must be in the order of T to ensure $\mathcal{O}(\sqrt{T})$ regret.

Gap-Free Bounds. In Proposition 1, we state a high-probability instance-dependent regret bound, and we will continue to do so for all the regret upper bounds that we present for the non-contextual variants. This type of result can be directly transformed to bound on $\mathbb{E}[\text{Reg}_T]$ by taking $\delta = 1/T$. Following the routine of separating Δ^a into two groups depending on their scale, most of our proofs can be modified to obtain a gap-free bound, which is usually in the order of $\mathcal{O}(L\sqrt{KT\log(T)})$. We will not state these results to avoid unnecessary repetitions.

4 Lower Bounds

In this section, we shortly discuss the necessity of our modeling assumptions for obtaining the improved regret bounds of Proposition 1. The complete discussion appears in Appendix D.

Intuitively, the regret can be improved both because the noise in the effect of an action is small, and because the observation of this effect does not heavily deteriorate with noise. These two points correspond respectively to assumptions **(I)** and **(II)**. Moreover, the knowledge on $(\mathcal{V}^a)_{a\in\mathcal{A}}$ allows the learner to distinguish between problems with different structures. Without such distinction, there is no chance that the learner can leverage the underlying structure. Therefore, the aforementioned three points are crucial for obtaining (4). Below, we establish this formally for algorithms that are *consistent* [24] over the class of 1-sub-Gaussian uplifting bandits, which means the induced regret of the algorithm on any uplifting bandit with 1-sub-Gaussian noise satisfies $\text{Reg}_t = o(t^p)$ for all $p > 0$.

Proposition 2. *Let π be a consistent algorithm over the class of 1-sub-Gaussian uplifting bandits that at most uses the knowledge of $\mathcal{P}^0$, $(\mathcal{V}^a)_{a\in\mathcal{A}}$, and the fact that the noise is 1-sub-Gaussian. Let $K, m > 0$ and sequence $(L^a, \Delta^a)_{1\le a\le K} \in ([m]\times\mathbb{R}_+)^K$ satisfy $\Delta^1 = 0$. Assume either of the following holds.*

(a) $L^a = m$ for all $a \in \mathcal{A}$, so that in the bandits considered below all actions affect all variables.

(b) Only the reward is observed.

(c) The algorithm π does not make use of any prior knowledge about the arms' expected payoffs $(\mu^a)_{a\in\mathcal{A}}$ (in particular, the knowledge of $(\mathcal{V}^a)_{i\in\mathcal{V}}$ is not used by π).

Then, there exists a 1-sub-Gaussian (K, m)-uplifting bandit whose suboptimality gaps and numbers of affected variables are respectively $(\Delta^a)_{a\in\mathcal{A}}$ and $(L^a)_{a\in\mathcal{A}}$, such that the regret induced by π on it satisfies: $\liminf_{T\to+\infty} \frac{\mathbb{E}[\mathrm{Reg}_T]}{\log T} \ge \sum_{a\in\mathcal{A}:\Delta^a>0} \frac{2m^2}{\Delta^a}$.

Proposition 2 states an instance-dependent lower-bound for a learner that may be equipped with full knowledge of the baseline distribution.[4] Its proof is presented in Appendix D. At this point, what remains unclear is whether similar improvement of the regret is still possible when the baseline is unknown or when the learner only has access to more restricted knowledge than $(\mathcal{V}^a)_{a\in\mathcal{A}}$. We give affirmative answers to these two questions in the next two sections.

5 Case of Unknown Baseline

In this section, we consider the situation where the learner knows the sets of affected variables $(\mathcal{V}^a)_{a\in\mathcal{A}}$, but not the baseline payoffs. Since the actual uplift at each round is never observed, the uplifts of the actions can hardly be estimated directly in this case. Instead, we follow a two-model approach. Leveraging the fact that $\mathcal{P}^a$ and $\mathcal{P}^0$ have the same marginal distribution on $\overline{\mathcal{V}^a}$, we can estimate the baseline payoffs by aggregating information from the feedback of different actions. This leads to

$$N_t^0(i) = \sum_{s=1}^{t} \mathbb{1}\{i \notin \mathcal{V}^{a_s}\}, \quad c_t^0(i) = \sqrt{\frac{2\log(1/\delta')}{N_t^0(i)}}, \quad \hat{\mu}_t^0(i) = \frac{\sum_{s=1}^{t} y_s(i)\,\mathbb{1}\{i \notin \mathcal{V}^{a_s}\}}{\max(1, N_t^0(i))}. \tag{5}$$

Compared to (3), we notice that both N_t^0 and c_t^0 are functions of i. This is because for each taken action a, we only increase the counters $N_t^0(i)$ for those $i \notin \mathcal{V}^a$, which causes a non-uniform increase of $(N_t^0(i))_{i\in\mathcal{V}}$. Then, by looking at all the rounds that variable i is not influenced by the chosen action, we manage to compute $\hat{\mu}_t^0(i)$, an estimate of $\mu^0(i)$. To proceed, we define the following UCB indices for all the pairs $(a, i) \in \mathcal{A}_0 \times \mathcal{V}$

$$U_t^a(i) = \begin{cases} 0 & \text{if } a = 0 \text{ and } i \in \bigcap_{a\in\mathcal{A}} \mathcal{V}^a, \\ \hat{\mu}_{t-1}^a(i) + c_{t-1}^a & \text{if } a \in \mathcal{A} \text{ and } i \in \mathcal{V}^a, \\ \hat{\mu}_{t-1}^0(i) + c_{t-1}^0(i) & \text{otherwise.} \end{cases} \tag{6}$$

The second and the third lines of (6) contain the usual definition of UCBs using the empirical estimates and the radii of the confidence intervals defined in (3) and (5). In the special case that a variable is affected by all the actions, it is impossible to estimate $U^0(i)$ but it is enough to compare $U^a(i)$ directly against $U^{a'}(i)$ for any two actions $a, a' \in \mathcal{A}$, so we just set $U_t^0(i)$ to 0 in this case.

We outline the proposed method, UpUCB, in Algorithm 1. The uplifting indices are given by $\tau_t^a = \sum_{i\in\mathcal{V}^a}(U_t^a(i) - U_t^0(i))$. It may be counter-intuitive to compare the differences between two UCBs. Indeed, τ_t^a is no longer an optimistic estimate of the uplifting effect r_{up}^a, but it captures the essential trade-off between learning action a's payoffs and learning the baseline μ^0. To provide some intuition, we give an informal justification of UpUCB in Fig. 2a: If a suboptimal action a is taken in round t, the estimates of all $U_t^a(i)$ move closer to the actual mean from above. As a result, τ_t^a decreases, since all $U_t^a(i)$ for $i \in \mathcal{V}^a$ do. Thus action a is less likely to be taken next. Moreover, $\tau_t^{a^\star}$ increases, since $U_t^0(i)$ decrease for any i affected by $a^\star$ but not a. Thus $a^\star$ is more likely to be taken next. The effectiveness of UpUCB is confirmed by the following theorem.

[4] Of course, the problem only becomes more challenging if the learner does not know the baseline distribution.

Algorithm 1 UPUCB

Input: Error probability δ', the sets of variables each action affects $\{\mathcal{V}^a : a \in \mathcal{A}\}$
Initialization: Take each action once
for $t = K+1, \dots, T$ **do**
  Compute the UCB indices following (6)
  For $a \in \mathcal{A}$, set $\tau_t^a \leftarrow \sum_{i \in \mathcal{V}^a} (U_t^a(i) - U_t^0(i))$
  Select action $a_t \in \arg\max_{a \in \mathcal{A}} \tau_t^a$

Theorem 1. *Let $\delta' = \delta/(4KLT)$. Then the regret of* UPUCB *(Algorithm 1) with probability at least $1-\delta$, satisfies:*

$$\mathrm{Reg}_T \leq \sum_{a \in \mathcal{A}: \Delta^a > 0} \left(\frac{8(L^a + L^{a^\star})^2 \log(4KLT/\delta)}{\Delta^a} + \Delta^a \right). \tag{7}$$

Idea of proof. The full proof of Theorem 1 is presented in Appendix C.2, and proceeds as following.

1. With concentration of measure, we show that with probability $1-\delta$, it holds $|\hat{\mu}_t^a(i) - \mu^a(i)| \leq c_t^a$ and $|\hat{\mu}_t^0(i) - \mu^0(i)| \leq c_t^0(i)$ for all relevant estimates. It is thus sufficient to show that (7) holds when these inequalities are satisfied.

2. A suboptimal action a can only be taken if its uplifting index is larger than that of $a^\star$, i.e., if $\sum_{i \in \mathcal{V}^a} (U_t^a(i) - U_t^0(i)) \geq \sum_{i \in \mathcal{V}^{a^\star}} (U_t^{a^\star}(i) - U_t^0(i))$. Rearranging, we get

$$\sum_{i \in \mathcal{V}^a} U_t^a(i) + \sum_{i \in \mathcal{V}^{a^\star} \setminus \mathcal{V}^a} U_t^0(i) \geq \sum_{i \in \mathcal{V}^{a^\star}} U_t^{a^\star}(i) + \sum_{i \in \mathcal{V}^a \setminus \mathcal{V}^{a^\star}} U_t^0(i). \tag{8}$$

Using the inequalities mentioned in the previous point and $\Delta^a = r_{\text{up}}^{a^\star} - r_{\text{up}}^a$, bounding the two sides of (8), we deduce $\Delta^a \leq 2L^a c_{t-1}^a + \sum_{i \in \mathcal{V}^{a^\star} \setminus \mathcal{V}^a} 2c_{t-1}^0(i)$.

3. Note that for all $i \in \mathcal{V}^{a^\star} \setminus \mathcal{V}^a$, whenever action a is taken, the count of $N_s^0(i)$ also increases by 1. We have thus $N_{t-1}^0(i) \geq N_{t-1}^a$ and accordingly $c_{t-1}^0(i) \leq c_{t-1}^a$. We then get $\Delta^a \leq 2(L^a + L^{a^\star}) c_{t-1}^a$. This allows us to bound the number of times that action a is taken and conclude. □

As in Proposition 1, the regret of Theorem 1 is in $\mathcal{O}(KL^2 \log T/\Delta)$. In fact, all decisions in UPUCB are made on uncertain estimates of at most L variables; thus the statistical efficiency scales with L and not m. A detailed comparison with Proposition 1 reveals that the difference is in the order of $K(L^{a^\star})^2 \log T/\Delta$; this is only significant when the optimal actions affect many more variables then the suboptimal ones. Hence, the price of not knowing the baseline payoffs is generally quite small.

6 Case of Unknown Affected Variables

Now we study the more challenging setting where the affected variables $(\mathcal{V}^a)_{a \in \mathcal{A}}$ are unknown to the learner. Proposition 2 states that improvement is impossible if the learner does not have any prior knowledge to exploit the structure. To circumvent this negative result, we study two weak assumptions motivated by practice: learner has access to (*i*) an upper bound on the number of affected variables or (*ii*) a lower bound on individual uplift. Due to space constraints, we only present the first setting here and defer the discussion of the other to Appendix E.

For the rest of this section, we assume that we know an upper bound on the number of affected variables L (i.e., $L \geq \max_{a \in \mathcal{A}} L^a$). We design algorithms with $\mathcal{O}(KL^2 \log T/\Delta)$ regret bounds that takes L as input. We consider the cases of known and unknown baseline payoffs.

Known Baseline Payoffs. To illustrate our ideas, we start by assuming that the baseline payoffs are known. We propose an optimistic algorithm that maintains a UCB on the total uplift with an overestimate in the order of L. Let N_t^a, c_t^a, and $\hat{\mu}_t^a(i)$ be defined as in (3). The UCB, uplifting indices, and the confidence intervals for each (action, variable) pair $(a, i) \in \mathcal{A} \times \mathcal{V}$ are

$$U_t^a(i) = \hat{\mu}_{t-1}^a(i) + c_{t-1}^a,\ \rho_t^a(i) = U_t^a(i) - \mu^0(i),\ \mathcal{C}_t^a(i) = [\hat{\mu}_{t-1}^a(i) - c_{t-1}^a,\ \hat{\mu}_{t-1}^a(i) + c_{t-1}^a]. \tag{9}$$

In the rest of the paper, we refer to $\rho_t^a(i)$ as the *individual uplifting index* of the pair (a, i). It is

Algorithm 2 UPUCB-nAff

1: **Input:** Error probability δ', Upper bound L on the number of variables that each action affects
2: **Initialization:** Take each action once
3: **for** $t = K+1, \ldots, T$ **do**
4: Choose $b_t \in \arg\max_{a \in \mathcal{A}} N_{t-1}^a$ and compute UCBs and confidence intervals using (9)
5: **for** $a \in \mathcal{A}$ **do**
6: Set $\widehat{\mathcal{V}}_t^a \leftarrow \{i \in \mathcal{V} : \mathcal{C}_t^a(i) \cap \mathcal{C}_t^{b_t}(i) = \emptyset\}$
7: For $i \in \mathcal{V}$, compute $\rho_t^a(i) \leftarrow U_t^a(i) - U_t^{b_t}(i)$
8: Set $L_t^a \leftarrow \max(0, 2L - \mathrm{card}(\widehat{\mathcal{V}}_t^a))$ and $\mathcal{L}_t^a \leftarrow \arg\max_{\mathcal{L} \subseteq \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a,\ \mathrm{card}(\mathcal{L}) \le L_t^a} \sum_{i \in \mathcal{L}} \rho_t^a(i)$
9: Compute uplifting index $\tau_t^a \leftarrow \sum_{i \in \widehat{\mathcal{V}}_t^a \cup \mathcal{L}_t^a} \rho_t^a(i)$
10: Select action $a_t \in \arg\max_{a \in \mathcal{A}} \tau_t^a$

an overestimate of the individual uplift $\mu_{\text{up}}^a(i)$. As for $\mathcal{C}_t^a(i)$, it is the confidence interval that $\mu^a(i)$ lies in with high probability. Our algorithm, UPUCB-nAff (b) with nAff for number of affected, leverages two important procedures to compute an optimistic estimate of the uplift r_{up}^a: identification of affected variables, and padding with variables with the highest individual uplifting indices.

To begin, UPUCB-nAff (b) constructs the set of *identified variables* $\widehat{\mathcal{V}}_t^a = \{i \in \mathcal{V} : \mu^0(i) \notin \mathcal{C}_t^a(i)\}$ which is contained in $\mathcal{V}^a$ with high probability. In fact, by concentration of measure, with high probability $\mu^a(i) \in \mathcal{C}_t^a(i)$, in which case $\mu^0(i) \notin \mathcal{C}_t^a(i)$ indicates $i \in \mathcal{V}^a$. However, $\widehat{\mathcal{V}}_t^a$ is not guaranteed to capture all the affected variables, so we also need to provide an upper bound for $\sum_{i \in \mathcal{V}^a \setminus \widehat{\mathcal{V}}_t^a} \mu_{\text{up}}^a(i)$, the uplift contributed by the unidentified affected variables. Since the individual uplifting index $\rho_t^a(i)$ is in fact a UCB on the individual uplift $\mu_{\text{up}}^a(i)$ here and $\mathrm{card}(\mathcal{V}^a) \le L$, we can simply choose the $L - \mathrm{card}(\widehat{\mathcal{V}}_t^a)$ variables in $\mathcal{V} \setminus \widehat{\mathcal{V}}_t^a$ with the largest $\rho_t^a(i)$. Let us refer to this set as $\mathcal{L}_t^a$. We then get a proper UCB on the uplift of action a by computing $\tau_t^a = \sum_{i \in \widehat{\mathcal{V}}_t^a \cup \mathcal{L}_t^a} \rho_t^a(i)$. This process is summarized in Algorithm 4 in Appendix A and illustrated in Fig. 2b.

Unknown Baseline Payoffs. Now we focus on the most challenging setting, where also the baseline payoffs are unknown. In this case, neither the sets of identified variables nor the uplifting indices of UPUCB-nAff (b) can be defined. We also cannot estimate the baseline payoffs using (5) since the sets of affected variables are unknown. To overcome these challenges, we note that for any two actions $a, a' \in \mathcal{A}$, μ^a and $\mu^{a'}$ only differ on $\mathcal{V}^a \cup \mathcal{V}^{a'}$, and $\mathrm{card}(\mathcal{V}^a \cup \mathcal{V}^{a'}) \le 2L$. Therefore, μ^a and $\mu^{a'}$ differ in at most $2L$ variables, and we recover a similar problem structure by taking the payoffs of any action as the baseline.

Combining this idea with the elements that we have introduced previously, we obtain UPUCB-nAff (Algorithm 2). In each round, UPUCB-nAff starts by picking a most frequently taken action b_t (Line 4) whose payoffs are treated as the baseline in that round. Then, in Line 6, UPUCB-nAff chooses variables that are guaranteed to be either in $\mathcal{V}^a$ or $\mathcal{V}^{b_t}$. This generalizes the identification step of UPUCB-nAff (b). The individual uplifting indices $\rho_t^a(i) = U_t^a(i) - U_t^{b_t}(i)$ are computed in Line 7. The differences of UCBs are inspired by a similar construction in UPUCB. Line 8 constitutes the padding step, during which variables with the highest uplifting indices are selected, and finally in Line 9 we combine the above to get the uplifting index of the action. To see why UPUCB-nAff is similar to UPUCB-nAff (b), suppose that one action has been taken frequently. Then the baseline payoffs are precisely estimated and do not change much between consecutive rounds.

Regret. Both UPUCB-nAff (b) and UPUCB-nAff choose $\mathcal{O}(L)$ variables for estimating the uplift of an action, and the decisions are based on these estimates. Therefore, the statistical efficiencies of these algorithms only scale with L and not m. This in turn translates into an improvement of the regret, as demonstrated by the theorem below.

Theorem 2. *Let $\delta' = \delta/(2KmT)$. Then the regret of* UPUCB-nAff *(Algorithm 2) (resp.* UPUCB-nAff (b)*, Algorithm 4), with probability at least $1 - \delta$, satisfies:*

$$\mathrm{Reg}_T \le \sum_{a \in \mathcal{A} : \Delta^a > 0} \left(\frac{\alpha L^2 \log(2KmT/\delta)}{\Delta^a} + \Delta^a \right), \tag{10}$$

where $\alpha = 512$ (resp. 32) in the above inequality.

Idea of proof. The full proof of Theorem 2 is presented in Appendix C.3. We outline below the main steps to prove the result for UPUCB-nAff (b).

1. With concentration of measure, with probability $1-\delta$ it holds $\mu^a(i) \in \mathcal{C}_t^a(i)$. Then, as explained in the text, $\widehat{\mathcal{V}}_t^a \subseteq \mathcal{V}^a$ and τ_t^a is an upper bound on r_{up}^a; especially $\tau_t^{a^\star} \geq r_{\text{up}}^{a^\star}$.
2. For $i \notin \widehat{\mathcal{V}}_t^a$, by definition $\mu^0(i) \in \mathcal{C}_t^a$, from which we deduce $0 \leq \rho_t^a(i) \leq 2c_{t-1}^a$. With $\widehat{\mathcal{V}}_t^a \subseteq \mathcal{V}^a$ and $\mathcal{L}_t^a \subseteq \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a$ we can then write
$$\tau_t^a = \sum_{i \in \widehat{\mathcal{V}}_t^a \cup \mathcal{L}_t^a} \rho_t^a(i) \leq \sum_{i \in \mathcal{V}^a} \rho_t^a(i) + \sum_{i \in \mathcal{L}_t^a} \rho_t^a(i) \leq r_{\text{up}}^a + 4Lc_{t-1}^a.$$
We have also used $\text{card}(\mathcal{V}^a) \leq L$, $\text{card}(\mathcal{L}_t^a) \leq L$, and $\rho_t^a(i) \leq \mu^a(i) + 2c_{t-1}^a$.
3. Whenever a suboptimal action a is taken, we have $\tau_t^a \geq \tau_t^{a^\star}$. Combined with the two previous point we then deduced $\Delta^a \leq 4Lc_{t-1}^a$. Subsequently, we bound the number of times that each suboptimal action a is taken to conclude. □

The proof of Theorem 2 is notable for two reasons. First, tracking of the identified variables guarantees that the uplifting index τ_t^a does not overestimate the uplift r_{up}^a too much. Take UPUCB-nAff (b) as an example. An alternative to constructing a UCB on r_{up}^a is to choose the L variables with the highest individual uplifting indices $\rho_t^a(i)$. However, this would result in a severe overestimate when a negative individual uplift is present. Second, to prove (10) for UPUCB-nAff, we use that the widths of confidence intervals of the chosen b_t are always smaller than those of the taken action. This is ensured by taking b_t as the most frequent action (Line 4 in Algorithm 2).

7 Contextual Extensions

In this section, we briefly discuss potential contextual extensions of our model; a detailed case study is presented in Appendix F. As in contextual bandits, context is a side information that helps the learner to make a more informed decision, which results in a higher reward. To incorporate context, one possibility is to associate each variable with a feature vector $x_t(i) \in \mathbb{R}^d$. The subscript t indicates that the context can change from one round to another. We also associate each action with a function f^a so that the expected payoff of action a acting on a variable with feature $x_t(i)$ is $f^a(x_t(i))$. The expected reward of choosing a at round t is then $r^a(x_t) = \sum_{i \in \mathcal{V}} f^a(x_t(i))$. The optimal action in round t is $a_t^\star = \arg\max_{a \in \mathcal{A}} r^a(x_t)$ and the regret of a learner that takes the actions $(a_t)_{t \in [T]}$ is given by $\text{Reg}_T = \sum_{t=1}^T \sum_{i \in \mathcal{V}} (f^{a_t^\star}(x_t(i)) - f^{a_t}(x_t(i)))$.

The key structure in our model (Section 2) is that there exists a baseline payoff vector μ^0 such that for any given action a, $\mu^a(i) = \mu^0(i)$ holds for most $i \in \mathcal{V}$. Given context, this translates into the existence of a baseline function f^0 such that for any given a and t, $f^a(x_t(i)) = f^0(x_t(i))$ holds for most $i \in \mathcal{V}$. The uplift of action a is defined as $r_{\text{up}}^a(x_t) = \sum_{i \in \mathcal{V}} f^a(x_t(i)) - f^0(x_t(i))$.

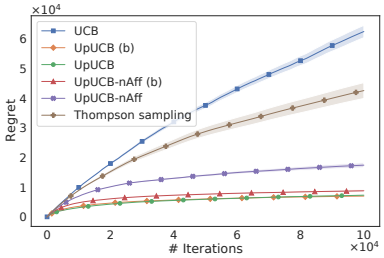
For concreteness, let us consider a model with linear payoffs. Then, each action is associated with an unknown parameter θ^a and the expected payoff is the scalar product of θ^a and the feature of the variable, i.e., $f^a(x_t(i)) = \langle \theta^a, x_t(i) \rangle$. We also assume the aforementioned equality to hold for the baseline function f^0 and we use $\mathcal{V}_t^a = \{i \in \mathcal{V} : \langle \theta^a, x_t(i) \rangle \neq \langle \theta^0, x_t(i) \rangle\}$ for the variables affected by action a at round t. One sufficient condition for $\mathcal{V}_t^a$ to be small is sparsity in both the parameter difference $\theta_{\text{up}}^a = \theta^a - \theta^0$ and the context vector $x_t(i)$. In fact, $\langle \theta^a, x_t(i) \rangle = \langle \theta^0, x_t(i) \rangle$ as long as the supports of θ_{up}^a and $x_t(i)$ are disjoint. We assume $\text{card}(\mathcal{V}_t^a)$ is uniformly bounded by L below.

Clearly, our algorithms can be directly applied as long as we can construct a UCB on $\langle \theta^a, x_t(i) \rangle$. This can for example be done using the construction of linear UCB [1]. In this way, the decision of the learner is again based on the uncertain estimates of at most $\mathcal{O}(L)$ variables, and we expect similar improvements as in our earlier theorems. As an example, when both θ^0 and $\mathcal{V}_t^a$ are known, UPUCB (b) adapted to this situation constructs UCB for $\sum_{i \in \mathcal{V}_t^a} \langle \theta_{\text{up}}^a, x_t(i) \rangle$, and it is straightforward to show that the regret of such algorithm can be as small as $\mathcal{O}(Ld\sqrt{KT})$. In contrast, if the learner works directly with the total reward, the regret is in $\mathcal{O}(md\sqrt{KT})$.[5]

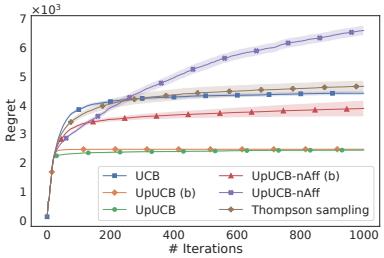
[5] We present gap-free bounds here and thus we get L versus m in the place of L^2 versus m^2. As in previous sections, these bounds apply to a potentially dependent noise.

8 Numerical Experiments

In this section, we present numerical experiments to demonstrate the benefit of estimating uplifts in our model. We compare our methods introduced in Sections 3, 5 and 6 against UCB and Thompson sampling with Gaussian prior and Gaussian noise that only use the observed rewards $(r_t)_{t\in[T]}$. To ensure a fair comparison, we tune all the considered algorithms and report results for the parameters that yield the best average performance. The detailed procedure, and additional experimental details and results, are provided in Appendices G and H. The experiments for the contextual extension in Section 7 are presented in Appendix F.3.



(a) Gaussian uplifting



(b) Bernoulli uplifting

Figure 3: Experimental results on a synthetic and real-world dataset. All the curves are averaged over 100 runs and the shaded areas represent the standard errors.

Gaussian Uplifting Bandit. We first study our algorithms in a Gaussian uplifting bandit with $K = 10$ actions, $m = 100$ variables, and $L^a = 10$ for all $a \in \mathcal{A}$ meaning that each action affects 10 variables.[6] The expected payoffs are contained in $[0, 1]$, and the covariance matrix of the noise is taken the same for all the actions. The suboptimality gap of the created problem is around 0.2, and the variance of the total noise $\sum_{i\in\mathcal{V}} \xi^a(i)$ is around 80.

Bernoulli Uplifting Bandit with Criteo Uplift. We use the Criteo Uplift Prediction Dataset [15] with 'visit' as the outcome variable to build a Bernoulli uplifting bandit, where the payoff of each variable has a Bernoulli distribution. This dataset is designed for uplift modeling, and has outcomes for both treated and untreated individuals. Thus it is suitable for our simulations. To build the model, we sample 10^5 examples from the dataset, and use K-means to partition these samples into 20 clusters of various sizes. The 10^5 examples are taken as our variables. We consider 20 actions that correspond to treating individuals of each cluster, and construct independent Bernoulli payoffs using the visit rates of the treated/untreated individuals of the clusters following a procedure detailed in Appendix G.2. Here, $L = 12654$ and Δ is around 30.

Results. Fig. 3 confirms that we can effectively achieve much smaller regret by restricting our attention to the uplift. Moreover, when the sets of affected are known, the loss of not knowing the baseline payoffs seems to be minimal. On the other hand, not knowing the affected variables has a more severe effect in the second experiment. In fact, the design of UpUCB-nAff and UpUCB-nAff (b) heavily rely on the additive structure of the uplifting index, and can thus hardly benefit from the payoff independence which allow the other four algorithms to achieve smaller regret in this case.

9 Concluding Remarks

This paper studies multi-armed bandit problems where the rewards are sums of observable variables. When each action only affects a limited number of these variables, much smaller regret can be achieved, and we developed algorithms with such guarantees under different forms of knowledge that the learner possesses.

While we study here a UCB-style algorithm, we believe that understanding how similar improvement can be achieved by other types of algorithms such as Thompson sampling and information directed sampling is an important question. Moreover, further extending our work to cope with non-stationary or even adversarial bandits is another promising direction to pursue. As for the former, a direct combination with existing techniques [19] can readily make our algorithms bypass the stationarity assumption that we make throughout the paper.

[6] By Gaussian uplifting bandit we mean that the noise vectors $(\xi^a)_{a\in\mathcal{A}}$ are Gaussian.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24:2312–2320, 2011.

[2] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *Artificial Intelligence and Statistics*, pages 1–9. PMLR, 2012.

[3] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

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

[5] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

[6] Jeroen Berrevoets, Sam Verboven, and Wouter Verbeke. Treatment effect optimisation in dynamic environments. *Journal of Causal Inference*, 10(1):106–122, 2022.

[7] Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on applications of multi-armed and contextual bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE, 2020.

[8] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning*, 5(1):1–122, 2012.

[9] Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

[10] Nicolo Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

[11] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159. PMLR, 2013.

[12] Kai Lai Chung and Kailai Zhong. *A course in probability theory*. Academic press, 2001.

[13] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, pages 355–366, 2008.

[14] Rémy Degenne and Vianney Perchet. Combinatorial semi-bandit with known covariance. In *Advances in Neural Information Processing Systems*, pages 2972–2980, 2016.

[15] Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In *International Conference on Knowledge Discovery and Data Mining*. ACM, 2018.

[16] Joseph L. Doob. *Stochastic processes*. John Wiley & Sons, 1953.

[17] Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(6), 2006.

[18] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, volume 23, pages 586–594, 2010.

[19] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for switching bandit problems. In *International Conference on Algorithmic Learning Theory*, pages 174–188. Springer, 2011.

[20] Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In *International Conference on Predictive Applications and APIs*, pages 1–13. PMLR, 2017.

[21] Botao Hao, Tor Lattimore, and Wei Deng. Information directed sampling for sparse linear bandits. In *Advances in Neural Information Processing Systems*, 2021.

[22] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.

[23] Joon Kwon, Vianney Perchet, and Claire Vernade. Sparse stochastic bandits. In *Conference on Learning Theory*, volume 65, 2017.

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

[25] Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in neural information processing systems*, 2016.

[26] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[27] Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? *Advances in Neural Information Processing Systems 31*, 31, 2018.

[28] Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pages 141–150. PMLR, 2020.

[29] Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Causal bandits with unknown graph structure. In *Advances in Neural Information Processing Systems*, 2021.

[30] Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 2017–2025. PMLR, 2021.

[31] Pierre Perrault, Etienne Boursier, Michal Valko, and Vianney Perchet. Statistical efficiency of thompson sampling for combinatorial semi-bandits. *Advances in Neural Information Processing Systems*, 33:5429–5440, 2020.

[32] Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 461–469. SIAM, 2014.

[33] Nicholas Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. *Direct Marketing Analytics Journal*, pages 14–21, 2007.

[34] Neela Sawant, Chitti Babu Namballa, Narayanan Sadagopan, and Houssam Nassif. Contextual multi-armed bandits for causal marketing. In *International Conference on Machine Learning*, 2018.

[35] István Szita and Csaba Szepesvári. Agnostic KWIK learning and efficient approximate reinforcement learning. In *Conference on Learning Theory*, pages 739–772. JMLR Workshop and Conference Proceedings, 2011.

[36] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[37] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommender systems. In *Fourteenth ACM Conference on Recommender Systems*, pages 426–431, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We have discussed potential extensions, which reflects the limitation of the current work.
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Provided in the supplmental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 8 and Appendices F.3, G and H.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix G
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix G
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix for "Uplifting Bandits"

Table of Contents

A Notation Table and Missing Pseudo Codes **15**

B Additional Related Work **16**

C Proofs for Upper Bound Results **17**

C.1 Concentration Bounds 17

C.2 Unknown Baseline 18

C.3 Unknown Affected Variables 20

D Lower Bounds **23**

D.1 Lower Bounds for Uplifting Bandits– Statements and Discussion 24

D.2 A General Information-theoretic Lower Bound 25

D.3 Lower Bounds for Uplifting Bandits– Proof 26

E Case of Unknown Affected Variables: Scenario 2 **27**

E.1 Known Baseline Payoffs 27

E.2 Unknown Baseline Payoffs 29

F Contextual Combinatorial Uplifting Bandits **32**

F.1 Model and Algorithm Template 32

F.2 Linear Expected Payoffs 33

F.3 Experiments for Targeted Campaign with Criteo Uplift Dataset 36

G Additional Experimental Details **38**

G.1 Choice of Algorithm Parameters 38

G.2 Payoff Construction for Bernoulli Uplifting Bandit with Criteo Uplift 39

H Ablation Study on Gaussian Uplifting Bandit **39**

H.1 Baseline UCB versus Baseline LCB. 40

H.2 Misspecification of L. 40

H.3 Ignoring Unaffected variables in UCB and Thompson Sampling 40

H.4 Actions with Different Covariance Matrices 41

A Notation Table and Missing Pseudo Codes

	Notation	Explanation
Bandits	K	Number of actions
	a	An action
	$a^\star$	An optimal action, $a^\star \in \arg\max_{a\in\mathcal{A}} r^a$
	$\mathcal{A}$	The set of actions, $\mathcal{A} = \{1, ..., K\}$
	t	Round index
	T	Number of rounds / Time horizon
	$[T]$	The set of all the rounds, $[T] = \{1, ..., T\}$
	r_t	Reward received at round t
	r^a	Expected reward of action a
	$r^\star$	Largest expected reward, $r^\star = r^{a^\star} = \max_{a\in\mathcal{A}} r^a$
	Δ^a	Suboptimality gap of action a, $\Delta^a = r^\star - r^a$
	Δ	Minimum non-zero suboptimality gap $\Delta = \min_{a\in\mathcal{A},\Delta^a>0} \Delta^a$
Uplifting Bandits	m	Number of underlying variables
	i	A variable
	$\mathcal{V}$	The set of variables, $\mathcal{V} = \{1, ..., m\}$
	L^a	Number of variables affected by action a
	L	Upper bound on number of affected variables $L \geq \max_{a\in\mathcal{A}} L^a$
	$\mathcal{V}^a$	The set of variables affected by action a
	0	A convenient notation to represent all quantities related to the baseline
	$\mathcal{A}_0$	A notation to include both the actions and the baseline
	r^0	Baseline reward
	r^a_{up}	Expected uplift of action a, $r^a_{\text{up}} = r^a - r^0$
	$r^\star_{\text{up}}$	Largest expected uplift, $r^\star_{\text{up}} = r^{a^\star}_{\text{up}} = \max_{a\in\mathcal{A}} r^a_{\text{up}}$
	y_t	Payoffs observed at round t, $y_t = (y_t(i))_{i\in\mathcal{V}}$
	μ^a	Expected payoffs of action a, $\mu^a = (\mu^a(i))_{i\in\mathcal{V}}$
	μ^0	Baseline payoffs, $\mu^0 = (\mu^0(i))_{i\in\mathcal{V}}$
	μ^a_{up}	Expected individual uplifts of action a, $\mu^a_{\text{up}} = (\mu^a_{\text{up}}(i))_{i\in\mathcal{V}}$
	$\mathcal{P}^a$	The distribution of the payoffs associated to action a
	$\mathcal{P}^0$	Baseline distribution of the payoffs
	Δ_{up}	Lower bound on individual uplift $\forall a \in \mathcal{A},\ i \in \mathcal{V}^a, \lvert\mu^a(i) - \mu^0(i)\rvert \geq \Delta_{\text{up}}$
UCB	N^a_t	Number of times action a is taken in the first t rounds
	$\hat{\mu}^a_t$	Empirical estimates of expected payoffs
	c^a_t	Widths of confidence intervals
	$\mathcal{C}^a_t$	Confidence intervals
	U^a_t	Upper confidence bounds
	τ^a_t	Uplifting index of action a
	ρ^a_t	Individual uplifting indices
	δ'	Error probability provided as input for UCB-type algorithms

Table 2: Main notations used in the paper.

Algorithm 3 UPUCB (b)

Input: Error probability δ', Baseline payoffs μ^0, Sets of affected variables $\{\mathcal{V}^a : a \in \mathcal{A}\}$
Initialization: Take each action once
for $t = K+1, \dots, T$ **do**
  Compute the statistics following (3)
  For $a \in \mathcal{A}$, compute uplifting index $\tau_t^a \leftarrow \sum_{i \in \mathcal{V}^a} (\hat{\mu}_{t-1}^a(i) + c_{t-1}^a - \mu^0(i))$
  Select action $a_t \in \arg\max_{a \in \mathcal{A}} \tau_t^a$

Algorithm 4 UPUCB-nAff (b)

Input: Error probability δ', Baseline payoffs μ^0, Upper bound L on the number of affected variables
Initialization: Take each action once
for $t = K+1, \dots, T$ **do**
  for $a \in \mathcal{A}$ **do**
    Compute UCBs, uplifting indices, and confidence intervals following (9)
    Set $\widehat{\mathcal{V}}_t^a \leftarrow \{i \in \mathcal{V} : \mu^0(i) \notin \mathcal{C}_t^a(i)\}$
    Set $L_t^a \leftarrow \max(0, L - \mathrm{card}(\widehat{\mathcal{V}}_t^a))$
    Set $\mathcal{L}_t^a \leftarrow \underset{\substack{\mathcal{L} \subseteq \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a \\ \mathrm{card}(\mathcal{L}) = L_t^a}}{\arg\max} \sum_{i \in \mathcal{L}} \rho_t^a(i)$
    Compute uplifting index $\tau_t^a \leftarrow \sum_{i \in \widehat{\mathcal{V}}_t^a \cup \mathcal{L}_t^a} \rho_t^a(i)$
  Select action $a_t \in \arg\max_{a \in \mathcal{A}} \tau_t^a$

B Additional Related Work

In this section we complement our related work section by providing comparison to causal bandits and more detailed discussion on how our work fits into and differs from the semi-bandit framework.

Causal Bandits. The term causal bandits captures the general idea that in a bandit model the generation of reward may be governed by some causal mechanism while the actions correspond to interventions on the underlying causal graph. This idea was first formalized in [25] and subsequently explored in a series of work [27, 28, 30] that cover multiples variants of the problem, including both the minimization of simple and cumulative regrets. Our problem fits into the causal bandit framework by regarding Fig. 1 as a simple causal graph where the variables consist of all the direct causes of the reward and the actions correspond to atomic interventions on the action nodes. The observability of these direct causes are also key assumptions in many of the aforementioned works. Moreover, in the unknown affected case our causal graph is in fact only partially unknown. This is an intriguing problem in causal bandits as argued by [29]. While we have shown improvement of the algorithms for this specific model, a interesting question is whether similar improvement can still be shown in more general causal bandit models under the assumption that each actions only affects a limited number of variables of the causal graph.

Combinatorial Bandits. The conceptual differences between our model and the semi-bandit ones [11, 22, 31] have been discussed in Section 1. The following paragraph thus treats this subject from a more technical viewpoint.

To begin, we provide a way to reduce our model to that of semi-bandits when the sets of affected variables are known. For this, we consider a semi-bandit problem whose ground set contains the following $m + \sum_{a \in \mathcal{A}} L^a$ items

- For each variable $i \in \mathcal{V}$, we have an item $e_{0,i}$ with mean $\mu^0(i)$. This item models the reward of variable i when it is unaffected by the taken action.
- For each action $a \in \mathcal{A}$ and $i \in \mathcal{V}^a$, we have an item $e_{a,i}$ with mean $\mu^a(i)$. This item models the reward of variable i when affected by action a.

Then, when action a is taken, we observe the stochastic reward of $e_{a,i}$ for $i \in \mathcal{V}^a$, and that of $e_{0,i}$ for $i \notin \mathcal{V}^a$. The total reward is the sum of the observed individual rewards.

In terms of algorithm, let us consider the UPUCB algorithm described in Algorithm 1, which is based on selecting the action with the largest uplift index $\tau_t^a = \sum_{i \in \mathcal{V}^a} (U_t^a(i) - U_t^0(i))$. We notice that the quantity $U_t^a = \sum_{i \in \mathcal{V}} U_t^a(i) = \tau_t^a + \sum_{i \in \mathcal{V}} U_t^0(i)$ is indeed an upper confidence bound on the reward of action a and UPUCB is equivalent to choosing the action with the largest U_t^a in each round. This observation effectively reduces UPUCB to CombUCB [11, 22] on the transformed model. Nonetheless, the UPUCB formulation has several advantages:

1. It captures the fact that we really care about is the uplifting effect and the variables affected by each action.
2. The uplifting interpretation facilitates algorithm design in more involved situations as we have shown in Sections 6 and 7.
3. While the general analysis of CombUCB yields a regret in $\mathcal{O}((m^2 + mKL) \log T/\Delta)$, we have shown in Theorem 1 that the regret of UPUCB is in $\mathcal{O}(KL^2 \log T/\Delta)$. Our analysis presented in Appendix C.2 also greatly benefits from the UPUCB formulation of the algorithm.

Finally, it is worth noticing that one of the most important contribution of our work is to consider the situations where the sets of affected variables are unknown. In the transformed model, this means that we do not know the composition of each action. We are not aware of any works that address this situation in the combinatorial bandit literature.

C Proofs for Upper Bound Results

In this section, we provide proofs of the regret bounds from Sections 5 and 6. Our proofs have the following structure.

1. Leveraging Lemmas 2 and 3 presented in Appendix C.1, we define a high-probability event $\mathcal{E}$ on which all the expected payoffs belong to their confidence intervals.
2. Using the fact that the uplifting index of the taken action is larger than that of any optimal action by design, we show that on event $\mathcal{E}$ a suboptimal action can only be taken when its associated width of confidence interval is sufficiently large.
3. The previous point allows us to bound the number of times that a suboptimal action is taken by $\tilde{\mathcal{O}}(L^2/(\Delta^a)^2)$ on $\mathcal{E}$.
4. We conclude with the following regret decomposition (which was already stated in (2)).

$$\mathrm{Reg}_T = \sum_{a \in \mathcal{A}} N_t^a \Delta^a. \tag{11}$$

C.1 Concentration Bounds

To prove the regret upper bounds, we will make use of concentration bounds that apply to quantities defined in (3) and (5). Underlying all these results is the following fundamental concentration inequality for σ-sub-Gaussian random variables [36]

$$\mathbb{P}(|X| \geq \gamma) \leq 2 \exp(-\gamma^2/(2\sigma^2)) \text{ for all } \gamma > 0. \tag{12}$$

However, the estimates in (3) and (5) aggregate observed values in a way that depend on the decisions that have been made. Therefore, concentration bounds for the sum of independent variables, which can be directly derived from (12), are not sufficient for our purposes. The standard trick to circumvent this issue in bandit literature is to combine Doob's optional skipping (Doob 16, Chap. 3, Chung and Zhong 12, Chap. 9) with concentration inequality for martingale difference sequence. The following lemma adapted from [35, Lem. A.1] is proved exactly using this technique.

Lemma 1. *Let $(\mathcal{F}_t)_{t \in \mathbb{N}}$ be a filtration and let $(\epsilon_t, X_t)_{t \in \mathbb{N}}$ be a sequence of $\{0,1\} \times \mathbb{R}$- valued random variables such that ϵ_t is $\mathcal{F}_{t-1}$-measurable, and X_t is $\mathcal{F}_t$-measurable and conditionally σ-sub-Gaussian, i.e., $\mathbb{E}[\exp(\gamma X_t) \mid \mathcal{F}_{t-1}] \leq \exp(\gamma^2\sigma^2/2)$ for all $\gamma \in \mathbb{R}$. Let $N > 0$ and let $\tau = \min\{t \geq 1 : \sum_{s=1}^t \epsilon_s = N\}$, where we take $\tau = +\infty$ when $\sum_{s=1}^{+\infty} \epsilon_s < N$. Then, for any $\delta \in (0,1)$,*

$$\mathbb{P}\left(\left|\sum_{t=1}^{\tau} \epsilon_t X_t\right| \geq \sigma\sqrt{2N \log(2/\delta)}\right) \leq \delta. \tag{13}$$

Proof. The proof of [35, Lem. A.1] essentially applies. We just need to replace the use of Hoeffding-Azuma inequality by the use of concentration inequality for conditionally sub-Gaussian martingale difference sequences, which is itself derived from (12) and the fact that the sum of conditionally sub-Gaussian martingale differences is sub-Gaussian. □

In the following, $(\mathcal{F}_t)_t$ will represent the natural filtration associated to $(y_t, a_{t+1})_t$ such that a_t is $\mathcal{F}_{t-1}$-measurable, and y_t is $\mathcal{F}_t$-measurable. Applying Lemma 1 to properly defined $(\epsilon_t, X_t)_t$ along with the use of a union bound gives the following concentration results that are key to our analyses.

Lemma 2. *Let $a \in \mathcal{A}$, $i \in \mathcal{V}$, and let $\mu_t^a(i)$ be defined as in (3). Then for any $\delta \in (0,1)$, it holds*

$$\mathbb{P}\left(\exists\, t \in T,\, |\hat{\mu}_t^a(i) - \mu^a(i)| \geq \sqrt{\frac{2\log(2/\delta)}{N_t^a}}\right) \leq T\delta.$$

Proof. For $N \in [T]$, we define the random variables

$$\tau_N = \min\left\{t \geq 1 : \sum_{s=1}^{t} \mathbb{1}\{a_s = a\} = N\right\} \quad \text{and} \quad S_N^a(i) = \sum_{s=1}^{\tau_N} y_s(i)\, \mathbb{1}\{a_s = a\}.$$

We then define the failure events

$$\mathcal{E}(i) = \left\{\exists\, t \in [T],\, |\hat{\mu}_t^a(i) - \mu^a(i)| \geq \sqrt{\frac{2\log(2/\delta)}{N_t^a}}\right\},$$

$$\widetilde{\mathcal{E}}(i) = \left\{\exists\, N \in [T],\, \left|S_N^a(i) - \mu^a(i)\left(\sum_{s=1}^{\tau_N} \mathbb{1}\{a_s = a\}\right)\right| \geq \sqrt{2N\log(2/\delta)}.\right\}$$

It holds $\mathcal{E}(i) \subseteq \widetilde{\mathcal{E}}(i)$. In fact, if $\mathcal{E}(i)$ happens then $|\hat{\mu}_t^a(i) - \mu^a(i)| \geq \sqrt{2\log(2/\delta)/N_t^a(i)}$ for some t and this is only possible when $N_t^a(i) \geq 1$. With the definition $N_t^a(i) = \sum_{s=1}^{t} \mathbb{1}\{a_s = a\}$, we further see that $N_t^a(i) \leq t \leq T$ and $\tau_{N_t^a} \leq t < +\infty$. The latter means that we indeed have $\sum_{s=1}^{\tau_{N_t^a}} \mathbb{1}\{a_s = a\} = N_t^a$. Thus multiplying $|\hat{\mu}_t^a(i) - \mu^a(i)| \geq \sqrt{2\log(2/\delta)/N_t^a(i)}$ by $N = N_t^a \in [T]$ gives exactly

$$\left|S_N^a(i) - \mu^a(i)\left(\sum_{s=1}^{\tau_N} \mathbb{1}\{a_s = a\}\right)\right| \geq \sqrt{2N\log(2/\delta)},$$

which shows that $\widetilde{\mathcal{E}}(i)$ happens as well. It is thus sufficient to prove that $\mathbb{P}(\widetilde{\mathcal{E}}(i)) \leq T\delta$.

For $N \in [T]$, we apply Lemma 1 to $\epsilon_t \leftarrow \mathbb{1}\{a_t = a\}$ and $X_t \leftarrow \xi_t^a(i) = y_t^a(i) - \mu^{a_t}(i)$. Since $\mu^{a_t}(i)\, \mathbb{1}\{a_t = a\} = \mu^a(i)\, \mathbb{1}\{a_t = a\}$, we recover the exact inequality that appears in the definition of $\widetilde{\mathcal{E}}(i)$. We conclude with a union bound taking N ranging from 1 to T. □

Lemma 3. *Let $i \in \mathcal{V}$, and let $\mu_t^0(i)$ be defined as in (5). Then for any $\delta \in (0,1)$, it holds*

$$\mathbb{P}\left(\exists\, t \in T,\, |\hat{\mu}_t^0(i) - \mu^0(i)| \geq \sqrt{\frac{2\log(2/\delta)}{N_t^0(i)}}\right) \leq T\delta.$$

Proof. Lemma 3 is proved in the same way as Lemma 2. In particular, we apply Lemma 1 to $\epsilon_t \leftarrow \mathbb{1}\{i \notin \mathcal{V}^{a_t}\}$ and $X_t \leftarrow \xi_t^0(i)$. □

C.2 Unknown Baseline

Theorem 1. *Let $\delta' = \delta/(4KLT)$. Then the regret of* UPUCB *(Algorithm 1) with probability at least $1-\delta$, satisfies:*

$$\mathrm{Reg}_T \leq \sum_{a \in \mathcal{A}:\Delta^a > 0} \left(\frac{8(L^a + L^{a^\star})^2 \log(4KLT/\delta)}{\Delta^a} + \Delta^a\right). \tag{7}$$

Proof. We first notice that while there are m variables, not all of them are involved in estimating the uplifts. Therefore, we only need to focus on the set of relevant variables $\mathcal{V}' = \bigcup_{a\in\mathcal{A}} \mathcal{V}^a$. Clearly, $\mathrm{card}(\mathcal{V}') \leq KL$. With this in mind, we define the events

$$\mathcal{E}_1 = \{\forall t \in [T],\ \forall a \in \mathcal{A},\ \forall i \in \mathcal{V}^a,\ |\hat{\mu}_t^a(i) - \mu^a(i)| \leq c_t^a\},$$
$$\mathcal{E}_2 = \{\forall t \in [T],\ \forall i \in \mathcal{V}',\ |\hat{\mu}_t^0(i) - \mu^0(i)| \leq c_t^0(i)\}.$$

Lemma 2 and Lemma 3 in Appendix C.1 along with union bounds over a and i guarantee

$$\mathbb{P}(\mathcal{E}_1 \cup \mathcal{E}_2) \geq 1 - \left(\sum_{a\in\mathcal{A}} L^a + \mathrm{card}(\mathcal{V}')\right) T(2\delta') \geq 1 - 4KLT\delta' = 1 - \delta.$$

It is thus sufficient to show that (7) holds on $\mathcal{E}_1 \cup \mathcal{E}_2$.

In the remainder of the proof, we consider an arbitrary realization of $\mathcal{E}_1 \cup \mathcal{E}_2$ and prove that (7) indeed holds for this realization. From (11) it is clear that we just need to bound $N_T^a = \sum_{t=1}^T \mathbb{1}\{a_t = a\}$ from above for all suboptimal action $a \in \mathcal{A}$ with $\Delta^a > 0$. This is done by providing a lower bound on c_t^a for any t such that $a_t = a$. To proceed, suppose that a suboptimal action a is taken for the last time at round $t \in \{K+1, ..., T\}$.[7] This implies $\tau_t^a \geq \tau_t^{a^\star}$ and $N_T^a = N_{t-1}^a + 1$. By the definition of the uplifting indices, we get

$$\sum_{i\in\mathcal{V}^a} (U_t^a(i) - U_t^0(i)) \geq \sum_{i\in\mathcal{V}^{a^\star}} (U_t^{a^\star}(i) - U_t^0(i)).$$

Rearranging and dropping $\sum_{i\in\mathcal{V}^a\cap\mathcal{V}^{a^\star}} U_t^0(i)$ from both sides, we get

$$\sum_{i\in\mathcal{V}^a} U_t^a(i) + \sum_{i\in\mathcal{V}^{a^\star}\setminus\mathcal{V}^a} U_t^0(i) \geq \sum_{i\in\mathcal{V}^{a^\star}} U_t^{a^\star}(i) + \sum_{i\in\mathcal{V}^a\setminus\mathcal{V}^{a^\star}} U_t^0(i). \tag{14}$$

This is similar to the type of inequality that defines standard UCB. In fact, the left and the right hand sides of (14) are respectively upper confidence bounds on $\sum_{i\in\mathcal{V}^a\cup\mathcal{V}^{a^\star}} \mu^a(i)$ and upper confidence bounds on $\sum_{i\in\mathcal{V}^a\cup\mathcal{V}^{a^\star}} \mu^{a^\star}(i)$. We next bound these two quantities respectively from above and from below. Since both $\mathcal{E}_1$ and $\mathcal{E}_2$ hold, we have the following inequalities for the UCB indices

$$\forall s \in [T],\ \forall a \in \mathcal{A},\ \forall i \in \mathcal{V}^a,\ \mu^a(i) + 2c_{s-1}^a \geq U_s^a(i) \geq \mu^a(i),$$
$$\forall s \in [T],\ \forall i \in \mathcal{V}' \setminus \bigcap_{a\in\mathcal{A}} \mathcal{V}^a,\ \mu^0(i) + 2c_{s-1}^0(i) \geq U_s^0(i) \geq \mu^0(i).$$

Combining the above, we obtain

$$\sum_{i\in\mathcal{V}^a} (\mu^a(i) + 2c_{t-1}^a) + \sum_{i\in\mathcal{V}^{a^\star}\setminus\mathcal{V}^a} (\mu^0(i) + 2c_{t-1}^0(i)) \geq \sum_{i\in\mathcal{V}^{a^\star}} \mu^{a^\star}(i) + \sum_{i\in\mathcal{V}^a\setminus\mathcal{V}^{a^\star}} \mu^0(i).$$

Adding $\sum_{i\in\mathcal{V}^a\cap\mathcal{V}^{a^\star}} \mu^0(i)$ to both sides and rearranging leads to

$$\sum_{i\in\mathcal{V}^{a^\star}} \mu^{a^\star}(i) - \sum_{i\in\mathcal{V}^a} \mu^a(i) + \sum_{i\in\mathcal{V}^a} \mu^0(i) - \sum_{i\in\mathcal{V}^{a^\star}} \mu^0(i) \leq \sum_{i\in\mathcal{V}^a} 2c_{t-1}^a + \sum_{i\in\mathcal{V}^{a^\star}\setminus\mathcal{V}^a} 2c_{t-1}^0(i).$$

From this we can derive an inequality between the suboptimal gap Δ^a and the the widths of the confidence intervals as follows

$$\begin{aligned}
\Delta^a &= r_{\text{up}}^{a^\star} - r_{\text{up}}^a \\
&= \sum_{i\in\mathcal{V}^{a^\star}} (\mu^{a^\star}(i) - \mu^0(i)) - \sum_{i\in\mathcal{V}^a} (\mu^a(i) - \mu^0(i)) \\
&= \sum_{i\in\mathcal{V}^{a^\star}} \mu^{a^\star}(i) - \sum_{i\in\mathcal{V}^a} \mu^a(i) + \sum_{i\in\mathcal{V}^a} \mu^0(i) - \sum_{i\in\mathcal{V}^{a^\star}} \mu^0(i) \\
&\leq 2L^a c_{t-1}^a + \sum_{i\in\mathcal{V}^{a^\star}\setminus\mathcal{V}^a} 2c_{t-1}^0(i).
\end{aligned} \tag{15}$$

[7] If the action is not taken anymore after the first T runs. We have $N_T^a = 1$ so the upper bound that we show later trivially holds.

We next argue that for all $i \in \mathcal{V}^{a^\star} \setminus \mathcal{V}^a$ we have $N^0_{t-1}(i) \geq N^a_{t-1}$ and hence $c^0_{t-1}(i) \leq c^a_{t-1}$. In fact, if $i \notin \mathcal{V}^a$, then whenever action a is taken, the count of $N^0_s(i)$ also increases by 1. Formally,

$$N^0_{t-1}(i) = \sum_{s=1}^{t-1} \mathbb{1}\{i \notin \mathcal{V}^{a_s}\} \geq \sum_{s=1}^{t-1} \mathbb{1}\{a_s = a\}\, \mathbb{1}\{i \notin \mathcal{V}^{a_s}\} = \sum_{s=1}^{t-1} \mathbb{1}\{a_s = a\} = N^a_{t-1}.$$

Therefore, we further deduce $\Delta^a \leq 2(L^a + L^{a^\star})c^a_{t-1}$. Equivalently,

$$N^a_{t-1} \leq \frac{8(L^a + L^{a^\star})^2 \log(1/\delta')}{(\Delta^a)^2}.$$

We conclude that action a is taken at most $\lfloor 8(L^a + L^{a^\star})^2 \log(1/\delta')/(\Delta^a)^2 \rfloor + 1$ times during the execution of the algorithm. Plugging this into (11) gives exactly (7). □

C.3 Unknown Affected Variables

Below we prove Theorem 2 separately for UPUCB-nAff (b) and UPUCB-nAff.

Proposition 3 (Regret of UPUCB-nAff (b))**.** *Let $\delta' = \delta/(2KmT)$. If the learner knows and an upper bound L on maximum number of variables an action affects and the baseline payoffs, then* UPUCB-nAff (b) *(Algorithm 4), with probability at least $1-\delta$, has regret bounded by:*

$$\mathrm{Reg}_T \leq \sum_{a \in \mathcal{A}: \Delta^a > 0} \left(\frac{32L^2 \log(2KmT/\delta)}{\Delta^a} + \Delta^a \right). \tag{16}$$

Proof. Let us consider the event

$$\mathcal{E} = \{\forall t \in [T],\ \forall a \in \mathcal{A},\ \forall i \in \mathcal{V},\ |\hat{\mu}^a_t(i) - \mu^a(i)| < c^a_t\}, \tag{17}$$

By Lemma 3 and a union bound we get immediately $\mathbb{P}(\mathcal{E}) \geq 1 - KmT(2\delta') = 1 - \delta$. In the following, we assume that $\mathcal{E}$ occurs and prove (16).

First, by the definition of $\widehat{\mathcal{V}}^a_t$ we have clearly $\widehat{\mathcal{V}}^a_t \subseteq \mathcal{V}^a$. In fact, if $i \notin \mathcal{V}^a$ then $\mu^0(i) = \mu^a(i) \in \mathcal{C}^a_t(i)$. The inclusion $\widehat{\mathcal{V}}^a_t \subseteq \mathcal{V}^a$ along with the condition $\mathrm{card}(\mathcal{V}^a) \leq L$ imply $\mathrm{card}(\mathcal{V}^a \setminus \widehat{\mathcal{V}}^a_t) \leq L - \mathrm{card}(\widehat{\mathcal{V}}^a_t)$; in particular, $L - \mathrm{card}(\widehat{\mathcal{V}}^a_t) \geq 0$ and thus $L^a_t = L - \mathrm{card}(\widehat{\mathcal{V}}^a_t)$.

Next, for any $i \in \mathcal{V} \setminus \widehat{\mathcal{V}}^a_t$, it holds $\mu^0(i) \in \mathcal{C}^a_t(i)$. This implies $\mu^0(i) \leq U^a_t(i)$, and accordingly $\rho^a_t(i) \geq 0$. We can then write

$$\sum_{i \in \mathcal{V}^a \setminus \widehat{\mathcal{V}}^a_t} \rho^a_t(i) \leq \max_{\substack{\mathcal{L} \subseteq \mathcal{V} \setminus \widehat{\mathcal{V}}^a_t \\ \mathrm{card}(\mathcal{L}) = \mathrm{card}(\mathcal{V} \setminus \widehat{\mathcal{V}}^a_t)}} \sum_{i \in \mathcal{L}} \rho^a_t(i) \leq \max_{\substack{\mathcal{L} \subseteq \mathcal{V} \setminus \widehat{\mathcal{V}}^a_t \\ \mathrm{card}(\mathcal{L}) = L - \mathrm{card}(\widehat{\mathcal{V}}^a_t)}} \sum_{i \in \mathcal{L}} \rho^a_t(i) = \sum_{i \in \mathcal{L}^a_t} \rho^a_t(i).$$

Since $U^a_t(i) \geq \mu^a(i)$ on $\mathcal{E}$, we have $\rho^a_t(i) \geq \mu^a_{\mathrm{up}}(i)$; subsequently

$$r^a_{\mathrm{up}} = \sum_{i \in \mathcal{V}^a} \mu^a_{\mathrm{up}}(i) = \sum_{i \in \widehat{\mathcal{V}}^a_t} \mu^a_{\mathrm{up}}(i) + \sum_{i \in \mathcal{V}^a \setminus \widehat{\mathcal{V}}^a_t} \mu^a_{\mathrm{up}}(i) \leq \sum_{i \in \widehat{\mathcal{V}}^a_t} \rho^a_t(i) + \sum_{i \in \mathcal{L}^a_t} \rho^a_t(i) = \tau^a_t.$$

This shows that τ^a_t is effectively an upper bound of r^a_{up}. With this in mind, we are ready to bound the number of times that a suboptimal action is taken.

Let $a \in \mathcal{A}$ with $\Delta^a > 0$ and assume that it is taken at round $t \in \{K+1, ..., T\}$, which means that $\tau^a_t \geq \tau^{a^\star}_t$. We have shown that $\tau^{a^\star}_t \geq r^{a^\star}_{\mathrm{up}}$. It remains to provide an upper bound for τ^a_t that involves r^a_{up}. For this, we again use $\widehat{\mathcal{V}}^a_t \subseteq \mathcal{V}^a$ and the fact that $\rho^a_t(i) \geq 0$ for all $i \in \mathcal{V} \setminus \widehat{\mathcal{V}}^a_t$ to decompose

$$\tau^a_t = \sum_{i \in \widehat{\mathcal{V}}^a_t \cup \mathcal{L}^a_t} \rho^a_t(i) \leq \sum_{i \in \mathcal{V}^a \cup \mathcal{L}^a_t} \rho^a_t(i) = \sum_{i \in \mathcal{V}^a} \rho^a_t(i) + \sum_{i \in \mathcal{L}^a_t \setminus \mathcal{V}^a} \rho^a_t(i). \tag{18}$$

The summation of the individual uplifting indices over $i \in \mathcal{V}^a$ is naturally related to r^a_{up}. Since $\mu^a(i) \geq \hat{\mu}^a_{t-1}(i) - c^a_{t-1}$, for all $i \in \mathcal{V}$ it holds

$$\rho^a_t(i) = U^a_t(i) - \mu^a(i) \leq \mu^a(i) - \mu^0(i) + 2c^a_{t-1} = \mu^a_{\mathrm{up}}(i) + 2c^a_{t-1}.$$

In consequence,

$$\sum_{i\in\mathcal{V}^a} \rho_t^a(i) \le \sum_{i\in\mathcal{V}^a} (\mu_{\text{up}}^a(i) + 2c_{t-1}^a) \le r_{\text{up}}^a + 2Lc_{t-1}^a. \tag{19}$$

What we need to show next is that the definition of $\widehat{\mathcal{V}}_t^a$ guarantees $\rho_t^a(i)$ to be small whenever $i \notin \widehat{\mathcal{V}}_t^a$. In fact, if $i \in \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a$, we have $\mu^0(i) \in \mathcal{C}_t^a(i)$, and in particular $\mu^0(i) \ge \hat{\mu}_{t-1}^a(i) - c_{t-1}^a$, implying that

$$\rho_t^a(i) = U_t^a(i) - \mu^0(i) \le (\hat{\mu}_{t-1}^a(i) + c_{t-1}^a) - (\hat{\mu}_{t-1}^a(i) - c_{t-1}^a) = 2c_{t-1}^a.$$

By definition, $\mathcal{L}_t^a \subseteq \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a$, and hence

$$\sum_{i\in\mathcal{L}_t^a\setminus\mathcal{V}^a} \rho_t^a(i) \le \sum_{i\in\mathcal{L}_t^a\setminus\mathcal{V}^a} 2c_{t-1}^a \le 2Lc_{t-1}^a. \tag{20}$$

Putting (18), (19), and (20) together, we get

$$\tau_t^a \le \sum_{i\in\mathcal{V}^a} \rho_t^a(i) + \sum_{i\in\mathcal{L}_t^a\setminus\mathcal{V}^a} \rho_t^a(i) \le r_{\text{up}}^a + 4Lc_{t-1}^a.$$

With $\tau_t^a \ge \tau_t^{a^\star} \ge r_{\text{up}}^{a^\star}$ we deduce

$$r_{\text{up}}^{a^\star} \le r_{\text{up}}^a + 4Lc_{t-1}^a.$$

Therefore, $\Delta^a = r_{\text{up}}^{a^\star} - r_{\text{up}}^a \le 4Lc_{t-1}^a$, and equivalently

$$N_{t-1}^a \le \frac{32L^2\log(1/\delta')}{(\Delta^a)^2}.$$

Following the argument in the proof of Theorem 1, this translates into an upper bound on N_T^a and we conclude by invoking the decomposition of (11). □

We notice that it is quite straightforward to adapt UPUCB-nAff (b) and its analysis to the setting where the learner knows an action-dependent upper bound $\overline{L^a} \ge L^a$. The only change in (16) is that L is replaced by $\overline{L^a}$. Thus we recover (4) up to a multiplicative constant when $\overline{L^a} = \Theta(L^a)$. We complete this section with the proof of the regret bound for UPUCB-nAff. It combines ideas from the proofs of Theorem 1 and Proposition 3.

Proposition 4 (Regret of UPUCB-nAff)**.** *Let $\delta' = \delta/(2KmT)$. Then the regret of* UPUCB-nAff *(Algorithm 2), with probability at least $1-\delta$, satisfies:*

$$\text{Reg}_T \le \sum_{a\in\mathcal{A}:\Delta^a>0} \left(\frac{512L^2\log(2KmT/\delta)}{\Delta^a} + \Delta^a \right), \tag{21}$$

Proof. We again place ourselves in the event $\mathcal{E}$ defined in (17), and prove (21) under this condition. Recall that $b_t \in \arg\max_{a\in\mathcal{A}} N_{t-1}^a$ is an action that is taken most frequently during the first $t-1$ rounds and the associated estimates serve as baselines at round t. We define $\widetilde{\mathcal{V}}_t^a = \mathcal{V}^a \cup \mathcal{V}^{b_t}$ as the set of variables that are affected by either action a or b_t. Clearly, $\text{card}(\widetilde{\mathcal{V}}_t^a) \le 2L$.

We first show that $\widehat{\mathcal{V}}_t^a \subseteq \widetilde{\mathcal{V}}_t^a$ whenever $\mathcal{E}$ occurs. Let $i \in \mathcal{V} \setminus \widetilde{\mathcal{V}}_t^a$. Under $\mathcal{E}$ we have both $\mu^0(i) = \mu^a(i) \in \mathcal{C}_t^a(i)$ and $\mu^0(i) = \mu^{b_t}(i) \in \mathcal{C}_t^b(i)$. Therefore, $\mathcal{C}_t^a(i) \cap \mathcal{C}_t^{b_t}(i) \ne \varnothing$; subsequently $i \notin \widehat{\mathcal{V}}_t^a$. The contrapositive of the above gives exactly $\widehat{\mathcal{V}}_t^a \subseteq \widetilde{\mathcal{V}}_t^a$.

We shall now bound N_T^a for any suboptimal action a, from which the proof concludes by applying (11). Let $a \in \mathcal{A}$ satisfying $\Delta^a > 0$. If it is taken at round $t \in \{K+1, ..., T\}$, we have $\tau_t^a \ge \tau_t^{a^\star}$. This can be written as

$$\sum_{i\in\widehat{\mathcal{V}}_t^a} \rho_t^a(i) + \sum_{i\in\mathcal{L}_t^a} \rho_t^a(i) \ge \sum_{i\in\widehat{\mathcal{V}}_t^{a^\star}} \rho_t^{a^\star}(i) + \sum_{i\in\mathcal{L}_t^{a^\star}} \rho_t^{a^\star}(i). \tag{22}$$

Decomposition of Suboptimality Gap using the New Baseline. Similar to before, we will use (22) to relate the widths of the confidence intervals to the suboptimality gap Δ^a. For this, we notice that the suboptimality gap of a can be rewritten as

$$\begin{aligned}\Delta^a &= \sum_{i\in\mathcal{V}} \mu^{a^\star}(i) - \sum_{i\in\mathcal{V}} \mu^a(i) \\ &= \sum_{i\in\mathcal{V}} (\mu^{a^\star}(i) - \mu^{b_t}(i)) - \sum_{i\in\mathcal{V}} (\mu^a(i) - \mu^{b_t}(i)) \\ &= \sum_{i\in\widetilde{\mathcal{V}}_t^{a^\star}} (\mu^{a^\star}(i) - \mu^{b_t}(i)) - \sum_{i\in\widetilde{\mathcal{V}}_t^{a}} (\mu^a(i) - \mu^{b_t}(i)) \end{aligned} \tag{23}$$

The last equality is true because $\mu^{a^\star}(i) = \mu^0(i)$ outside $\widetilde{\mathcal{V}}_t^{a^\star}$, and the same holds when we replace $a^\star$ by a. Therefore, our goal is to lower bound the right-hand side of (22) by an expression that contains $\sum_{i\in\widetilde{\mathcal{V}}_t^{a^\star}} (\mu^{a^\star}(i) - \mu^{b_t}(i))$ and upper bound the left-hand side of (22) by an expression that contains $\sum_{i\in\widetilde{\mathcal{V}}_t^{a}} (\mu^a(i) - \mu^{b_t}(i))$.

Lower Bound on RHS of (22). Since $\sum_{i\in\mathcal{L}_t^{a^\star}} \rho_t^{a^\star}(i)$ is the maximum padding, we can lower bound the right-hand side of the inequality by $\sum_{i\in\widetilde{\mathcal{V}}_t^{a^\star}} \rho_t^{a^\star}(i)$. Formally, with $\widehat{\mathcal{V}}_t^{a^\star} \subseteq \widetilde{\mathcal{V}}_t^{a^\star}$ we have $\mathrm{card}(\widetilde{\mathcal{V}}_t^{a^\star} \setminus \widehat{\mathcal{V}}_t^{a^\star}) \le 2L - \mathrm{card}(\widehat{\mathcal{V}}_t^{a^\star})$ and $L_t^{a^\star} = 2L - \mathrm{card}(\widehat{\mathcal{V}}_t^{a^\star})$. Thus, by definition of $\mathcal{L}_t^{a^\star}$,

$$\sum_{i\in\mathcal{L}_t^{a^\star}} \rho_t^{a^\star}(i) = \max_{\substack{\mathcal{L}\subseteq\mathcal{V}^a\setminus\widehat{\mathcal{V}}_t^a \\ \mathrm{card}(\mathcal{L})\le 2L-\mathrm{card}(\widehat{\mathcal{V}}_t^{a^\star})}} \sum_{i\in\mathcal{L}} \rho_t^{a^\star}(i) \ge \sum_{i\in\widetilde{\mathcal{V}}_t^{a^\star}\setminus\widehat{\mathcal{V}}_t^{a^\star}} \rho_t^{a^\star}(i).$$

To further provide a lower bound on $\sum_{i\in\widetilde{\mathcal{V}}_t^{a^\star}} \rho_t^{a^\star}(i)$, we recall that $\rho_t^{a^\star}(i) = U_t^{a^\star}(i) - U_t^{b_t}(i)$, so with $U_t^{a^\star}(i) \ge \mu^{a^\star}(i)$, $U_t^{b_t}(i) \le \mu^{b_t}(i) + 2c_{t-1}^{b_t}$, and $\mathrm{card}(\widetilde{\mathcal{V}}_t^{a^\star}) \le 2L$, we get

$$\begin{aligned}\sum_{i\in\widehat{\mathcal{V}}_t^{a^\star}} \rho_t^{a^\star}(i) + \sum_{i\in\mathcal{L}_t^{a^\star}} \rho_t^{a^\star}(i) &\ge \sum_{i\in\widetilde{\mathcal{V}}_t^{a^\star}} \rho_t^{a^\star}(i) \\ &\ge \sum_{i\in\widetilde{\mathcal{V}}_t^{a^\star}} (\mu^{a^\star}(i) - \mu^{b_t}(i) - 2c_{t-1}^{b_t}) \\ &\ge \sum_{i\in\widetilde{\mathcal{V}}_t^{a^\star}} (\mu^{a^\star}(i) - \mu^{b_t}(i)) - 4Lc_{t-1}^{b_t}. \end{aligned} \tag{24}$$

Upper Bound on LHS of (22). To make the term $\sum_{i\in\widetilde{\mathcal{V}}_t^{a}} (\mu^a(i) - \mu^{b_t}(i))$ appear, we first relate it to the sum of the uplifting indices using $U_t^a(i) \le \mu^a(i) + 2c_{t-1}^a$.

$$\sum_{i\in\widetilde{\mathcal{V}}_t^{a}} \rho_t^a(i) \le \sum_{i\in\widetilde{\mathcal{V}}_t^{a}} (\mu^a(i) + 2c_{t-1}^a - \mu^{b_t}(i)) \le \sum_{i\in\widetilde{\mathcal{V}}_t^{a}} (\mu^a(i) - \mu^{b_t}(i)) + 4Lc_{t-1}^a. \tag{25}$$

The last inequality uses $\mathrm{card}(\widetilde{\mathcal{V}}_t^a) \le 2L$. Next, to relate $\sum_{i\in\widetilde{\mathcal{V}}_t^{a}} \rho_t^a(i)$ to the left-hand side of (22), we want to show (*i*) $\sum_{i\in\mathcal{L}_t^{a}} \rho_t^a(i)$ is small, and (*ii*) We can add additional terms from $\widetilde{\mathcal{V}}_t^a \setminus \widehat{\mathcal{V}}_t^a$ to the left-hand side of (22). These two points correspond respectively to inequalities (20) and (18) in the proof of Proposition 3. Again, this can be done by looking at the definition of $\widehat{\mathcal{V}}_t^a$. In fact, if $i \in \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a$, then by definition $\mathcal{C}_t^a(i) \cap \mathcal{C}_t^{b_t}(i) \neq \varnothing$, implying that

$$\begin{aligned}\hat{\mu}_{t-1}^{b_t}(i) + c_{t-1}^{b_t} &\ge \hat{\mu}_{t-1}^{a}(i) - c_{t-1}^{a}, \\ \hat{\mu}_{t-1}^{a}(i) + c_{t-1}^{a} &\ge \hat{\mu}_{t-1}^{b_t}(i) - c_{t-1}^{b_t}.\end{aligned}$$

The above two inequalities respectively translate into

$$\begin{aligned}U_t^a(i) - U_t^{b_t}(i) &\le 2c_{t-1}^a, \\ U_t^a(i) - U_t^{b_t}(i) &\ge -2c_{t-1}^{b_t},\end{aligned}$$

In other words, for any $i \in \mathcal{V} \setminus \widehat{\mathcal{V}}_t^a$ we have $-2c_{t-1}^{b_t} \leq \rho_t^a(i) \leq 2c_{t-1}^a$. Subsequently,

$$\sum_{i \in \mathcal{L}_t^a} \rho_t^a(i) \leq \sum_{i \in \mathcal{L}_t^a} 2c_{t-1}^a, \tag{26}$$

$$\sum_{i \in \widetilde{\mathcal{V}}_t^a \setminus \widehat{\mathcal{V}}_t^a} -2c_{t-1}^{b_t} \leq \sum_{i \in \widetilde{\mathcal{V}}_t^a \setminus \widehat{\mathcal{V}}_t^a} \rho_t^a(i). \tag{27}$$

Therefore

$$\begin{aligned}
\sum_{i \in \widehat{\mathcal{V}}_t^a} \rho_t^a(i) + \sum_{i \in \mathcal{L}_t^a} \rho_t^a(i) &\leq \sum_{i \in \widehat{\mathcal{V}}_t^a} \rho_t^a(i) + \sum_{i \in \mathcal{L}_t^a} 2c_{t-1}^a + \sum_{i \in \widetilde{\mathcal{V}}_t^a \setminus \widehat{\mathcal{V}}_t^a} (\rho_t^a(i) + 2c_{t-1}^{b_t}) \\
&= \sum_{i \in \widetilde{\mathcal{V}}_t^a} \rho_t^a(i) + \sum_{i \in \mathcal{L}_t^a} 2c_{t-1}^a + \sum_{i \in \widetilde{\mathcal{V}}_t^a \setminus \widehat{\mathcal{V}}_t^a} 2c_{t-1}^{b_t} \\
&\leq \sum_{i \in \widetilde{\mathcal{V}}_t^a} \rho_t^a(i) + 4Lc_{t-1}^a + 4Lc_{t-1}^{b_t}.
\end{aligned} \tag{28}$$

The first inequality makes use of (26) and (27), and in particular the last term on the right-hand side is non-negative thanks to (27). The equality in the second line is true because $\widehat{\mathcal{V}}_t^a \subseteq \widetilde{\mathcal{V}}_t^a$. Finally, for the last inequality we use the fact that $\mathrm{card}(\mathcal{L}_t^a) \leq 2L$ and $\mathrm{card}(\widetilde{\mathcal{V}}_t^a) \leq 2L$. Along with (28) we get an upper bound on $\sum_{i \in \widehat{\mathcal{V}}_t^a} \rho_t^a(i) + \sum_{i \in \mathcal{L}_t^a} \rho_t^a(i)$ that involves $\sum_{i \in \widetilde{\mathcal{V}}_t^a} (\mu^a(i) - \mu^{b_t}(i))$.

Conclusion. Combining (22), (24), (25), and (28), we obtain

$$\sum_{i \in \widetilde{\mathcal{V}}_t^a} (\mu^a(i) - \mu^{b_t}(i)) + 8Lc_{t-1}^a + 4Lc_{t-1}^{b_t} \geq \sum_{i \in \widetilde{\mathcal{V}}_t^{a^\star}} (\mu^{a^\star}(i) - \mu^{b_t}(i)) - 4Lc_{t-1}^{b_t}.$$

By the choice of b_t we have $c_{t-1}^a \geq c_{t-1}^{b_t}$. With (23), it follows that $16Lc_{t-1}^a \geq \Delta^a$ and thus

$$N_{t-1}^a \leq \frac{512L^2 \log(1/\delta')}{(\Delta^a)^2}.$$

Repeating the argument in the proof of Theorem 1 (i.e., choose t to be the last time that action a is taken if it is taken more than once and apply decomposition (11)) gives the desired result. □

Theorem 2. *Let $\delta' = \delta/(2KmT)$. Then the regret of* UPUCB-nAff *(Algorithm 2) (resp.* UPUCB-nAff (b)*, Algorithm 4), with probability at least $1 - \delta$, satisfies:*

$$\mathrm{Reg}_T \leq \sum_{a \in \mathcal{A}: \Delta^a > 0} \left(\frac{\alpha L^2 \log(2KmT/\delta)}{\Delta^a} + \Delta^a \right), \tag{10}$$

where $\alpha = 512$ (resp. 32) in the above inequality.

Proof. The proof follows by combining Propositions 3 and 4. □

D Lower Bounds

In this section, we prove Proposition 2 that justifies our key modeling assumptions. In particular, these lower bounds show that there would be a higher regret in absence of any of the following: (*i*) Limited number of affected variables, (*ii*) Side information about variables beyond just reward, and (*iii*) Some prior knowledge on how the variables are affected. Moreover, we also derive matching lower bounds for Proposition 1. We discuss these results in Appendix D.1. Subsequently, we introduce a general information-theoretic lower bounds for bandit problems with similar design in Appendix D.2, based on which we prove the lower bounds in Appendix D.3.

D.1 Lower Bounds for Uplifting Bandits– Statements and Discussion

To begin, we restate Proposition 2 from Section 4 and introduce associated notations. For sake of conciseness, in the following we say that an uplifting bandit has parameters $(K, m, (\Delta^a, L^a)_{a\in\mathcal{A}})$ if the number of actions, the number of variables, the suboptimality gaps, and the number of affected variables of this instance are respectively K, m, $(\Delta^a)_{a\in\mathcal{A}}$, and $(L^a)_{a\in\mathcal{A}}$. We focus on instance-dependent lower-bounds and assume that the learner has full knowledge of the baseline distribution. We also recall that our lower bounds are derived for the algorithms that are *consistent* [24] over the class of 1-sub-Gaussian uplifting bandits, meaning the induced regret of the algorithm on any uplifting bandit with 1-sub-Gaussian noise satisfies $\mathrm{Reg}_t = o(t^p)$ for all $p > 0$.

Proposition 2. *Let π be a consistent algorithm over the class of 1-sub-Gaussian uplifting bandits that at most uses knowledge of $\mathcal{P}^0$, $(\mathcal{V}^a)_{a\in\mathcal{A}}$, and the fact that the noise is 1-sub-Gaussian. Let $K, m > 0$ and sequence $(L^a, \Delta^a)_{1\le a\le K} \in ([m] \times \mathbb{R}_+)^K$ satisfy $\Delta^1 = 0$. Assume either of the following holds.*

(a) $L^a = m$ for all $a \in \mathcal{A}$, so that in the bandits considered below all actions affect all variables.

(b) Only the reward is observed.

(c) The algorithm π does not make use of any prior knowledge about the arms' expected payoffs $(\mu^a)_{a\in\mathcal{A}}$ (in particular, the knowledge of $(\mathcal{V}^a)_{i\in\mathcal{V}}$ is not used by π).

Then, there exists a 1-sub-Gaussian uplifting bandit with parameters $(K, m, (\Delta^a, L^a)_{a\in\mathcal{A}})$ such that the regret induced by π on it satisfies:

$$\liminf_{T\to+\infty} \frac{\mathbb{E}[\mathrm{Reg}_T]}{\log T} \ge \sum_{a\in\mathcal{A}:\Delta^a>0} \frac{2m^2}{\Delta^a}.$$

The conditions (a) and (b) echo assumptions **(I)** and **(II)**. Note that, however, the condition (b) does not imply the necessity of observing all the payoffs. For example, observing the aggregated payoff $\sum_{i\in\mathcal{V}^{a_t}} y_t(i)$ is sufficient for computing the uplifting indices of UPUCB (b). Having observation of the entire payoff vector $(y_t(i))_{i\in\mathcal{V}}$ is particularly helpful for the case of unknown affected variables, as under this condition we can identify the affected variables relatively easily.

Let us now provide some more intuition on the meaning of (c). For this, consider two bandits with exactly the same baseline and noise distributions and very close expected payoffs of the arms. Then, with finite observations, they provide very similar feedback which prevents the learner from distinguishing between the two problems. However, the two bandits may have completely different structures, and in particular, one of them may have all variables affected by all actions since slight perturbation to the baseline would mean that a variable is affected. Therefore, the lower bound from (a) applies to this instance. Now, provided that the learner is doomed to treat the two instances equally, the lower bound applies to the other instance as well. Of course, one can argue that it is unnatural that the regrets of two similar problems differ significantly even when the learner is provided some prior knowledge. We believe this is the drawback of the asymptotic analysis that we look at here, since these differences may indeed be small within a reasonable number of runs. This is exactly related to the problem of misspecified model that we mentioned in Footnote 3.

The next proposition refines the lower-bound of Proposition 2 to show the optimality of Proposition 1.

Proposition 5. *Let π be a consistent algorithm over the class of 1-sub-Gaussian uplifting bandits that at most uses knowledge of $\mathcal{P}^0$, $(\mathcal{V}^a)_{a\in\mathcal{A}}$, and the fact that the noise is 1-sub-Gaussian. Then, for any $K, m > 0$ and sequence $(L^a, \Delta^a)_{1\le a\le K} \in ([m] \times \mathbb{R}_+)^K$ with $\Delta^1 = 0$, there exists a 1-sub-Gaussian uplifting bandit $(K, m, (\Delta^a, L^a)_{a\in\mathcal{A}})$ such that the regret induced by π on it satisfies*

$$\liminf_{T\to+\infty} \frac{\mathbb{E}[\mathrm{Reg}_T]}{\log T} \ge \sum_{a\in\mathcal{A}:\Delta^a>0} \frac{2(L^a)^2}{\Delta^a}. \tag{29}$$

Proposition 5 essentially posits that the regret must scale with the magnitudes of the noises, and the stated lower bound (29) matches an expected version of (4) up to a constant factor. One way to argue this lower bound holds is by reducing a carefully designed uplifting bandits with parameters $(K, m, (\Delta^a, L^a)_{a\in\mathcal{A}})$ to a K-arms bandit with the corresponding noise scales. We will take a different approach below which allows us prove all the lower bounds following the same schema.

Remark 1. Our lower bounds are derived with respect to the worst correlation structure of the noise. It is also possible to obtain lower bounds that depend on the covariance of the noise. For example, if the variables $(y^a(i))_{i\in\mathcal{V}}$ are independent, the lower bound in (29) only scales linearly with L^a.

D.2 A General Information-theoretic Lower Bound

To prove Proposition 2 and Proposition 5, we establish here a general lemma for deriving instance-dependent lower-bounds for bandit problems with underlying variables, side observations, and prior knowledge on the distribution of the variables. In particular, uplifting bandit can be viewed as a special case of this model.

Let $y = (y(1), \dots, y(m))$ be a random vector of m variables that entirely determines the reward so that we can write $r = \varphi(y)$ for some deterministic reward function $\varphi\colon \mathbb{R}^m \to \mathbb{R}$. Let $\mathcal{P}^a$ be the distribution of action a on the underlying variables and $\mathcal{P}^0$ be the baseline distribution on these variables. At each round after taking an action, the learner observes a vector in $\mathbb{R}^d$ that is itself a function of the variables, written as $o(y)$ where $o\colon \mathbb{R}^m \to \mathbb{R}^d$ is the observation function. The decision of which action to take can then only be based on the interaction history $(a_1, o(y_1), \dots, a_{t-1}, o(y_{t-1}))$ and some prior knowledge of the learner about the distributions of the variables $\Upsilon(\mathcal{P}^0, \dots, \mathcal{P}^K)$.[8]

A class of action distributions is defined as $\mathcal{B} = \mathcal{M}^0 \times \cdots \times \mathcal{M}^K$ where each $\mathcal{M}^a$ is a set of distributions on $\mathbb{R}^m$. We say that $\mathcal{B}$ is indistinguishable under Υ if for all $\mathcal{P}_* = (\mathcal{P}^a)_{a\in\mathcal{A}_0}, \mathcal{Q}_* = (\mathcal{Q}^a)_{a\in\mathcal{A}_0} \in \mathcal{B}$ we have $\Upsilon(\mathcal{P}^0, \dots, \mathcal{P}^K) = \Upsilon(\mathcal{Q}^0, \dots, \mathcal{Q}^K)$. A bandit problem with underlying variables, side observations, and prior knowledge on the variable distribution is thus defined by the quadruple $(\mathcal{B}, \varphi, o, \Upsilon)$. A policy π is said to be *consistent* over $(\mathcal{B}, \varphi)$ if for all $p > 0$ and all instance of $(\mathcal{B}, \varphi)$, the regret of π on the instance satisfies $\mathrm{Reg}_t = o(t^p)$. A policy π is compatible with observation o and prior knowledge Υ if it can be implemented by a learner that observes $o(y)$ and has prior knowledge $\Upsilon(\mathcal{P}^0, \dots, \mathcal{P}^K)$.

By abuse of notation, for any distributions $\mathcal{P}$ on $\mathbb{R}^m$ we will write $\varphi(\mathcal{P}) = \mathbb{E}_{y\sim\mathcal{P}}(\varphi(y))$ as the expected reward when the variables follow distribution $\mathcal{P}$ whenever this quantity is defined. We also write $\mathcal{P}_o$ as the pushforward of $\mathcal{P}$ along o. Let $\mathcal{M}$ be a set of distributions such that $\varphi(\mathcal{P})$ is well-defined for any $\mathcal{P} \in \mathcal{M}$. Let $r^\star \in \mathbb{R}$ and $\mathcal{P} \in \mathcal{M}$ such that $\varphi(\mathcal{P}) < r^\star$. The following quantity is crucial to the analysis

$$d_{\inf}(\mathcal{P}, \mathcal{M}, r^\star, \varphi, o) = \inf_{\mathcal{Q}\in\mathcal{M}} \{D_{\mathrm{KL}}(\mathcal{P}_o, \mathcal{Q}_o) : \varphi(\mathcal{Q}) > r^\star\},$$

where $D_{\mathrm{KL}}(\mathcal{P}_o, \mathcal{Q}_o)$ denotes the KL divergence of $\mathcal{P}_o$ from $\mathcal{Q}_o$. Intuitively, it quantities how difficult it is to learn that the expected reward $\varphi(\mathcal{P})$ is smaller than $r^\star$ (smaller the value of $d_{\inf}$ the more difficult it is). The next lemma is a straightforward adaptation of [9, Th. 1]/[26, Th. 16.2] to our model. It provides a lower bound on the asymptotic regret

Lemma 4. *Let $\mathcal{B} = \mathcal{M}^0 \times \cdots \times \mathcal{M}^K$ be a class of action distributions and φ be the reward function. Let π be a consistent algorithm over $(\mathcal{B}, \varphi)$ that is compatible with prior knowledge function Υ and observation function o. Then, if $\mathcal{B}$ is indistinguishable under Υ, for all $\mathcal{P}_* \in \mathcal{B}$, it holds that*

$$\liminf_{T\to+\infty} \frac{\mathbb{E}[\mathrm{Reg}_T]}{\log T} \geq \sum_{a\in\mathcal{A}:\Delta^a>0} \frac{\Delta^a}{d_{\inf}(\mathcal{P}^a, \mathcal{M}^a, r^\star, \varphi, o)}. \tag{30}$$

Proof. The proof follow closely the one presented in [26]. The key observation is that the same proof still carries out if we have a random vector and our reward and observation depends on this random vector. In particular, let $\mathbb{Q}(\pi, \mathcal{P}_*)$ and $\mathbb{Q}(\pi, \mathcal{Q}_*)$ be the probability measure on the actions and the observations induced by the interaction of the policy with the bandit instances $(\mathcal{P}_*, \varphi)$ and $(\mathcal{Q}_*, \varphi)$ for $\mathcal{P}_*, \mathcal{Q}_* \in \mathcal{B}$. The assumption that π is compatible with o and Υ and that $\mathcal{B}$ is indistinguishable under Υ allows us to write $q_{\pi,\mathcal{P}_*}(a_1, o_1, \dots, a_T, o_T) = \prod_{t=1}^T \pi_t^{\mathcal{B}}(a_t|a_1, o_1, \dots, a_{t-1}, o_{t-1}) p_o^{a_t}(o_t)$, where $o_1, \dots, o_T$ are the observations of the learner, $q_{\pi,\mathcal{P}_*}$ and $p_o^{a_t}$ are respectively the probability density functions of $\mathbb{Q}(\pi, \mathcal{P}_*)$ and $\mathcal{P}_o^a$, and $\pi_t^{\mathcal{B}} = \pi_{t,\Upsilon(\mathcal{P}_*)} = \pi_{t,\Upsilon(\mathcal{Q}_*)}$ is a probability kernel that can be chosen using prior knowledge.[9] We have a similar decomposition for $q_{\pi,\mathcal{Q}_*}$. Then, following

[8] For our purpose, we can regard this as a partition of the set of all bandit instances that the learner may encounter.

[9] Many technical details are omitted here. We refer the interested readers to [26] for a rigorous treatment of such proof.

the proof of [26, Lem. 15.1] we get immediately

$$D_{\mathrm{KL}}(\mathbb{Q}(\pi, \mathcal{P}_*), \mathbb{Q}(\pi, \mathcal{Q}_*)) = \sum_{a \in \mathcal{A}} \mathbb{E}_{\pi,\mathcal{P}}[N_T^a] D_{\mathrm{KL}}(\mathcal{P}_o^a, \mathcal{Q}_o^a).$$

In the above, $\mathbb{E}_{\pi,\mathcal{P}}$ means that the expectation is taken with respect to $\mathbb{Q}(\pi, \mathcal{P}_*)$. The proof can be completed in the same way as done for [26, Th. 16.2]. □

D.3 Lower Bounds for Uplifting Bandits– Proof

The following proposition embraces both Proposition 2 and Proposition 5.

Proposition 6. *Let π be a consistent algorithm over the class of 1-sub-Gaussian uplifting bandits that at most uses knowledge of $\mathcal{P}^0$, $(\mathcal{V}^a)_{a\in\mathcal{A}}$, and the fact that the noise is 1-sub-Gaussian. Then, for any $K, m > 0$ and sequence $(L^a, \Delta^a)_{1\le a\le K} \in ([m] \times \mathbb{R}_+)^K$ with $\Delta^1 = 0$, there exists a 1-sub-Gaussian uplifting bandit $(K, m, (\Delta^a, L^a)_{a\in\mathcal{A}})$ such that the regret induced by π on it satisfies*

$$\liminf_{T\to+\infty} \frac{\mathbb{E}[\mathrm{Reg}_T]}{\log T} \ge \sum_{a\in\mathcal{A}:\Delta^a>0} \frac{2(L^a)^2}{\Delta^a}. \tag{31}$$

Moreover, if π can be implemented under either of the following conditions

(a) Only the reward is observed.

(b) The learner does not have any prior knowledge about the arms' expected payoffs.

Then, there exists a 1-sub-Gaussian uplifting bandit with parameters $(K, m, (\Delta^a, L^a)_{a\in\mathcal{A}})$ where the regret of π is

$$\liminf_{T\to+\infty} \frac{\mathbb{E}[\mathrm{Reg}_T]}{\log T} \ge \sum_{a\in\mathcal{A}:\Delta^a>0} \frac{2m^2}{\Delta^a}. \tag{32}$$

Proof. Let $r^\star = 1 + \max_{a\in\mathcal{A}} \Delta^a > 0$. Throughout the proofs, the lower bound will be shown for problem instances with $\mu^0(i) \equiv 0$ and the following mean values for $a \in \mathcal{A}$

$$\mu^a(i) = \begin{cases} \dfrac{r^\star - \Delta^a}{L^a} & \text{if } 1 \le i \le L^a, \\ 0 & \text{if } L^a + 1 \le i \le m. \end{cases} \tag{33}$$

By construction, we have $\mu^a(i) \ne \mu^0(i)$ if and only if $1 \le i \le L^a$ and thus the number of mean-affected variables of a is exactly L^a. Clearly, 1 is an optimal action and the suboptimality gap of action a is Δ^a.

Next, we will need to define the classes of action distributions $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$ that we will use for proving (31), (a)-(32), and (b)-(32). To begin, we want these classes to be indistinguishable under a certain prior knowledge function, and since we allow the learner to know the baseline distribution, $\mathcal{M}^0$ must be a singleton. As for $\mathcal{M}^a$, it will take the form $\mathcal{M}^a = \{\mathcal{N}(\mu, \Sigma^a) : \mu \in \mathcal{U}^a\}$ for some mean vector set $\mathcal{U}^a \subset \mathbb{R}^m$ and covariance matrix $\Sigma^a \in [0, 1]^{m\times m}$ to be specified. This implies that every noise variable $\xi^a(i)$ is 1-sub-Gaussian and thus π is consistent in each of these classes. Moreover, the classes will also be defined in such a way that when the mean values are taken as in (33), $\mathcal{P}^0$ and $\mathcal{P}^a$ indeed have the same marginal distribution on all but the first L^a variables.

The Importance of Limited Number of Affected Variables. To prove (31), we let the unique element of $\mathcal{M}_0$ be the Dirac measure that assigns full weight to the point $(0, \dots, 0)$. For any $a \in \mathcal{A}$, we define $\mathcal{U}^a = (\mathbb{R} \setminus \{0\})^{L^a} \times \{0\}^{m-L^a}$ and block matrix Σ^a such that $\Sigma^a(i, j) = \mathbb{1}\{\max(i, j) \le L^a\}$. In other words, the random variable y^a can be written in the following form

$$y^a(i) = \begin{cases} \mu^a(i) + \xi^a & \text{if } 1 \le i \le L^a, \\ 0 & \text{if } L^a + 1 \le i \le m, \end{cases}$$

where ξ^a is Gaussian with zero mean and unit variance. Notice that the same noise applies to all the affected variables. By construction, every $\mathcal{P}^a \in \mathcal{M}^a$ has exactly the first L^a variables affected, i.e., $\mathcal{V}^a = \{1, \dots, L^a\}$. Thus $\mathcal{B}_1$ defined in this way is indistinguishable under prior knowledge on

baseline distribution $\mathcal{P}^0$ and the sets of affected variables $(\mathcal{V}^a)_{a\in\mathcal{A}}$. To conclude, we would like to apply Lemma 4. For this, we observe that for any $\mathcal{P}^a, \mathcal{Q}^a \in \mathcal{M}^a$ with means μ^a and ν^a, it holds[10]

$$D_{\mathrm{KL}}(\mathcal{P}^a, \mathcal{Q}^a) = \begin{cases} \dfrac{\tilde{\Delta}^2}{2} & \text{if } \mu^a(i) - \nu^a(i) = \tilde{\Delta} \text{ for all } i \in \{1, ..., L^a\}, \\ +\infty & \text{otherwise.} \end{cases} \tag{34}$$

With this we see immediately that for $\mathcal{P}^a$ with mean values (33), we have $d_{\inf}(\mathcal{P}^a, \mathcal{M}^a, r^\star, (+), \mathrm{Id}) = \Delta^a/(2(L^a)^2)$, where $(+)$ and Id are respectively the sum and the identity function that gives the reward and the observations of the learner. Plugging this into (30) gives exactly (31).

The importance of Side Observation. To construct $\mathcal{B}_2$, we draw the same noise for all the variables. In terms of covariance matrix, this means all the entries of Σ is equal to 1. The baseline distribution is $\mathcal{P}^0 = \mathcal{N}(\mathbf{0}, \Sigma)$ and for every a, we set $\Sigma^a = \Sigma$ and choose again $\mathcal{U}^a = (\mathbb{R}\backslash\{0\})^{L^a} \times \{0\}^{m-L^a}$. This time, only the reward, that is, the sum of the variables, is observed, and the distribution of the reward at action a is $\mathcal{N}(r^\star - \Delta^a, m^2)$. For any two real numbers γ, γ', one has $D_{\mathrm{KL}}(\mathcal{N}(\gamma, m^2), \mathcal{N}(\gamma', m^2)) = (\gamma - \gamma')^2/(2m^2)$. Therefore $d_{\inf}(\mathcal{P}^a, \mathcal{M}^a, r^\star, (+), (+)) = \Delta^a/(2m^2)$ for $\mathcal{P}^a$ with mean values (33). Notice that $\mathcal{B}_2$ is also indistinguishable under prior knowledge on the baseline distribution and on the sets of affected variables $(\mathcal{V}^a)_{a\in\mathcal{A}}$, and thus Lemma 4 can be applied, leading to (32).

The Importance of Some Prior Knowledge on How the Variables are Affected. The class $\mathcal{B}_3$ is defined almost in the same way as $\mathcal{B}_2$ except for the fact that we now set $\mathcal{U}^a = \mathbb{R}^m$ for every $a \in \mathcal{A}$. Our condition (b) postulates that $\mathcal{B}_3$ is indistinguishable under the prior knowledge of the learner. Similar to (34), for any $\mathcal{P}^a, \mathcal{Q}^a \in \mathcal{M}^a$ with means μ^a and ν^a, we now have

$$D_{\mathrm{KL}}(\mathcal{P}^a, \mathcal{Q}^a) = \begin{cases} \dfrac{\tilde{\Delta}^2}{2} & \text{if } \mu^a(i) - \nu^a(i) = \tilde{\Delta} \text{ for all } i \in \{1, ..., m\}, \\ +\infty & \text{otherwise.} \end{cases}$$

In particular, if μ^a is defined as in (33) and $\nu^a(i) = \mu^a(i) + (\Delta^a + \varepsilon)/m$ for all i for some $\varepsilon \in \mathbb{R}$, we have $D_{\mathrm{KL}}(\mathcal{P}^a, \mathcal{Q}^a) = (\Delta^a + \varepsilon)^2/(2m^2)$. We deduce immediately $d_{\inf}(\mathcal{P}^a, \mathcal{M}^a, r^\star, (+), \mathrm{Id}) = \Delta^a/(2m^2)$. Invoking Lemma 4 completes the proof. □

E Case of Unknown Affected Variables: Scenario 2

In this second part of Section 6, we assume a lower bound on individual uplift is known. This means the learner has access to $\Delta_{\text{up}} > 0$ satisfying that for all $a \in \mathcal{A}$ and $i \in \mathcal{V}^a$, $|\mu^a(i) - \mu^0(i)| \geq \Delta_{\text{up}}$. This gives us an indicator on how many times we need to take each action in order to identify all the affected variables. Similar assumption were made by Lu et al. [29] to identify which node in the causal graph affects the reward variable.

Again in this scenario, we derive regret bounds for both when the baseline payoffs are known or not to the learner. Our algorithms combine UCB, successive elimination [17], and the idea that the affected variables can be identified after an action is taken sufficiently many times. Our regret bounds feature the ratio $\Delta^a/\Delta_{\text{up}}$, suggesting that this knowledge is helpful when Δ_{up} is relatively large.

Hereinafter, for any $x, \alpha, \beta \in \mathbb{R}$ with $\alpha \leq \beta$, we use $\mathrm{clip}(x, \alpha, \beta)$ to denote the clipping function that restricts x to the interval $[\alpha, \beta]$, i.e., $\mathrm{clip}(x, \alpha, \beta) = \max(\alpha, \min(\beta, x))$.

E.1 Known Baseline Payoffs

As we have seen in Section 6, if the baseline payoffs μ^0 are known, we can directly check whether $\mu^0(i)$ is in the confidence interval. Moreover, with the knowledge $|\mu^a(i) - \mu^0(i)| \geq \Delta_{\text{up}}$ for all $i \in \mathcal{V}^a$, we deduce that $\mathcal{V}^a$ can be identified correctly with probability $1 - 2m\delta'$ after action a is taken $N_0 = \lceil (8/\Delta_{\text{up}}^2) \log(1/\delta') \rceil$ times. In fact, let $y_1^a, \ldots, y_{N_0}^a$ be N_0 i.i.d. realizations of $y^a \sim \mathcal{P}^a$ and $\tilde{\mu}_{N_0}^a(i) = (\sum_{t=1}^{N_0} y_t^a(i))/N_0$ be the empirical mean of variable i. Applying the sub-Gaussian concentration inequality (12) gives

$$\mathbb{P}\left(|\tilde{\mu}_{N_0}^a(i) - \mu^a(i)| \geq \frac{\Delta_{\text{up}}}{2}\right) \leq 2\exp\left(-\frac{N_0 \Delta_{\text{up}}^2}{8}\right) \leq 2\delta'. \tag{35}$$

[10] This can for example be proved by applying the chain rule of KL divergence.

Algorithm 5 UPUCB-iLift (b)

Input: Error probability δ', Baseline payoffs μ^0, Lower bound on individual uplift Δ_{up}
Initialization: Take each action once; Set $N_0 \leftarrow \lceil (8/\Delta_{\text{up}}^2)\log(1/\delta')\rceil$ and $\widehat{\mathcal{V}}^a \leftarrow \mathcal{V}$
for $t = K+1, \dots, T$ **do**
  Compute the UCB indices following (9)
  For $a \in \mathcal{A}$, compute uplifting index $\tau_t^a \leftarrow \sum_{i \in \widehat{\mathcal{V}}^a}(U_t^a(i) - \mu^0(i))$
  Select action $a_t \in \arg\max_{a\in\mathcal{A}} \tau_t^a$
  if $N_t^{a_t} \geq N_0$ **then**
    Set $\widehat{\mathcal{V}}^{a_t} \leftarrow \{i \in \mathcal{V} : |\hat{\mu}_t^{a_t}(i) - \mu^0(i)| > \Delta_{\text{up}}/2\}$

With a union bound we conclude that with probability at least $1 - 2m\delta'$ it holds for all i that $|\tilde{\mu}_{N_0}^a(i) - \mu^a(i)| < \Delta_{\text{up}}/2$. We distinguish between the following two situations

1. $i \in \mathcal{V}^a$, then $|\tilde{\mu}_{N_0}^a(i) - \mu^0(i)| \geq |\mu^a(i) - \mu^0(i)| - |\tilde{\mu}_{N_0}^a(i) - \mu^a(i)| > \Delta_{\text{up}} - \dfrac{\Delta_{\text{up}}}{2} = \dfrac{\Delta_{\text{up}}}{2}$.
2. $i \notin \mathcal{V}^a$, then $|\tilde{\mu}_{N_0}^a(i) - \mu^0(i)| = |\tilde{\mu}_{N_0}^a(i) - \mu^a(i)| < \dfrac{\Delta_{\text{up}}}{2}$.

This implies that with the choice $\widehat{\mathcal{V}}^a = \{i \in \mathcal{V} : |\tilde{\mu}_{N_0}^a(i) - \mu^0(i)| > \Delta_{\text{up}}/2\}$, we have effectively $\mathbb{P}(\widehat{\mathcal{V}}^a = \mathcal{V}^a) \geq 1 - 2m\delta'$. One natural algorithm is to thus first pull each arm N_0 times and run UPUCB (b) (Algorithm 3) in the remaining rounds by regarding $\widehat{\mathcal{V}}^a$ as the variables that action a affects. A direct computation shows that the regret of this algorithm is in $\mathcal{O}(\sum_{a\in\mathcal{A}:\Delta^a>0}(\Delta^a/\Delta_{\text{up}} + L^a)^2 \log T/\Delta^a)$.

However, N_0 can be arbitrarily large when Δ_{up} gets closer to 0, and taking an action much fewer times is probably sufficient to deduce that action is not optimal (with high probability). We thus propose to implement the above idea in a more flexible manner: We run UPUCB (b) with the uplifting indices $\tau_t^a = \sum_{i\in\widehat{\mathcal{V}}_t^a}(U_t^a(i) - \mu^0(i))$, where

$$\widehat{\mathcal{V}}_t^a = \begin{cases} \mathcal{V} & \text{if } N_t^a < N_0, \\ \{i \in \mathcal{V} : |\hat{\mu}_t^{a_t}(i) - \mu^0(i)| > \Delta_{\text{up}}/2\} & \text{otherwise.} \end{cases}$$

This algorithm is summarized in Algorithm 5 and has the following regret guarantee.

Theorem 3. *Let $\delta' = \delta/(2KmT)$. Then the regret of* UPUCB-*iLift* (b) *(Algorithm 5), with probability at least $1-\delta$, satisfies:*

$$\text{Reg}_T \leq \sum_{a\in\mathcal{A}:\Delta^a>0}\left(\frac{8\,\text{clip}(\Delta^a/\Delta_{\text{up}}, L^a, m)^2 \log(2KmT/\delta)}{\Delta^a} + \Delta^a\right). \tag{36}$$

Proof. As before, we prove (36) under the condition that $\mathcal{E}$ defined in (17) occurs. Recall that we have shown in the proof of Proposition 3 that $\mathbb{P}(\mathcal{E}) \geq 1-\delta$ with our choice of δ', so this effectively leads to a high-probability regret bound.

In the following we assume that $\mathcal{E}$ occurs and show that for any suboptimal action $a \in \mathcal{A}$, it holds

$$N_T^a \leq \max\left(\min\left(\frac{8\log(1/\delta')}{\Delta_{\text{up}}^2}, \frac{8m^2\log(1/\delta')}{(\Delta^a)^2}\right), \frac{8(L^a)^2\log(1/\delta')}{(\Delta^a)^2}\right) + 1. \tag{37}$$

The inequality (36) then follows directly from (11) and a rearrangement of the terms.

We first claim that for any $a \in \mathcal{A}$ if $N_t^a \geq N_0$ then $\widehat{\mathcal{V}}_t^a = \mathcal{V}^a$. In fact, by the definition of N_0, $N_t^a \geq N_0$ implies that $c_t^a < \Delta_{\text{up}}/2$. Therefore, as argued in the text, $i \in \widehat{\mathcal{V}}_t^a$ if and only if $i \in \mathcal{V}^a$. Since $\rho_t^a(i) \geq \mu_{\text{up}}^a(i)$ for any a and i, this also proves that $\rho_t^a \geq r_{\text{up}}^a$ always holds (we have either $\widehat{\mathcal{V}}_t^a = \mathcal{V}$ or $\widehat{\mathcal{V}}_t^a = \mathcal{V}^a$).

Let $t \in \{K+1, ..., T\}$ be the last time that a suboptimal action a is taken so that $N_T^a = N_{t-1}^a + 1$. This indicates $\tau_t^a \geq \tau_t^{a^\star}$, and hence $\tau_t^a \geq r_{\text{up}}^{a^\star}$. We distinguish between the following two cases:

Algorithm 6 UPUCB-iLift

Input: Error probability δ', Lower bound on effect gap Δ_{up}
Initialization: Set $t \leftarrow 0$, $\mathcal{A}_1 \leftarrow \mathcal{A}$, and $N_0 \leftarrow \lceil (32/\Delta_{\text{up}}^2) \log(1/\delta') \rceil$
`Phase I: Successive elimination`
for $\ell = 1, \ldots, N_0$ **do**
  if $T - t \geq \text{card}(\mathcal{A}_\ell)$ **then**
    Take each action $a \in \mathcal{A}_\ell$ once
    Set $t \leftarrow t + \text{card}(\mathcal{A}_\ell)$
    Compute the empirical estimates $\tilde{r}_\ell^a$ for $a \in \mathcal{A}_\ell$ and the radius of confidence interval c_ℓ
    Update active set $\mathcal{A}_{\ell+1} \leftarrow \{a \in \mathcal{A}_\ell : \tilde{r}_\ell^a + 2c_\ell \geq \max_{a'} \tilde{r}_\ell^{a'}\}$
  else
    Take $T - t$ actions from $\mathcal{A}_\ell$ and terminate the algorithm
`Phase II: UpUCB`
Construct the sets $\widehat{\mathcal{S}}(i)$ as in (38) and $\widehat{\mathcal{V}}^a \leftarrow \{i \in \mathcal{V} : a \notin \widehat{\mathcal{S}}(i)\}$
for $t = t + 1, \ldots, T$ **do**
  Compute the UCB indices following (9),(39)
  For $a \in \mathcal{A}_{N_0+1}$, compute uplifting index $\tau_t^a \leftarrow \sum_{i \in \widehat{\mathcal{V}}^a} (U_t^a(i) - U_t^0(i))$
  Select action $a_t \in \arg\max_{a \in \mathcal{A}_\ell} \tau_t^a$

1. $N_{t-1}^a < N_0$: Then $\widehat{\mathcal{V}}_t^a = \mathcal{V}$ so $\tau_t^a = \sum_{i \in \mathcal{V}} (U_t^a(i) - \mu^0) \leq r_{\text{up}}^a + 2mc_{t-1}^a$, which in turn leads to $c_{t-1}^a \geq \Delta^a/(2m)$, or equivalently $N_{t-1}^a \leq 8m^2 \log(1/\delta')/(\Delta^a)^2$.
2. $N_{t-1}^a \geq N_0$: Then $\widehat{\mathcal{V}}_t^a = \mathcal{V}^a$ so $\tau_t^a = \sum_{i \in \mathcal{V}^a} (U_t^a(i) - \mu^0) \leq r_{\text{up}}^a + 2L^a c_{t-1}^a$, which in turn leads to $c_{t-1}^a \geq \Delta^a/(2L^a)$, or equivalently $N_{t-1}^a \leq 8(L^a)^2 \log(1/\delta')/(\Delta^a)^2$.

Since we also have $N_{t-1}^a < N_0$ in the first case, the above gives

$$N_{t-1}^a \leq \max\left(\min\left(N_0 - 1, \frac{8m^2 \log(1/\delta')}{(\Delta^a)^2}\right), \frac{8(L^a)^2 \log(1/\delta')}{(\Delta^a)^2}\right).$$

Plugging in the definition of N_0 we get immediately (37). □

Theorem 3 shows that Algorithm 5 provides a smooth transition between standard UCB and the strategy that identifies the affected variables of all the actions. When $\Delta^a/\Delta_{\text{up}} \geq m$, the loss of taking action a is so large that UCB prevents it from being taken further even if $N_t^a \leq N_0$. Otherwise, the affected variables get identified after an action is taken sufficient number of times and the algorithm benefits from this knowledge to improve the estimate of the uplift. It is also quite straightforward to combine Algorithm 4 with Algorithm 5 when both L and Δ_{up} are known; this results in a regret bound that replaces $\text{clip}(\Delta^a/\Delta_{\text{up}}, L^a, m)$ by $\text{clip}(\Delta^a/\Delta_{\text{up}}, L^a, L)$ in (36).

E.2 Unknown Baseline Payoffs

Several additional challenges emerge when we want to adapt the aforementioned strategies to the case where the baseline payoffs are unknown. First, without additional assumption, it is impossible to tell whether two arbitrarily close estimates indicate the same effect of two actions on a variable. Then, without a proper baseline to compare with, we can never exclude the possibility that all the variables are affected by all the actions. To circumvent this issue, we make the following more stringent assumption.

Assumption 1. The effect of any two actions on a variable differ by at least Δ_{up} as long as one of them affects this variable, i.e., for all $a \in \mathcal{A}$, $i \in \mathcal{V}^a$, and $a' \in \mathcal{A} \setminus \{a\}$, it holds $|\mu^a(i) - \mu^{a'}(i)| \geq \Delta_{\text{up}}$.

In other words, Δ_{up} is not only a lower bound on individual uplift, but also a lower bound on (individual) effect gap. Second, with a UCB-style algorithm, the optimal actions may not be taken sufficiently many times, in which case we can only conservatively assume that they affect all the variables. This would incur additional regret in the analysis when μ^0 is unknown. Therefore, we would like to ensure that all the (potentially optimal) actions are taken the same number of times at

the point that the affected variables get identified. This leads to a two-stage method as described in Algorithm 6.

In the first stage, we perform successive elimination until the uneliminated actions are taken sufficiently many times. Here $\tilde{r}_\ell^a$ is the empirical estimate of the reward associated to action a and $c_\ell = m\sqrt{2\log(1/\delta')/\ell}$ is the associated radius of confidence interval. Subsequently, for each variable $i \in \mathcal{V}$ we construct the set

$$\widehat{\mathcal{S}}(i) = \{a \in \mathcal{A} : \exists\, a' \in \mathcal{A} \setminus \{a\}, \mathcal{C}_{t_0+1}^a(i) \cap \mathcal{C}_{t_0+1}^{a'}(i) \neq \varnothing\}, \tag{38}$$

where t_0 is the number of iterations in the elimination phase and the confidence intervals are defined as in (9). If Assumption 1 is verified, then with high probability $a \in \widehat{\mathcal{S}}(i)$ only if $i \notin \mathcal{V}^a$. Therefore, the observed results from these arms can be used to estimate the baseline payoff of i. The identified variables for action $a \in \mathcal{A}$ are then $\widehat{\mathcal{V}}^a = \{i \in \mathcal{V} : a \notin \widehat{\mathcal{S}}(i)\}$.

The second stage of the algorithm reuses the idea from UPUCB. We define the UCB indices following (9) for $a \in \mathcal{A}$ and for the baseline estimate we choose among $\widehat{\mathcal{S}}(i)$ the action that is taken the most frequently whenever this set is non-empty. Otherwise we arbitrarily set it to 0. That is,

$$U_t^0(i) = \begin{cases} 0 & \text{if } \widehat{\mathcal{S}}(i) = \varnothing, \\ U_t^{a_t^0(i)}(i) \text{ where } a_t^0(i) = \arg\max_{a \in \widehat{\mathcal{S}}(i)} N_{t-1}^a & \text{otherwise.} \end{cases} \tag{39}$$

The uplifting indices that guide the decision of the algorithm are defined by $\tau_t^a = \sum_{i \in \widehat{\mathcal{V}}_t^a} (U_t^a(i) - U_t^0(i))$.

As shown below, the regret of Algorithm 6 takes roughly the same form as the one in Theorem 3, but as in Theorem 1, the number of variables that an optimal action affects also plays a crucial role in the regret bound.

Theorem 4. *Let Assumption 1 hold and $\delta' = \delta/(2KmT)$. Then the regret of* UPUCB-iLift *(Algorithm 6), with probability at least $1-\delta$, satisfies:*

$$\mathrm{Reg}_T \leq \sum_{a \in \mathcal{A}: \Delta^a > 0} \left(\frac{8\ \mathrm{clip}(2\Delta^a/\Delta_{\text{up}}, L^a + L^{a^\star}, 2m)^2 \log(2KmT/\delta)}{\Delta^a} + \Delta^a \right). \tag{40}$$

Proof. We shall again consider the event $\mathcal{E}$ defined in (17), which holds with probability $1-\delta$ as argued in the proof of Proposition 3. Leveraging the decomposition in (11) it is sufficient to prove that on $\mathcal{E}$ we have

$$N_T^a \leq \max\left(\min\left(\frac{32\log(1/\delta')}{\Delta_{\text{up}}^2}, \frac{32m^2\log(1/\delta')}{(\Delta^a)^2} \right), \frac{8(L^a + L^{a^\star})^2 \log(1/\delta')}{(\Delta^a)^2} \right) + 1. \tag{41}$$

We assume $\mathcal{E}$ happens in the rest of the proof.

Let us first focus on the elimination phase, notice that event $\mathcal{E}$ implies $|\tilde{r}_\ell^a - r^a| < c_\ell$ for every $a \in \mathcal{A}_\ell$. In this case, the optimal actions never get eliminated. In fact, if $a^\star \in \mathcal{A}_\ell$, then for all $a \in \mathcal{A}_\ell$ it holds that

$$\tilde{r}_\ell^{a^\star} + 2c_\ell \geq r^{a^\star} + c_\ell \geq r^a + c_\ell \geq \tilde{r}_\ell^a,$$

As a consequence, we also have $a^\star \in \mathcal{A}_{\ell+1}$ and we conclude by the principle of induction. With this we further deduce that all action a with $\Delta^a \geq 4c_\ell$ gets eliminated after it is taken at most ℓ times. To see this, notice that for such action we have

$$\tilde{r}_\ell^a + 2c_\ell < r^a + 3c_\ell = r^{a^\star} - \Delta^a + 3c_\ell < \tilde{r}_\ell^{a^\star} - \Delta^a + 4c_\ell \leq \tilde{r}_\ell^{a^\star}.$$

Therefore, for all t in phase I of the algorithm, it is true that

$$N_t^a \leq \min\left(\left\lceil \frac{32\log(1/\delta')}{\Delta_{\text{up}}^2} \right\rceil, \left\lceil \frac{32m^2\log(1/\delta')}{(\Delta^a)^2} \right\rceil \right). \tag{42}$$

This suggests that inequality (41) holds if the algorithm terminates in the elimination phase.

Otherwise, the algorithm constructs $\mathcal{S}(i)$'s, $\widehat{\mathcal{V}}^a$'s, and switches to the UPUCB phase. Let us denote by $\mathcal{S}(i) = \{a \in \mathcal{A} : i \notin \mathcal{V}^a\}$ the set of actions that do not affect variable i and let $\mathcal{A}_{\text{res}} = \mathcal{A}_{N_0+1}$ be the actions that remain after phase I. We argue that the sets $\widehat{\mathcal{S}}(i)$ are constructed such that when $\mathcal{E}$ occurs,

(*i*) $\widehat{\mathcal{S}}(i) \subseteq \mathcal{S}(i)$.

(*ii*) If there exist $a, a' \in \mathcal{A}_{\text{res}} \cap \mathcal{S}(i)$ such that $a \neq a'$ then we also have $a, a' \in \widehat{\mathcal{S}}(i)$.

The second point is proved by observing that under $\mathcal{E}$ we must have $\mu^0(i) \in \mathcal{C}^a_{t_0+1}(i) \cap \mathcal{C}^{a'}_{t_0+1}(i)$ if $a, a' \in \mathcal{A}_{\text{res}} \cap \mathcal{S}(i)$. To show the first point, notice that after the elimination phase, each of the action in $\mathcal{A}_{\text{res}}$ is taken exactly $\lceil (32/\Delta^2_{\text{up}}) \log(1/\delta') \rceil$ times and thus $c^a_{t_0} \leq \Delta_{\text{up}}/4$. By Assumption 1, if $a \notin \mathcal{S}(i)$, then for all $a' \in \mathcal{A}_{\text{res}} \setminus \{a\}$ we have $|\mu^a(i) - \mu^{a'}(i)| \geq \Delta_{\text{up}}$, which along with the definition of $\mathcal{E}$ implies $\mathcal{C}^a_{t_0+1}(i) \cap \mathcal{C}^{a'}_{t_0+1}(i) = \varnothing$, and therefore $a \notin \widehat{\mathcal{S}}(i)$.

The rest of the proof follows closely that of Theorem 1. Suppose that a suboptimal action a is taken at round t during phase II of the algorithm. This means $\tau^a_t \geq \tau^{a^\star}_t$ (recall that $a^\star \in \mathcal{A}_{\text{res}}$ since $\mathcal{E}$ happens), i.e.,

$$\sum_{i \in \widehat{\mathcal{V}}^a} (U^a_t(i) - U^0_t(i)) \geq \sum_{i \in \widehat{\mathcal{V}}^{a^\star}} (U^{a^\star}_t(i) - U^0_t(i)).$$

Rearranging the terms we get

$$\sum_{i \in \widehat{\mathcal{V}}^a} U^a_t(i) + \sum_{i \in \widehat{\mathcal{V}}^{a^\star} \setminus \widehat{\mathcal{V}}^a} U^0_t(i) \geq \sum_{i \in \widehat{\mathcal{V}}^{a^\star}} U^{a^\star}_t(i) + \sum_{i \in \widehat{\mathcal{V}}^a \setminus \widehat{\mathcal{V}}^{a^\star}} U^0_t(i).$$

If $i \in \widehat{\mathcal{V}}^{a^\star} \setminus \widehat{\mathcal{V}}^a$ then $a \in \widehat{\mathcal{S}}(i)$. In particular, $\widehat{\mathcal{S}}(i) \neq \varnothing$ so $a' = a^0_t(i) = \arg\max_{\tilde{a} \in \widehat{\mathcal{S}}(i)} N^{\tilde{a}}_{t-1}$ is well-defined and $U^0_t(i) = U^{a'}_t(i)$. With $\widehat{\mathcal{S}}(i) \subseteq \mathcal{S}(i)$ we know that $\mu^{a'}(i) = \mu^0(i)$ and accordingly $|\mu^{a'}_{t-1}(i) - \mu^0(i)| \leq c^{a'}_{t-1}$ by the fact that $\mathcal{E}$ happens. This shows $U^0_t(i) \geq \mu^0(i)$. Moreover, $N^{a'}_t \geq N^a_t$ by the choice of a'; hence, $U^0_t(i) \leq \mu^0(i) + 2c^{a'}_{t-1} \leq \mu^0(i) + 2c^a_{t-1}$. The same argument applies to $i \in \widehat{\mathcal{V}}^a \setminus \widehat{\mathcal{V}}^{a^\star}$, and the UCB indices of a and $a^\star$ can be bounded from above and below as in previous proofs. In summary, we obtain

$$\sum_{i \in \widehat{\mathcal{V}}^a} (\mu^a(i) + 2c^a_{t-1}) + \sum_{i \in \widehat{\mathcal{V}}^{a^\star} \setminus \widehat{\mathcal{V}}^a} (\mu^0(i) + 2c^a_{t-1}) \geq \sum_{i \in \widehat{\mathcal{V}}^{a^\star}} \mu^{a^\star}(i) + \sum_{i \in \widehat{\mathcal{V}}^a \setminus \widehat{\mathcal{V}}^{a^\star}} \mu^0(i).$$

Equivalently,

$$2 \operatorname{card}\left(\widehat{\mathcal{V}}^a \cup \widehat{\mathcal{V}}^{a^\star}\right) c^a_{t-1} \geq \sum_{i \in \widehat{\mathcal{V}}^a \cup \widehat{\mathcal{V}}^{a^\star}} \mu^{a^\star}(i) - \sum_{i \in \widehat{\mathcal{V}}^a \cup \widehat{\mathcal{V}}^{a^\star}} \mu^a(i). \tag{43}$$

To conclude, we claim that

$$\widehat{\mathcal{V}}^a \cup \widehat{\mathcal{V}}^{a^\star} = \mathcal{V}^a \cup \mathcal{V}^{a^\star}. \tag{44}$$

Let us first show $\mathcal{V}^a \subseteq \widehat{\mathcal{V}}^a \subseteq \mathcal{V}^a \cup \mathcal{V}^{a^\star}$. The first inclusion is a consequence of $\widehat{\mathcal{S}}(i) \subseteq \mathcal{S}(i)$:

$$i \in \mathcal{V}^a \implies a \notin \mathcal{S}(i) \implies a \notin \widehat{\mathcal{S}}(i) \implies i \in \widehat{\mathcal{V}}^a.$$

As for the second inclusion we prove it's contrapositive:

$$i \notin \mathcal{V}^a \cup \mathcal{V}^{a^\star} \implies a, a^\star \in \mathcal{A}_{\text{res}} \cap \mathcal{S}(i) \implies a, a^\star \in \widehat{\mathcal{S}}(i) \implies i \notin \widehat{\mathcal{V}}^a$$

Notice that the second implication holds because a is suboptimal and in particular $a \neq a^\star$. The inclusion $\mathcal{V}^{a^\star} \subseteq \widehat{\mathcal{V}}^{a^\star} \subseteq \mathcal{V}^a \cup \mathcal{V}^{a^\star}$ can be proved in a similar way, and combining these results gives (44). Therefore, the right-hand side of (43) is exactly Δ^a while the left-hand side of (43) is bounded from above by $2(L^a + L^{a^\star})c^a_{t-1}$. In consequence,

$$N^a_{t-1} \leq \frac{8(L^a + L^{a^\star})^2 \log(1/\delta')}{(\Delta^a)^2}.$$

Therefore, any suboptimal action that gets taken in the UpUCB phase satisfies

$$N^a_T \leq \frac{8(L^a + L^{a^\star})^2 \log(1/\delta')}{(\Delta^a)^2} + 1.$$

On the other hand, if an suboptimal action is not taken in the UpUCB phase then (42) applies, so in either case (41) is verified, which concludes the proof. □

Algorithm 7 C2UPUCB: Case of single treatment

Input: Budget constraint L, Parameters for computing the UCBs indices
for $t = 1, \dots, T$ **do**
  for $i \in \mathcal{V}$ **do**
    Compute UCB indices $(U_t^z(i))_{z\in\{0,1\}}$
    Compute uplifting index $\rho_t(i)$ $\{\rho_t(i) = U_t^1(i) - U_t^0(i)\}$
  Set $\mathcal{V}_t \leftarrow \arg\max_{\mathcal{S}\subset\mathcal{V},\mathrm{card}(\mathcal{S})\leq L} \sum_{i\in\mathcal{S}} \rho_t(i)$
  Select action $a_t = (\mathbb{1}\{i \in \mathcal{V}_t\})_{i\in\mathcal{V}}$ {this defines the set of treated customers}

Remark 2. A natural candidate for $U_t^0(i)$ is the one that uses aggregated information as in (5). The problem of this choice is that how we aggregate the observations collected in the first phase are only determined at the end of the the phase, while this choice itself depends on these observations. Mathematically, this prevents us from defining a suitable martingale difference sequence for which we establish concentration bound. An easy fix is to only aggregate observations from the second phase, while the use of the observations from the first phase should be limited to those coming from just one arm.

F Contextual Combinatorial Uplifting Bandits

To demonstrate the flexibility of our framework, we consider in this section another contextual extension that tackles the problem of targeted campaign. Throughout the section, we refer to the variables as individuals (customers) that we may want to treat (serve) with some treatment (campaign). To improve the overall return on investment, it is often beneficial to treat only a subset of the individuals. The bandit problem thus consists of finding the right set of individuals to treat at each round. We formalize this model and provide a template algorithm for solving the problem in Appendix F.1. Subsequently, we specify the algorithm for the case of linear expected payoffs and present some associated analysis in Appendix F.2. Numerical experiments conducted with the Criteo Uplift Prediction Dataset are presented in Appendix F.3.

F.1 Model and Algorithm Template

Let us consider a pool of m individuals that may change from round to round, associated with (time-dependent) contexts $(x_t(i))_{i\in\mathcal{V}}$ with $x_t(i) \in \mathbb{R}^d$. At each round, the learner selects at most L individuals to treat. If there is only a single treatment, the action set can be written as $\mathcal{A} = \{a \in \{0,1\}^m : \|a\|_0 \leq L\}$, where $\|a\|_0$ is the number of non-zero coordinates of a. In this form, the i-th coordinate of a is 1 if and only if individual i receives the treatment. Similar to before, the learner observes the payoffs of all the individuals after treatments are applied.

We also make a practical *no interference* assumption, that the payoff of each individual only depends on whether they are treated or not (i.e., independent of others). Then, with the notations defined in Section 7, we can define functions g^z for $z \in \{0,1\}$ so that $f^a(x_t(i)) = g^{a(i)}(x_t(i))$.[11] Considering the action of not treating any individual as the baseline, the uplift of action a in round t is given by $r^a_{\text{up}}(x_t) = \sum_{i\in\mathcal{V}^a} g^{a(i)}(x_t(i)) - g^0(x_t(i))$, where $\mathcal{V}^a$ is the set of individuals treated by action a. For UCB-type methods, we should thus provide upper confidence bounds $U_t^z(i)$ for $g^z(x_t(i))$. Depending on whether the baseline function is known or not, we may subtract $g^0(x_t(i))$ or $U_t^0(i)$ from $U_t^1(i)$ to compute the uplifting indices for each individual. The algorithm then chooses the L individuals with the highest uplifting indices. We formalize the above explanation in Algorithm 7, and refer to the resulting algorithm as C2UPUCB, where C2 stands for "contextual combinatorial".

The Case of Multiple Treatments. More generally, there may be a set of different treatments that can be applied. Let us denote by $\mathcal{Z} = \{1, ..., q\}$ this set of treatments, and write $\mathcal{Z}_0 = \mathcal{Z} \cup \{0\}$, where, as before, 0 indicates that no treatment is applied to an individual. The underlying action set is now $\mathcal{A} = \{a \in (\mathcal{Z}_0)^m : \|a\|_0 \leq L\}$. Similarly, we associate each treatment with a function g^z. The reward and the uplift of choosing action a in round t are given then by

$$r^a(x_t) = \sum_{i\in\mathcal{V}} g^{a(i)}(x_t(i)), \quad r^a_{\text{up}}(x_t) = \sum_{i\in\mathcal{V}^a} g^{a(i)}(x_t(i)) - g^0(x_t(i)).$$

[11] Strictly speaking, f^a is a function of $(x_t(i), i)$ here.

Algorithm 8 C2UPUCB-M: Case of multiple treatments

Input: Budget constraint L, Parameters for computing the UCBs indices, Treatment set $\mathcal{Z}_0$
for $t = 1, \ldots, T$ **do**
  for $i \in \mathcal{V}$ **do**
    Compute UCB indices $(U_t^z(i))_{z\in\mathcal{Z}_0}$ for $(g^z(x_t(i)))_{z\in\mathcal{Z}_0}$ {use (47) for linear payoffs}
    Compute uplifting indices $(\rho_t^z(i))_{z\in\mathcal{Z}}$ $\{\rho_t^z(i) = U_t^z(i) - U_t^0(i)\}$
    Compute presumed optimal treatment $z_t(i)$ and uplifting index $\rho_t(i)$ {use (45)}
  Set $\mathcal{V}_t \leftarrow \arg\max_{\mathcal{S}\subset\mathcal{V},\mathrm{card}(\mathcal{S})\le L} \sum_{i\in\mathcal{S}} \rho_t(i)$ {this defines the set of treated customers}
  Select action $a_t \leftarrow (z_t(i)\,\mathbb{1}\{i \in \mathcal{V}_t\})_{i\in\mathcal{V}}$ $\{a_t(i) = z_t(i)$ if $i \in \mathcal{V}_t$; $a_t(i) = 0$ otherwise$\}$

The new problem consists in finding both the set of customers to treat and the treatments to apply to these customers at each round. To adapt C2UPUCB to this situation, we define a treatment-dependent uplifting index $\rho_t^z(i) = U_t^z(i) - U_t^0(i)$ where $U_t^z(i)$ and $U_t^0(i)$ are respectively upper confidence bounds on $g^{a(i)}(x_t(i))$ and $g^0(x_t(i))$. Again, if g^0 is known, $U_t^0(i)$ can be simply taken as $g^0(x_t(i))$. Next, at each time step t, we define the presumed optimal treatment for individual i and the associated uplifting index as

$$z_t(i) = \arg\max_{z\in\mathcal{Z}} \rho_t^z(i), \qquad \rho_t(i) = \rho_t^{z_t(i)}(i) = \max_{z\in\mathcal{Z}} \rho_t^z(i). \tag{45}$$

The adapted C2UPUCB algorithm then chooses the L customers with the largest uplifting index and for each of these customers chooses their presumed optimal treatment. The resulting algorithm (referred to as C2UPUCB-M) is summarized in Algorithm 8 .

F.2 Linear Expected Payoffs

To provide a concrete algorithm and its analysis thereof, let us assume that g^z is linear in the feature vector with some unknown coefficient vector θ^z, i.e., $g^z(x_t(i)) = \langle \theta^z, x_t(i)\rangle$. Then, we can construct upper confidence bounds for $g^z(x_t(i))$ following the ideas of linear UCB [1]. To begin, for each $z \in \mathcal{Z}_0$ and $t \in \{1, ..., T\}$, we define

$$\mathcal{V}_t^z = \{i \in \mathcal{V} : a_t(i) = z\}$$

as the set of customers that receive treatment z in round t. Fix a regularization parameter $\lambda > 0$, we further define the following $d \times d$ matrix and d-dimensional vector

$$V_t^z = \lambda I + \sum_{s=1}^{t} \sum_{i\in\mathcal{V}_s^z} x_s(i) x_s(i)^\top, \qquad b_t^z = \sum_{s=1}^{t} \sum_{i\in\mathcal{V}_s^z} y_s(i) x_s(i). \tag{46}$$

The regularized least-square estimate of θ^z is then given by $\hat{\theta}_t^z = (V_t^z)^{-1} b_t^z$. The UCB index is of the form

$$U_t^z(i) = \langle \hat{\theta}_{t-1}^z, x_t(i)\rangle + \alpha_t \|x_t(i)\|_{(V_{t-1}^z)^{-1}}. \tag{47}$$

for some α_t that is properly chosen.

Regret Guarantee. In this part, we provide regret analysis for Algorithm 8 when *i*) the baseline payoffs are unknown, *ii*) the expected payoffs are linear, and *iii*) the UCB indices are computed following (47). For technical reasons, we will assume the noises $\xi_t(i)$ to be mutually independent (here $\xi_t(i) = y_t(i) - \langle \theta^z, x_t(i)\rangle$). We comment on this assumption in Remark 3.

The theorem below establishes a high-probability regret bound for C2UPUCB-M when the above conditions are satisfied.

Theorem 5. *Assume that $\|x_t(i)\| \le 1$ for all t and i and $(\xi_t(i))_{i\in\mathcal{V}}$ are mutually independent and 1-sub-Gaussian (conditioning on the past). Let $S > 0$ such that $\|\theta^z\| \le S$ for all $z \in \mathcal{Z}_0$. Then if* C2UPUCB-M *(Algorithm 8) is run with the UCB indices defined in (47) under the parameter choices*

$$\alpha_t = \sqrt{\lambda} S + \sqrt{2\log\left(\frac{q+1}{\delta}\right) + d\log\left(1 + \frac{mt}{d\lambda}\right)},$$

and $\lambda \geq L$, it holds, with probability at least $1-\delta$, that

$$\mathrm{Reg}_T \leq 4\sqrt{dL(q+1)T\log\left(1+\frac{LT}{d\lambda}\right)}\left(\sqrt{\lambda}S+\sqrt{2\log\left(\frac{q+1}{\delta}\right)+d\log\left(1+\frac{mT}{d\lambda}\right)}\right). \tag{48}$$

Theorem 5 provides a regret bound in $\tilde{\mathcal{O}}(\sqrt{\max(d,L)dqLT})$. We observe that the dependence on m is only logarithmic. In fact, m plays no role in either the number of parameters to estimate (which is $(q+1)d$) or the scale of the noise (which is in $\mathcal{O}(\sqrt{L})$). Ideally, we would like to remove completely the dependence on m. However, we are not able to achieve this in our proof, and we leave whether this is possible or not as an open question. Finally, due the mutual independence of noises, the regret only scales with $\sqrt{L}$ as long as $L \leq d$.

Remark 3. The $\log(m)$ factor appears because $\hat{\theta}_t^0$ is estimated using feedback from roughly mT samples (see Lemma 5 below). This is also the reason why we need mutual independence of the noises, since otherwise, this estimation may cause the regret to get $\sqrt{m}$ larger. It can be easily verified that the dependence on m can be completely removed when the baseline payoffs are known and used for computing the uplifting indices. In this case, we can also show if the noises are not mutually independent, multiplying the second term of α_t by $\sqrt{L}$ still guarantees a regret in $\tilde{\mathcal{O}}(dL\sqrt{qT})$. What is at stake here is Lemma 5 that we present below, as the size of the confidence ellipsoid varies for different underlying assumptions.

Analysis. To prove Theorem 5, we need to show that $U_t^a(i)$ as defined in (47) is indeed an upper confidence bound on $\langle \theta^z, x_t(i)\rangle$. This step is based on the following lemma which establishes the confidence ellipsoids of the estimated parameters. In fact, $U_t^a(i)$ defined in (47) can be regarded as the largest estimated payoff that one can get from choosing a parameter from the confidence ellipsoid.

Lemma 5. *Assume that $\|x_t(i)\| \leq 1$ for all t and i and $(\xi_t(i))_{i\in\mathcal{V}}$ are mutually independent and 1-sub-Gaussian (conditioning on the past). For all $z \in \mathcal{Z}_0$ and $\delta' \in (0,1)$, with probability at least $1-\delta'$ it holds for all t that*

$$\begin{aligned}\|\theta^z - \hat{\theta}_t^z\|_{V_t^z} &\leq \sqrt{\lambda}\|\theta^z\| + \sqrt{2\log\left(\frac{1}{\delta'}\right)+\log\left(\frac{\det(V_t^z)}{\lambda^d}\right)} \\ &\leq \sqrt{\lambda}\|\theta^z\| + \sqrt{2\log\left(\frac{1}{\delta'}\right)+d\log\left(1+\frac{mt}{d\lambda}\right)}.\end{aligned}$$

Proof. This is an immediate corollary of [1, Th. 2]. Since $(\xi_t(i))_{i\in\mathcal{V}}$ are mutually independent, we can apply this theorem by arranging $(\xi_t(i))_{i\in\mathcal{V}_t^z}$ in some arbitrary order $\xi_t(i_1),\dots,\xi_t(i_{\mathrm{card}(\mathcal{V}_t^z)})$ so that $\xi_t(i_k)$ is 1-sub-Gaussian conditioning on $(\bigcup_{s<t}\{\xi_s(i) : i \in \mathcal{V}_s^z\}) \cup \{\xi_t(i_{k'})\}_{1\leq k'\leq k}$.

For the second inequality, we notice that $\mathrm{tr}(V_t^z) \leq d\lambda + \sum_{s=1}^t \sum_{i\in\mathcal{V}_s^z}\|x_t(i)\|^2 \leq d\lambda + mt$. It thus follows from the AM–GM inequality on the eigenvalues of V_t^z, $\det(V_t^z) \leq (\mathrm{tr}(V_t^z)/d)^d$. □

The lemma below will also be useful for bounding the regret.

Lemma 6. *Consider subsets of the variables $\mathcal{V}_1,\dots,\mathcal{V}_T \subset \mathcal{V}$ and define $V_t = \lambda I + \sum_{s=1}^t\sum_{i\in\mathcal{V}_s} x_s(i)x_s(i)^\top$. If $\lambda \geq \max_{t\in[T]}\mathrm{card}(\mathcal{V}_t)$ and $\|x_t(i)\| \leq 1$ for all t and i, it holds that*

$$\sum_{t=1}^T\sum_{i\in\mathcal{V}_t}\|x_t(i)\|^2_{V_{t-1}^{-1}} \leq 2\log\left(\frac{\det(V_T)}{\lambda^d}\right) \leq 2d\log\left(1+\frac{\sum_{t=1}^T\mathrm{card}(\mathcal{V}_t)}{d\lambda}\right).$$

Proof. For all $t \geq 1$, we have

$$V_t = V_{t-1} + \sum_{i\in\mathcal{V}_t} x_t(i)x_t(i)^\top = V_{t-1}^{1/2}\left(I + \sum_{i\in\mathcal{V}_t} V_{t-1}^{-1/2}x_t(i)x_t(i)^\top V_{t-1}^{-1/2}\right)V_{t-1}^{1/2}$$

Thus, with $w_t(i) = V_{t-1}^{-1/2} x_t(i)$, we get

$$\det(V_t) = \det(V_{t-1}) \det\left(I + \sum_{i \in \mathcal{V}_t} w_t(i) w_t(i)^\top\right). \tag{49}$$

Regarding $\{w_t(i)\}_{i \in \mathcal{V}_t}$ as column vectors, we denote by their horizontal concatenation as $W \in \mathbb{R}^{d \times m_t}$, where $m_t = \operatorname{card}(\mathcal{V}_t)$. We have $WW^\top = \sum_{i \in \mathcal{V}_t} w_t(i) w_t(i)^\top$. Let the eigenvalues of $WW^\top$ be denoted by $\lambda_1, \ldots, \lambda_d$, then clearly $\det(I + WW^\top) = \prod_{k=1}^d (1 + \lambda_k)$. On the other hand, $\sum_{k=1}^d \lambda_k = \operatorname{tr}(WW^\top) = \operatorname{tr}(W^\top W) = \sum_{i \in \mathcal{V}_t} \|w_t(i)\|^2$. From the non-negativity of the eigenvalues of $WW^\top$, we obtain

$$\det(I + WW^\top) = \prod_{k=1}^d (1 + \lambda_k) \geq 1 + \sum_{k=1}^d \lambda_k = 1 + \sum_{i \in \mathcal{V}_t} \|w_t(i)\|^2 = 1 + \sum_{i \in \mathcal{V}_t} \|x_t(i)\|^2_{V_{t-1}^{-1}}. \tag{50}$$

Applying (49) and (50) recursively leads to

$$\det(V_T) \geq \det(\lambda I) \prod_{t=1}^T \left(1 + \sum_{i \in \mathcal{V}_t} \|x_t(i)\|^2_{V_{t-1}^{-1}}\right) = \lambda^d \prod_{t=1}^T \left(1 + \sum_{i \in \mathcal{V}_t} \|x_t(i)\|^2_{V_{t-1}^{-1}}\right). \tag{51}$$

We now turn back to bound $\sum_{i \in \mathcal{V}_t} \|x_t(i)\|^2_{V_{t-1}^{-1}}$. Since $\lambda \geq \max_{t \in [T]} \operatorname{card}(\mathcal{V}_t)$ and $\|x_t(i)\| \leq 1$, we have

$$\sum_{i \in \mathcal{V}_t} \|x_t(i)\|^2_{V_{t-1}^{-1}} = \min\left(1, \sum_{i \in \mathcal{V}_t} \|x_t(i)\|^2_{V_{t-1}^{-1}}\right) \leq 2 \log\left(1 + \sum_{i \in \mathcal{V}_t} \|x_t(i)\|^2_{V_{t-1}^{-1}}\right). \tag{52}$$

Combining (51) and (52) we get the first inequality. The second inequality is proved as in Lemma 5. □

We are now ready to prove Theorem 5. The first part of the proof reuses the trick that we have seen several times before, notably in the proof of Theorem 1. We rearrange the UCB indices and show that the suboptimality gaps can be bounded from above by the sum of the radii of certain confidence intervals. The details will be omitted. Next, we apply Lemma 6 to bound the sum of these upper bounds and conclude.

Proof of Theorem 5. For any $z \in \mathcal{Z}_0$, choosing $\delta' = \delta/(q+1)$ in Lemma 5 gives

$$\mathbb{P}\left(\{\exists t, \|\theta^z - \hat{\theta}_t^z\|^2_{V_t^z} > \alpha_t\right) \geq \frac{\delta}{q+1}.$$

With a union bound we thus deduce that with probability at least $1-\delta$, the inequality $\|\theta^z - \hat{\theta}_t^z\|^2_{V_t^z} \leq \alpha_t$ holds for all t and z. This also implies that for all t and z, we have

$$|\langle \theta^z, x_t(i)\rangle - \langle \hat{\theta}_{t-1}^z, x_t(i)\rangle| \leq \alpha_t \|x_t(i)\|_{(V_{t-1}^z)^{-1}} \tag{53}$$

In the following, we will prove (48) under the condition that (53) holds, which happens with probability at least $1 - \delta$ as we just showed.

Let $t \geq 1$ and $a_t^\star \in \arg\max_{a \in \mathcal{A}} \sum_{i \in \mathcal{V}} \langle \theta^{a(i)}, x_t(i)\rangle$ be an optimal action of this round. The algorithm chooses a_t that maximizes $\rho_t^a = \sum_{i \in \mathcal{V}^a} (U_t^a(i) - U_t^0(i))$, and thus

$$\sum_{i \in \mathcal{V}^{a_t}} (U_t^{a_t(i)}(i) - U_t^0(i)) \geq \sum_{i \in \mathcal{V}^{a_t^\star}} (U_t^{a_t^\star(i)}(i) - U_t^0(i)). \tag{54}$$

On the other hand, inequality (53) can be rewritten as

$$U_t^z(i) - 2\alpha_t \|x_t(i)\|_{(V_{t-1}^z)^{-1}} \leq \langle \theta^z, x_t(i)\rangle \leq U_t^z(i). \tag{55}$$

Combining (54) and (55) and rearranging, we get

$$\sum_{i \in \mathcal{V}} \langle \theta^{a_t^\star(i)} - \theta^{a_t(i)}, x_t(i)\rangle = \sum_{i \in \mathcal{V}^{a_t} \cup \mathcal{V}^{a_t^\star}} \langle \theta^{a_t^\star(i)} - \theta^{a_t(i)}, x_t(i)\rangle \leq \sum_{i \in \mathcal{V}^{a_t} \cup \mathcal{V}^{a_t^\star}} 2\alpha_t \|x_t(i)\|_{(V_{t-1}^{a_t(i)})^{-1}}.$$

Summing the above from $t = 1$ to T and applying the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
\mathrm{Reg}_T &\le \sum_{t=1}^{T} \sum_{i \in \mathcal{V}^{a_t} \cup \mathcal{V}^{a_t^\star}} 2\alpha_t \|x_t(i)\|_{(V_{t-1}^{a_t(i)})^{-1}} \\
&\le 2\alpha_T \sqrt{\left(\sum_{t=1}^{T} \mathrm{card}(\mathcal{V}^{a_t} \cup \mathcal{V}^{a_t^\star})\right) \left(\sum_{t=1}^{T} \sum_{i \in \mathcal{V}^{a_t} \cup \mathcal{V}^{a_t^\star}} \|x_t(i)\|^2_{(V_{t-1}^{a_t(i)})^{-1}}\right)} \\
&\le 2\alpha_T \sqrt{2LT \sum_{t=1}^{T} \left(\sum_{z \in \mathcal{Z}} \sum_{i \in \mathcal{V}_t^z} \|x_t(i)\|^2_{(V_{t-1}^z)^{-1}} + \sum_{i \in \mathcal{V}^{a_t^\star} \setminus \mathcal{V}^{a_t}} \|x_t(i)\|^2_{(V_{t-1}^0)^{-1}}\right)}.
\end{aligned}$$

Since $\lambda \ge L$, by Lemma 6, for all $z \in \mathcal{Z}$ it holds that

$$\sum_{t=1}^{T} \sum_{i \in \mathcal{V}_t^z} \|x_t(i)\|^2_{(V_{t-1}^z)^{-1}} \le 2d \log\left(1 + \frac{LT}{d\lambda}\right).$$

Similarly, let us define $\widetilde{V}_t^0 = \lambda I + \sum_{s=1}^{t} \sum_{i \in \mathcal{V}^{a_t^\star} \setminus \mathcal{V}^{a_t}} x_s(i) x_s(i)^\top$. Using $\widetilde{V}_t^0 \preccurlyeq V_t^0$ and invoking Lemma 6 with $\mathcal{V}_t \leftarrow \mathcal{V}^{a_t^\star} \setminus \mathcal{V}^{a_t}$ gives

$$\sum_{t=1}^{T} \sum_{i \in \mathcal{V}^{a_t^\star} \setminus \mathcal{V}^{a_t}} \|x_t(i)\|^2_{(V_{t-1}^0)^{-1}} \le \sum_{t=1}^{T} \sum_{i \in \mathcal{V}^{a_t^\star} \setminus \mathcal{V}^{a_t}} \|x_t(i)\|^2_{(\widetilde{V}_{t-1}^0)^{-1}} \le 2d \log\left(1 + \frac{LT}{d\lambda}\right).$$

Putting the above three inequalities together, we get exactly (48). □

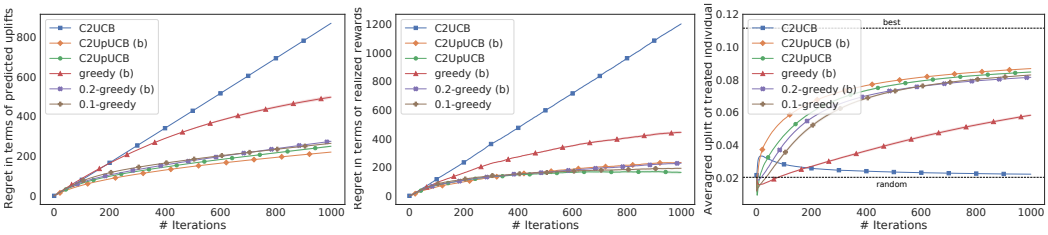
F.3 Experiments for Targeted Campaign with Criteo Uplift Dataset

In this section, we provide experiments for the contextual combinatorial uplifting model using the Criteo Uplift dataset. At each round, 100 individuals are sampled from the dataset and the problem consists in choosing 10 of them to treat so that the number of visits over the 100 individuals is maximized. Here, we compare our methods based on C2UPUCB (Algorithm 7) with approaches based on greedy and ε-greedy (see [26] for a description) that are more appropriate baselines in this setup. Again, our results are averaged over 100 independent runs and the shaded areas of the plots represent the standard errors.

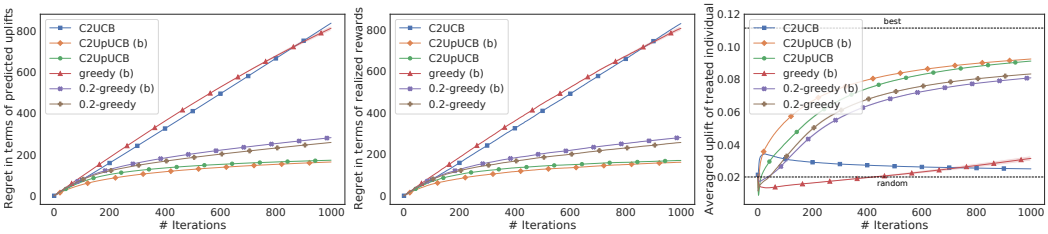
Algorithms. In all the considered algorithms, the learner learns a logistic model based on the feedback they have received from the individuals they select to treat.[12] If the baseline model is unknown, the learner also learns a baseline logistic model based on the feedback received from the individuals they do not treat. In the following, we will denote respectively by $\hat{g}_t^1$ and $\hat{g}_t^0$ the conditional probability functions of the two fitted models using data up to and including round t. When the baseline model is known, we denote by g^0 its conditional probability function. The methods that we look into in this experiment mainly differ in how they select individuals in each round t, as we detail below. As before, we use the suffix (bl) to indicate modification of the respective algorithm to the case where the baseline payoffs are known.

- **Greedy (bl):** The learner chooses the 10 individuals with the largest estimated uplifts $\hat{g}_{t-1}^1(x) - g^0(x)$.
- **ε-greedy (bl):** The learner reserves a exploration budget b where b is sampled from $B(10, \varepsilon)$ the binomial distribution with parameters 10 and ε. $10 - b$ individuals are chosen greedily as in the greedy strategy while the remaining b individuals are randomly sampled from those that have not been selected.
- **ε-greedy:** For the greedy selection part in ε-greedy, the estimated uplifts are computed by $\hat{g}_{t-1}^1(x) - \hat{g}_{t-1}^0(x)$.

[12] Logistic models were also considered in the original Criteo dataset paper [15] to compute uplifts.



(a) Logistic models only used to generate counterfactual outcomes.



(b) Logistic models used to generate all the outcomes.

Figure 4: Results on the targeted campaign experiment. For the top row the logistic models are only used to generate counterfactual outcomes, and in the bottom row all the outcomes are generated using the predicted conditional probabilities of the logistic models. We plot regret computed using the predicted uplifts and the realized rewards, and the predicted uplift averaged over all the treated individuals.

- **C2UPUCB (bl):** For each individual i we define their uplifting index $\rho_t^a(i) = \hat{g}_{t-1}^1(x) + \beta\|x_t(i)\|_{(V_{t-1}^1)^{-1}} - g^0(x)$ where V_{t-1}^1 is defined following (46) with $\lambda = 1$ and β is a exploration parameter to be tuned. The algorithm consists in choosing the 10 individuals with the largest uplifting indices.
- **C2UPUCB:** It works in the same way as C2UPUCB with baseline knowledge but the uplifting index is now defined by $\rho_t^a(i) = \hat{g}_{t-1}^1(x) + \beta\|x_t(i)\|_{(V_{t-1}^1)^{-1}} - \hat{g}_{t-1}^0(x) - \beta\|x_t(i)\|_{(V_{t-1}^0)^{-1}}$; that is, it is computed as the difference between two upper confidence bounds.
- **C2UCB:** It works in the same way as C2UPUCB but the baseline is ignored, so $\rho_t^a(i) = \hat{g}_{t-1}^1(x) + \beta\|x_t(i)\|_{(V_{t-1}^1)^{-1}}$. The linear variant of this algorithm was introduced in [32] for contextual combinatorial bandits.

Note that as the reward comes from a logistic model, we use a logistic regression for the bandit model based on the idea of GLM-UCB [18]. For the results, we select the optimal parameters from $\varepsilon \in \{0.05, 0.1, 0.2, 0.3\}$ and $\beta \in \{0.5, 1, 2, 3\}$. More details on the tuning of the parameters are provided in Appendix G.1, and the chosen optimal parameters are reported in Table 3.

Data Preprocessing. For this experiment, we normalize the data so that the features are contained in the range $[0, 1]$. Moreover, we also subsample the dataset to increase the visit rate to 0.25. In fact, we found the logistic models to perform very poorly on the original dataset (with a f1-score around 0.4) due to its small visit rate at 0.04.

Feedback. We consider two ways to use the fitted logistic models

1. The logistic models are only used to generate the counterfactual outcomes. That is, the feedback of an individual is sampled from the Bernoulli distribution with mean predicted by either of the models only if the learner decides to treat the individual while the individual is not treated in the dataset or vice-versa.
2. The logistic models are used to generate all the outcomes.

Results. In Fig. 4 (Right), we plot the uplift averaged over all the individuals that have been treated up to a time horizon by an algorithm. Higher this value is better it is because it means we are selecting the individuals for which the treatment is the most effective. For comparison, we also indicate the average uplift of random selection and an oracle strategy that always chooses the 10 individuals with the highest uplifts. Both C2UPUCB and ε-greedy beat greedy and random selection, which demonstrates the benefit of considering a bandit model for this problem. ε-greedy turns out to be

Table 3: Parameter values used for the reported experimental results. TS stands for Thompson sampling. Please refer to the text for a explanation on what is the parameter that we tune for each of the algorithms.

Non-contextual	UCB	UPUCB (b)	UPUCB	UPUCB-nAff (b)	UPUCB-nAff	TS
Gaussian	2	5	3	5	3	$8 \cdot 10^{-2}$
Bernoulli	$7 \cdot 10^{-7}$	$8 \cdot 10^{-5}$	$8 \cdot 10^{-5}$	$5 \cdot 10^{-7}$	$2 \cdot 10^{-6}$	$2 \cdot 10^{-7}$
Appendix H.4	0.6	0.5	0.4	1	0.9	$7 \cdot 10^{-2}$
Contextual		**C2UCB**	**C2UPUCB (b)**	**C2UPUCB**	**ε-greedy (b)**	**ε-greedy**
Logistic for counterfactual		2	2	2	0.2	0.1
Logistic for all		2	2	2	0.2	0.2

quite competitive, but C2UPUCB still has a slight advantage no matter whether the baseline is known or unknown.

For further comparison, we also plot the regret in terms of both the predicted uplifts (Left) and the rewards that the learner actually receives (Middle) in Fig. 4.[13] For the two regrets, we define the optimal action as the one that selects the 10 individuals with the highest predicted uplifts. Then, if the logistic models correctly predict the expected rewards, the regret in terms of the predicted uplifts should be the expectation of the regret in terms of the realized rewards, and thus we should get similar curves after averaging over 100 runs. This is indeed the case in Fig. 4b. Nonetheless, when the outcomes also come from true data, selecting the samples with higher predicted uplifts does not necessarily result in higher rewards, which explains the discrepancy between Fig. 4a (Left) and Fig. 4a (Middle). C2UPUCB and ε-greedy perform quite similarly here when we look at the realized rewards. Since the ground truth outcomes cannot be perfectly predict by logistic models, we also observe linear regret of all the algorithms in Fig. 4a (Left), while in Fig. 4b (Left), both C2UPUCB-M and C2UPUCB achieves sublinear regret and performs much better than ε-greedy.

In all the cases, Greedy and C2UCB perform poorly. This points out the importance of exploration (which is taken into account by bandit algorithms) and uplift modeling in this problem. Finally, to explain the surprising fact that an algorithm might perform better without knowledge of the baseline, we look into the data and find out that samples with higher baselines often have higher uplifts. Therefore, when the learner is provided with the baselines of the samples, they may avoid choosing these samples at the beginning, which results a slower exploration and a worse performance.

G Additional Experimental Details

In this section, we present missing experimental details from Section 8. The code implementation of the proposed methods is attached in supplementary material. All our experiments are run on amazon m4.xlarge EC2 instances. The Criteo Uplift Prediction Dataset that we use for our experiments has the 'CC0: Public Domain' license.

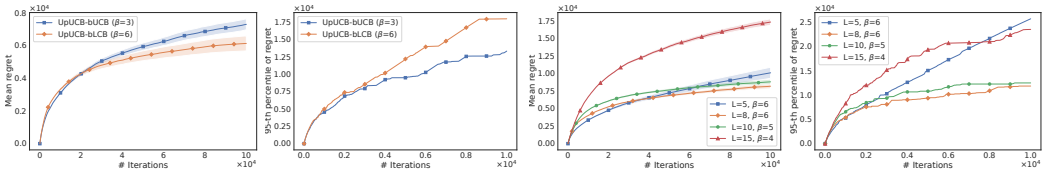
G.1 Choice of Algorithm Parameters

As mentioned in the main text, we report the results for the parameters that yield the best average performance. Precisely, we conduct 100 independent runs for each experimental setup, and in each run, we fix the random seed and test all the candidate algorithms with their parameters chosen from a predefined set. The performance of a specific parameter choice is then evaluated by the sum of the mean and one standard deviation of the regrets of the final run. In fact, when the algorithm puts too much weight on exploitation, we may get lower mean regret while the standard deviation gets extremely large because the algorithm fails more frequently.

Below we explain in detail what are the parameters that we tune in the non-contextual experiments. The optimal parameters that we have found are summarized in Table 3.

UCB and UPUCBs. For all UCB-type methods, we tune the exploration parameter $\beta = 1/\delta'$, which plays a crucial role in determining the widths of the confidence intervals– see (3) and (5). Since c_t^a

[13] We use occasionally the term predicted uplift here to remind that they are computed using logistic models fitted on the dataset and we do not have access to any type of *ground-truth* uplift.



(a) Baseline UCB versus Baseline LCB

(b) Consequence of misspecification of L

Figure 5: Ablation study on Gaussian uplifting bandit to investigate two aspects of our algorithms. Besides mean regret with standard error, we also plot the 95-th percentile of the regrets as an indicator for the robustness of a method.

represents the width of the confidence interval of a single variable's payoff in UpUCB algorithms, we set the widths of the confidence intervals of UCB to $mc_t^a = m\sqrt{2\beta/N_t^a}$, i.e., the UCB indices in UCB are $U_t^a = \hat{r}_t^a + mc_t^a$, where $\hat{r}_t^a$ is the empirical mean of action a's reward up to time t. In this way, we can choose exploration parameters of roughly the same order for different algorithms. In the Gaussian uplifting bandit experiment, we pick $\beta \in \{k \times 10^{-\gamma} : k \in [10], \gamma \in \{0, 1, 2\}\}$. As for the Bernoulli uplifting bandit experiment with Criteo dataset, we select $\beta \in \{k \times 10^{-\gamma} : k \in [10], \gamma \in \{5, 6, 7\}\}$. The optimal performances are achieved with much smaller exploration parameters here because the noises in the variables are independent.

Thompson Sampling. We use Thompson sampling with Gaussian prior and Gaussian noise. The mean and the variance of the prior are set to the mean and the variance of the actions' rewards. Therefore, Thompson sampling is actually provided with more information than UCB. Similar to before, the variance of the noise is set to $m^2\sigma^2$ since the noise in the total reward scales with m. The parameter σ^2 plays a similar role as β in UCB. In the Gaussian and Bernoulli uplifting experiments we respectively choose $\sigma^2 \in \{k \times 10^{-\gamma} : k \in [10], \gamma \in \{0, 1, 2\}\}$ and $\sigma^2 \in \{k \times 10^{-\gamma} : k \in [10], \gamma \in \{5, 6, 7\}\}$.

G.2 Payoff Construction for Bernoulli Uplifting Bandit with Criteo Uplift

To construct the payoffs in the Bernoulli uplifting experiment, we assume that each cluster contains a single type of users that react in the same pattern. That is, if a user of a certain type is treated, their payoff is sampled from a Bernoulli distribution with mean corresponding to the visit rate of the treated individuals of that cluster, otherwise their payoff is sampled from another Bernoulli distribution with mean corresponding to the visit rate of the untreated individuals of that cluster. The payoffs of the variables are sampled independently.

For sake of completeness, we report below the visit rate of the treated and the untreated individuals of the clusters (rounded to three decimal places) that we obtained in our experiments

$$\begin{aligned}\mu_{\text{treated}} = [&.001\ .037\ .003\ .001\ .003\ .377\ .237\ .309\ .071\ .287\\ &.531\ .044\ .007\ .086\ .002\ .019\ .028\ .007\ .265\ .013]\\ \mu_{\text{untreated}} = [&.001\ .023\ .002\ .002\ .004\ .289\ .206\ .229\ .073\ .289\\ &.464\ .035\ .004\ .052\ .001\ .011\ .022\ .004\ .165\ .000].\end{aligned}$$

The sizes of the clusters, which also correspond to $(L^a)_{a\in\mathcal{A}}$, are

$$\begin{aligned}[&10600\ 2764\ 7222\ 11128\ 6385\ 1630\ 2806\ 1089\ 3018\ 4594\\ &594\ 7020\ 12654\ 2186\ 9609\ 5101\ 3714\ 4569\ 1158\ 2159].\end{aligned}$$

We also provide the uplifts of the actions (rounded to one decimal place)

$$\begin{aligned}r_{\text{up}} = [&0.6\ 39.3\ 6.6\ -5.0\ -5.7\ 143.5\ 86.0\ 87.1\ -4.5\ -6.9\\ &39.8\ 66.3\ 34.7\ 75.4\ 4.0\ 40.3\ 21.1\ 12.3\ 115.7\ 28.3].\end{aligned}$$

H Ablation Study on Gaussian Uplifting Bandit

In this section, we present additional experimental results omitted in Section 8 that are conducted with the Gaussian uplifting bandit model. We follow the experimental procedures detailed in Appendix G.1 to find the optimal parameters for each considered algorithm.

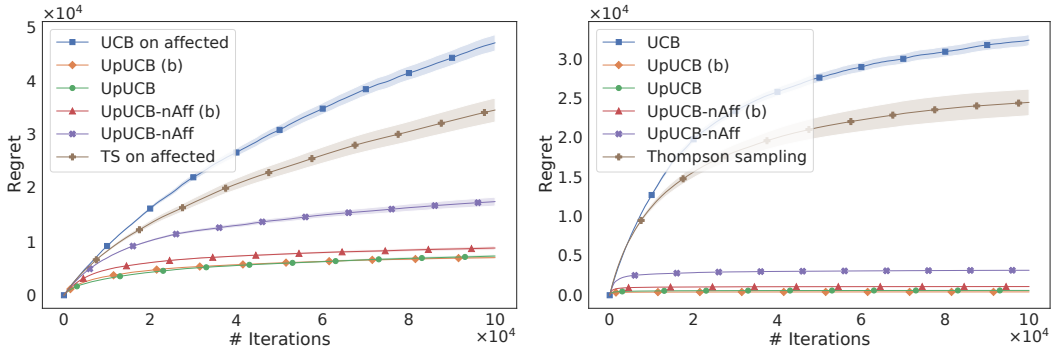


Figure 6: More experiments on Gaussian uplifting bandit. Left: UCB and Thompson sampling are run with rewards summed over the variables that are affected by at least on action. Right: test case for actions with difference covariance matrices.

H.1 Baseline UCB versus Baseline LCB.

In Algorithm 1, we suggest subtracting the UCBs of the baseline payoffs from the UCBs of the actions. This could be counter-intuitive because to get an optimistic estimate of the uplift, we should instead subtract the LCBs of the baseline payoffs. However, the latter strategy prevents the learner from learning more about the baselines that are badly estimated, and can thus lead to failure more frequently. In the following, we refer to the two strategies respectively as UpUCB (b)-bUCB and UpUCB (b)-bLCB.

In Fig. 5a (Left), we illustrate the mean regret of the two designs when their parameters are optimally tuned. It turns out there the two strategies perform similarly here. We would just like to highlight that to achieve the optimal performance, UpUCB (b)-bLCB needs to use a larger exploration parameter. Therefore, in terms of average performance, it seems that the effect of using LCBs of the baseline is close to a shrinkage of the exploration parameter.

Next, we further inspect the robustness of the algorithms by plotting the 95th percentile of the regrets (which roughly corresponds to the 5th largest regret among those recorded in the 100 runs). The same parameters give the optimal values for this metric. However, as shown in Fig. 5a (Right), it turns out that UpUCB (b)-bLCB achieves a much higher 95th percentile of regret, which indicates that it indeed fails more frequently.

H.2 Misspecification of L.

We also investigate the effect of misspecification of L on Algorithm 4. For the bandit instance that we employ here, $\max_{a \in \mathcal{A}} L^a = 10$ and thus according to Proposition 3, Algorithm 4 guarantees sublinear regret as long as it is provided with $L \geq 10$. On the flip side, the algorithm may incur linear regret if L is underestimated.

To verify this, we run the algorithm with $L \in \{5, 8, 10, 15\}$ and plot the averaged regret and the 95-th percentile of regret for optimally tuned exploration parameters in Fig. 5b. We effectively observe sublinear regret when L is overspecified, but the regret becomes much larger because the algorithm is more conservative and explores more. On the other hand, under severe underspecification of L, the algorithm is doomed to fail whatever the exploration parameter used. Finally, it seems that slight underspecification of L can be compensated by choosing a larger exploration parameter. In fact, in the figure we see that for $L = 8$ choosing $\beta = 6$ gives the smallest regret among the presented curves.

H.3 Ignoring Unaffected variables in UCB and Thompson Sampling

In the problem instance that we use for our Gaussian uplifting bandit experiments, not all the variables are affected by the actions. In fact, only $m' = 68$ out of the $m = 100$ variables are affected by at least one action. Since the remaining 32 variables contribute equally to all the actions, we can improve the efficiency of both UCB and Thompson sampling by using only rewards summed over the payoffs of the 68 variables that are actually affected by the actions. Nevertheless, as shown in Fig. 6 (left), the

resulting algorithms still achieve much larger regret than the ones that focus on the uplift, as L is still an order smaller than m'.

H.4 Actions with Different Covariance Matrices

For completeness, we further conduct experiments for the case where the actions have different covariance matrices. The results are reported in Fig. 6 (right). Here, the variances of the noises in the rewards vary from 16 to 80. Since the noises of some actions become smaller, we observe smaller regrets for all the algorithms. Nonetheless, the proposed methods are still much more effective than the algorithms that are agnostic to the structure of the problem.