Journal of Machine Learning Research 19 (2018) 1-40 Submitted 11/17; Revised 6/18; Published 9/18

Profile-Based Bandit with Unknown Profiles

Sylvain Lamprier sylvain.lamprier@lip6.fr
Sorbonne Universités, UPMC Paris 06, LIP6, CNRS UMR 7606

Thibault Gisselbrecht thibault.gisselbrecht@snips.ai
SNIPS, 18 rue Saint Marc, 75002 Paris

Patrick Gallinari patrick.gallinari@lip6.fr
Sorbonne Universités, UPMC Paris 06, LIP6, CNRS UMR 7606

Editor: Peter Auer

Abstract

Stochastic bandits have been widely studied since decades. A very large panel of settings have been introduced, some of them for the inclusion of some structure between actions. If actions are associated with feature vectors that underlie their usefulness, the discovery of a mapping parameter between such profiles and rewards can help the exploration process of the bandit strategies. This is the setting studied in this paper, but in our case the action profiles (constant feature vectors) are unknown beforehand. Instead, the agent is only given sample vectors, with mean centered on the true profiles, for a subset of actions at each step of the process. In this new bandit instance, policies have thus to deal with a doubled uncertainty, both on the profile estimators and the reward mapping parameters learned so far. We propose a new algorithm, called *SampLinUCB*, specifically designed for this case. Theoretical convergence guarantees are given for this strategy, according to various profile samples delivery scenarios. Finally, experiments are conducted on both artificial data and a task of focused data capture from online social networks. Obtained results demonstrate the relevance of the approach in various settings.

Keywords: Stochastic Linear Bandits, Profile-based Exploration, Upper Confidence Bounds

1. Introduction

Multi-armed bandits (MAB) correspond to online decision problems where, at each step of a sequential process, an agent has to choose an action - or arm - among a set of K actions, with the aim to maximize some cumulative reward function. In the so-called stochastic MAB setting, rewards collected for a given arm through time are assumed to be independently and identically distributed, following some hidden stationary distribution on every individual arm. The problem is therefore to deal with a tradeoff between exploitation - selecting actions according to some estimations about their usefulness - and exploration - selecting actions in order to increase the knowledge of their reward distribution. However, with classical stochastic bandit policies, the convergence towards the optimal arms can be slow when the number of actions becomes large.

On another hand, contextual bandits correspond to MAB settings where some side information can be leveraged to improve estimations of reward distributions. In these settings, a decision context is observed before selecting actions. This context can either

correspond to a global decision feature vector or to specific feature vectors observed for each single action. Depending on the setting, these features can vary over time, which can help to predict reward fluctuations, or they can correspond to constant features on actions - that we call action profiles in the following - whose structure can be used to improve exploration policies over non-contextual approaches. In this paper we address the latter case, where we assume stationary distributions of rewards, but where distributions depend on constant profiles associated each arm. Reward distributions from the different arms are connected by a common unknown parameter to be learned (Filippi et al., 2010).

However, we introduce a new scenario where, for various possible reasons (technical, political, etc...), profile vectors are not available a priori. Instead, the agent gets sample vectors, centered on the true profiles, for a subset of actions at each decision step. This can happen in various situations where some restrictions limit our knowledge of the whole decision environment. For example in a focused data capture or technology intelligence scenario on social media, where an agent is asked to collect relevant information w.r.t. to a given need. Because of the extremely large number of accounts on media such as Twitter, the agent needs to focus on a subset of relevant users to follow at each time step (Gisselbrecht et al., 2015). However, given the strict restrictions set by the media, no knowledge about users is available beforehand. Profiles have therefore to be constructed from users activities, which are only observed for a small fraction of users at each step. As we will see later, the process which delivers profile samples can either be independent - e.g., an external process delivers activity samples for randomly chosen users at each step in the case of data capture from Twitter - or be included in the decision process - e.g., activity samples are only collected for followed users at the current time step.

To the best of our knowledge, this instance of contextual bandit has not been studied in the literature. Existing bandit approaches do not fit with this new setting. First, even if traditional algorithms such as `UCB` (Auer et al., 2002) could be applied, the information provided by the sample profile vectors would be entirely ignored. Our claim is that important benefits can arise from taking this available side-information into account. On the other hand, existing contextual bandit policies do not take into account uncertainty on context vectors, while we face here a bandit problem where uncertainty not only arises from regression parameters, as classically considered by contextual approaches, but also from the estimated profiles which serve as inputs for reward predictions at each step. The aim is to propose an approach able to leverage structural distributions of arms, based on noisy observations of their profiles, to improve over existing exploitation/exploration policies in bandit settings with partial side-information.

The contribution of this paper is threefold:

- We propose a new instance of the contextual bandit problem, based on constant contexts for each action, where action profiles are not known beforehand, but built from samples obtained at each iteration (3 cases are investigated regarding the sampling process);
- We design the *SampLinUCB* algorithm to solve this problem, for which we demonstrate some theoretical convergence guarantees;
- We experiment our proposal for both an artificial setting and a real-world task of focused data capture from Twitter, to empirically demonstrate the benefits of such a

profile-based approach with ony partial knowledge for exploitation/exploration problems.

The paper is organized as follows. In section 2, we present some background and related works. In section 3, we formalize the problem, propose our algorithm and derive the regret bound. Finally, section 4 reports our experiments.

2. Background: the Linear Stochastic Bandit

The multi-armed bandit problem, originally introduced in (Lai and Robbins, 1985) in its stationary form, has been widely studied in the literature. This learning problem aims at tackling the trade off between exploration and exploitation in decision processes where, at each step, an agent must choose an action - or arm - among a finite set of size K. After each decision, it receives a reward which quantifies the quality of the chosen action. The aim for the agent is to maximize the cumulative reward trough time, or equivalently to minimize the cumulative regret R_T at step T defined as:

$$R_T = \max_{i \in \{1,2,\dots,K\}} \sum_{t=1}^{T} r_{i,t} - \sum_{t=1}^{T} r_{i_t,t} \tag{1}$$

where i_t stands for the action selected at step t and $r_{i,t}$ the reward obtained by playing the action i at step t. This represents the amount of rewards that has been lost by selecting i_t at each step t, compared to what could be obtained by playing the optimal arm from the beginning to the end of the process. Note that, at each step t, only the reward for the chosen action $r_{i_t,t}$ is observed in practice, other ones remain unknown.

In the so-called stochastic case, one assume that rewards of an arm i are identically and independently sampled from a distribution with mean ν_i. Therefore, one usually rather consider the pseudo-regret of a policy, which introduces expectations of regret in the previous definition:

$$\hat{R}_T = T\nu_{i^*} - \sum_{t=1}^{T} \nu_{i_t} \tag{2}$$

where $i^\star$ stands as the arm with the best reward expectation.

One of the simplest and most straightforward algorithms to deal with the stochastic bandit problem is the well-known ϵ-greedy algorithm (Auer et al., 2002). This algorithm selects the arm with the best reward mean empirical estimation with probability $1 - \epsilon$ and uniformly selects an arm among the whole set regardless their current estimations with probability ϵ. This guarantees to regularly reconsider estimations of all arms and therefore prevents from getting stuck on sub-optimal arms. However, the reward loss resulting from these blind selections prevents from ensuring a sub-linear upper bound of the pseudo-regret, unless setting an appropriate decay on ϵ. But this requires to know a lower bound on the difference of reward expectations between the best and the second best action (Auer et al., 2002).

Upper Confidence Bound algorithms (UCB) is another family of bandit approaches which define confident intervals for the reward expectations of each arm. Based on some concentration inequalities (Hoeffding, Bernstein, etc.), they propose optimistic policies which

consider possible deviations of the estimated mean of each arm. By using upper bounds of confidence intervals as selection scores, they ensure a cleaver balance between exploitation and exploration. Many extensions of the famous `UCB` algorithm proposed in (Auer et al., 2002) are known to guarantee a sub-linear bound of the pseudo-regret (see `UCBV` in (Audibert et al., 2009), `MOSS` in (Audibert and Bubeck, 2009) or `KL-UCB` in (Garivier, 2011)).

At last, Thompson sampling algorithms, originally proposed in (Thompson, 1933), develop a Bayesian approach to deal with uncertainty. By sampling from posterior distributions for the reward parameters, their exploration/exploitation mechanism is also proved to ensure a sub-linear regret (see (Kaufmann et al., 2012b) and (Agrawal and Goyal, 2012)).

The contextual bandit setting is an instance of the bandit problem where context vectors are observed before each decision step. Typically, contextual bandits assume a linear relation between context features and reward expectations. Formally, if we observe a context vector $x_{i,t} \in \mathbb{R}^d$ for each action $i \in \mathcal{K}$ at each time-step t, we consider the following assumption:

$$\exists \beta \in \mathbb{R}^d \text{ such that } r_{i,t} = x_{i,t}^\top \beta + \eta_{i,t} \tag{3}$$

where β is a mapping parameter between contexts and rewards, $\eta_{i,t}$ is a zero-mean conditionally R sub-Gaussian random noise, with constant $R > 0$ i.e: $\forall \lambda \in \mathbb{R} : \mathbb{E}[e^{\lambda \eta_{i,t}} | \mathcal{H}_{t-1}] \leq e^{\lambda^2 R^2/2}$, with $\mathcal{H}_{t-1} = \{(i_s, x_{i_s,s}, r_{i_s,s})\}_{s=1..t-1}$.

In this context, given a set $\mathcal{K}$ of K actions, any contextual bandit algorithm proceeds at each step $t \in \{1, 2, 3, \ldots, T\}$ as follows:

1. Observation of the context vector $x_{i,t} \in \mathbb{R}^d$ for each $i \in \{1, ..., K\}$;
2. According to the current estimate of β, selection of an action i_t and reception of the associated reward $r_{i_t,t}$;
3. Improvement of the selection policy by considering the new input $(i_t, x_{i_t,t}, r_{i_t,t})$ for the estimation of β.

Various contextual algorithms have been proposed in the literature. The first contextual bandit algorithm was introduced in (Auer, 2003). More recently the well-known `LinUCB` algorithm has been proposed for a task of personalized recommendation in (Li et al., 2010) and analyzed in (Chu et al., 2011). Both of these algorithms are UCB-like policies, each of them selecting the action whose upper bound of its reward confidence bound is the highest. Many other UCB approaches have been developed since then. In particular, algorithms such as `OFUL` or `ConfidenceBall` proposed in (Abbasi-Yadkori et al., 2011) and (Dani et al., 2008) have the advantage to enjoy a tighter regret upper bound (see also (Kaufmann et al., 2012a) and (Rusmevichientong and Tsitsiklis, 2010)). As in the stochastic bandit setting, Thompson sampling algorithms have also been designed for the contextual case, which also proved to be powerful, first empirically in (Chapelle and Li, 2011) and then theoretically in (Agrawal and Goyal, 2013) and (May et al., 2012).

In this paper, we consider a variant of the contextual bandit problem where contexts of actions are constant, which we call action profiles in the following. Hence, in our setting we assume that each action $i \in \mathcal{K}$ is associated with a profile vector $\mu_i \in \mathbb{R}^d$. The linear assumption of equation 3 becomes:

$$\exists \beta \in \mathbb{R}^d \textit{ such that } r_{i,t} = \mu_i^\top \beta + \eta_{i,t} \tag{4}$$

Thus, in this setting contexts cannot be used to anticipate some variations in the rewards expectations as it is traditionally the case in the literature about contextual bandit, but they can be leveraged to improve the exploration process, the use of a shared mapping parameter β allowing one to define areas of interest in the representation space of the actions. To illustrate this, the figure 1 represents the selection scores of a contextual algorithm (such as *OFUL* that we rely on in the following) at a given step for a simple case where $K = 4$ and $d = 2$. In this figure, green areas correspond to high scores, whereas red ones correspond to low scores areas. Color variations render the latent structure inferred by the model. In this setting, the algorithm would select the action 1, since its profile is located in the most promising area of the space. On the other hand, the action 3 is located in an area that is greatly less promising. The fact of using a common mapping parameter β allows one to perform a mutual learning, where observations on some actions inform on the usefulness of similar ones. This allows one to improve the exploration process by focusing more quickly on the useful areas: Imagine that a great number of actions are located in the red area of the figure. In that case, a classical bandit algorithm such as *UCB* would need to consider each of these actions several times to reveal their low reward expectation. On the other hand, a contextual bandit algorithm such as *OFUL* is able to avoid these actions really more quickly because of the proximity with other bad actions.

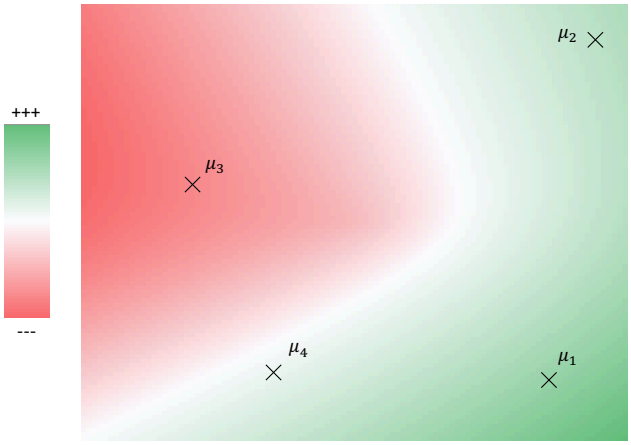


Figure 1: Illustration of `OFUL` scores for a profiles representation space.

This structured bandit setting has already been investigated in (Filippi et al., 2010). This work showed that great improvements could indeed be obtained by exploiting the structure of actions, since observations on some actions inform on the usefulness of similar ones. This comes down to a classical stochastic bandit where the pseudo-regret can be defined as follows:

$$\hat{R}_T = \sum_{t=1}^{\top} \mu_{i^\star}^\top \beta - \mu_{i_t}^\top \beta \tag{5}$$

where $\mu_i \in \mathbb{R}^d$ stands for the profile of the action $i \in \mathcal{K}$ and $\mu_{i^\star} = \arg\max_{\mu_i, i=1..K} \mu_i^\top \beta$ corresponds to the profile of the optimal action $i^\star$.

This is the setting which is studied in this paper. However, in our case the profile vectors are **unknown** to the agent beforehand, they have to be discovered iteratively during the process. Our problem differs from existing instances by the following two main aspects:

1. Action profiles are not directly available, one only get samples centered on them during the process;

2. At each step, one only get samples for a subset of actions.

In the following, we derive an UCB-based policy for this new setting.

3. Profile-Based Bandit with Unknown Profiles

In this section, we explore our new setting in which the set of profiles $\{\mu_1, .., \mu_K\}$ is not directly observed. Instead, at each iteration t, the agent is given a subset of actions $\mathcal{O}_t$ such that for every $i \in \mathcal{O}_t$, a sample $x_{i,t}$ of a random variable centered on μ_i is revealed. By assuming the same linear hypothesis as described in the previous section (formula 4), the relation of rewards with profiles can be re-written as follows for any time t, in order to introduce profile samples:

$$\begin{aligned} \forall s \leq t : r_{i,s} &= \mu_i^\top \beta + \eta_{i,s} \\ &= \hat{x}_{i,t}^\top \beta + (\mu_i - \hat{x}_{i,t})^\top \beta + \eta_{i,s} \\ &= \hat{x}_{i,t}^\top \beta + \epsilon_{i,t}^\top \beta + \eta_{i,s} \end{aligned} \tag{6}$$

where $\epsilon_{i,t} = \mu_i - \hat{x}_{i,t}$, $\hat{x}_{i,t} = \frac{1}{n_{i,t}} \sum_{s \in T_{i,t}^{obs}} x_{i,s}$, with $T_{i,t}^{obs} = \{s \leq t, i \in \mathcal{O}_s\}$ and $n_{i,t} = |T_{i,t}^{obs}|$. In words, $n_{i,t}$ corresponds to the number of times a sample has been obtained for the action i until step t and $\hat{x}_{i,t}$ corresponds to the empirical mean of observed samples for i at time t. $\epsilon_{i,t}$ corresponds to the deviation of the estimator $\hat{x}_{i,t}$ from the true profile μ_i.

Compared to traditional contextual bandits, the uncertainty is double : as classically it arises from the β parameter estimator, but also from the profile estimators, since the algorithm must both estimate β and the profile vectors $\{\mu_1, .., \mu_K\}$ from observations. Figure 2[1] illustrates this new setting. Contrary to figure 1 where profiles are known, here we only get confidence areas for them, represented by circles centered on their corresponding empirical mean (represented by a blue cross). From the law of large numbers, the more observations for a given action we get, the lower the deviation between its true profile and its empirical estimator is. Therefore, the more knowledge we get about a given action, the smaller its confidence circle is. The best action is still the action 1, whose true profile (represented by a black cross) is in the greenest area. However, this information is unknown from the agent. A naive solution would be to directly use the empirical mean for each action in order to determine the selection scores. From the figure, this would lead to select the sub-optimal

1. Note that this is only an illustration of the general principle, in practice the surface of selection scores should also differ from figure 1, since β is estimated from biased inputs.

action 2, whose empirical mean is located in a greener area than the one of other actions. We propose to include the additional uncertainty in the selection scores by using the best location inside the confidence ellipsoid it is possible to reach for each action. This allows one to define an optimistic policy which would select the optimal action 1 in the example figure, whose confidence area contains the most promising profiles (i.e., includes the most green locations in the figure).

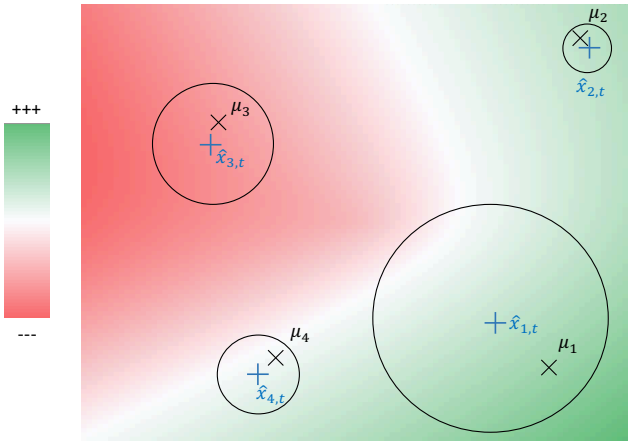


Figure 2: Illustration of the additional uncertainty arising from profile estimators.

In the following we propose to define an algorithm fitted for this setting. After deriving the algorithm in a generic context of samples delivery, we consider three different cases:

- **Case 1**: Every action delivers a profile sample at each step t (i.e., $\forall t$, $\mathcal{O}_t = \{1, ..., K\}$);
- **Case 2**: Each action i owns a probability p_i of sample delivery: at each step t, an action is included in $\mathcal{O}_t$ with probability p_i;
- **Case 3**: At each step t, only the action selected at the previous step delivers a sample (i.e., $\forall t$, $\mathcal{O}_t = i_{t-1}$).

The first case corresponds to the simplest case, where $\mathcal{O}_t$ is constant over time. The second case includes an additional difficulty since before any step, every action has not been observed the same number of times, which leads to different levels of uncertainty. The last case, probably the most interesting one for real-world applications, is the most difficult since decisions at each step not only affect knowledge about reward distributions but also profile estimations. For that case, it appears mandatory to take the uncertainty about profiles into account in the selection policy to guarantee the process to converge towards optimal actions.

After deriving confidence intervals for this new bandit instance, this section describes the proposed algorithm and analyzes its regret for the three settings listed above. Then, we consider the case where multiple actions can be performed at each step.

3.1. Confidence Intervals

In the following, we derive a series of propositions which will allow us to define a selection policy for our setting of profile-based bandit with unknown profiles. First, it is needed to define an estimator for the mapping parameter β, and the associated confidence ellipsoid. For that purpose, we rely on results from the theory of the self-normalized process (de la Peña et al., 2009). The next step is to define a way to mix it with the uncertainty on profiles to define an optimistic policy.

Proposition 1 *Let us consider that for any action i, all profile samples $x_{i,t} \in \mathbb{R}^d$ are iid from a distribution with mean $\mu_i \in \mathbb{R}^d$. Let us also assume that there exists a real number $L > 0$ such that $||x_{i,t}|| \leq L$ and a real number $S > 0$ such that $||\beta|| \leq S$. Then, for any $i \in \mathcal{K}$ and any step $s \leq t$, the random variable $\eta_{i,t,s} = \epsilon_{i,t}^\top \beta + \eta_{i,s}$ is conditionally sub-Gaussian with constant $R_{i,t} = \sqrt{R^2 + \frac{L^2 S^2}{n_{i,t}}}$.*

Proof Available in appendix A.1. ■

In this proposition, R is the constant of the sub-Gaussian random noise of the rewards (η_s from equation 6) and the notation $||x||$ stands as the norm of a vector x (i.e., $||x|| = \sqrt{x^T x}$). Since the noise $\eta_{i,t,s}$ is sub-Gaussian, it will be possible to apply the theory of self-normalized process for defining a confidence ellipsoid for β.

At step t, we can use the following set of observations to find an estimator for β: $\{(\hat{x}_{i_s,t}, r_{i_s,s})\}_{s=1..t-1}$ (i.e., at any decision step t, the reward $r_{i_s,s}$ observed at any previous step $s < t$ is associated with the profile of the selected action at step s, estimated knowing samples observed from step 1 to step t). The following notations are used in the remaining of the paper:

- $\eta'_{t-1} = (\eta_{i_s,s} + \epsilon_{i_s,t}^\top \beta)_{s=1..t-1}^\top$ the vector of noises of size $t-1$.
- $X_{t-1} = (\hat{x}_{i_s,t}^\top)_{s=1..t-1}$ the $(t-1) \times d$ matrix containing the empirical means of the selected actions, where the s-th row corresponds to the estimator at step t of the action selected at step s.
- $Y_{t-1} = (r_{i_s,s})_{s=1..t-1}^\top$ the rewards vector of size $t-1$.
- $A_{t-1} = diag(1/R_{i_s,t})_{s=1..t-1}$ the diagonal $(t-1) \times (t-1)$ matrix, where the s-th diagonal element equals $1/R_{i_s,t}$. Note that, for a specific action, the value of its corresponding coefficient increases with the number of observed samples for this action.

With these notations, the linear application from profiles to rewards can be written as:

$$Y_{t-1} = X_{t-1}\beta + \eta'_{t-1} \tag{7}$$

Proposition 2 *We note $\hat{\beta}_{t-1}$ the least square estimator of the parameter β at step t, according to the following l^2-regularized regression problem, where each element is weighted*

by the corresponding coefficient $1/R_{i_s,t}$*:*

$$\hat{\beta}_{t-1} = \arg\min_{\beta} \sum_{s=1}^{t-1} \frac{1}{R_{i_s,t}} (\beta^\top \hat{x}_{i_s,t} - r_{i_s,s})^2 + \lambda ||\beta||^2 \tag{8}$$

where $\lambda > 0$ *is the l2-regularization constant.*

We have:

$$\hat{\beta}_{t-1} = (X_{t-1}^\top A_{t-1} X_{t-1} + \lambda I)^{-1} X_{t-1}^\top A_{t-1} Y_{t-1} \tag{9}$$

Proof Let us rewrite the minimization problem such as:
$\hat{\beta}_{t-1} = \arg\min_{\beta} L$ with $L = (Y_{t-1} - X_{t-1}\beta)^\top A_{t-1}(Y_{t-1} - X_{t-1}\beta) + \lambda \beta^\top \beta$.
The gradient is given by:
$\nabla_\beta L = -2X_{t-1}^\top A_{t-1}(Y_{t-1} - X_{t-1}\beta) + 2\lambda\beta = 2(X_{t-1}^\top A_{t-1} X_{t-1} + \lambda I)\beta - 2X_{t-1}^\top A_{t-1} Y_{t-1}$
By canceling this gradient, we get the announced result. ■

This estimator of β uses empirical means of observed samples as inputs. Weighting each element according to the corresponding value $R_{i_s,t}$ allows one to consider the uncertainty associated with this approximation. It renders the confidence we have in the weighted input. Note that this coefficient tends towards a constant when the number of observed samples increases for the corresponding action. It allows one, according to the following proposition, to define a confidence ellipsoid for the estimator of β.

Proposition 3 *Let us define* $V_{t-1} = \lambda I + X_{t-1}^\top A_{t-1} X_{t-1} = \lambda I + \sum_{s=1}^{t-1} \frac{\hat{x}_{i_s,t} \hat{x}_{i_s,t}^\top}{R_{i_s,t}}$*. With the same assumptions as in proposition 1, for any* $0 < \delta < 1$*, with a probability at least equal to* $1 - \delta$*, the estimator* $\hat{\beta}_{t-1}$ *verifies for all* $t \geq 0$*:*

$$||\hat{\beta}_{t-1} - \beta||_{V_{t-1}} \leq \sqrt{2 \log \left(\frac{det(V_{t-1})^{1/2} det(\lambda I)^{-1/2}}{\delta} \right)} + \sqrt{\lambda} S = \alpha_{t-1} \tag{10}$$

where $||x||_V = \sqrt{x^\top V x}$ *is the V-norm of the vector* x*.*

Proof Available in appendix A.2. ■

This bound is very similar to the one defined in the `OFUL` algorithm (Abbasi-Yadkori et al., 2011) to build its confidence ellipsoid. However, a notable difference lies in the definition of the matrix V_t, in which weights in A_t are applied to cope with confidence differences between profile estimators. Without this weighting, no confidence ellipsoid could be found for β since no common bound could be defined for the various noises η_s' (see the proof of proposition 3 in appendix).

The following proposition can easily be deduced from the previous one to bound the expectation of reward with known profiles.

Proposition 4 *For every $i \in \mathcal{K}$, with probability greater than $1-\delta$, we have for all $t \geq 0$:*

$$\beta^\top \mu_i \leq \hat{\beta}_{t-1}^\top \mu_i + \alpha_{t-1} ||\mu_i||_{V_{t-1}^{-1}} \tag{11}$$

Proof Available in appendix A.3 ■

This upper-bound for the expected reward contains two distinct terms: while the former corresponds to a classical exploitation term which estimates the expected reward with the current parameters, the latter corresponds to an exploration term since it takes into account the uncertainty on the reward parameter. If profiles were known, this could directly be used as a selection score for an UCB-like policy. However, in our setting, profiles are unknown. We have to consider confidence ellipsoids for the profiles of actions too. The following proposition defines confidence bounds for the profile estimators.

Proposition 5 *For every $i \in \mathcal{K}$ and any $t > 0$, with probability greater than $1 - \delta/t^2$, we have:*

$$||\hat{x}_{i,t} - \mu_i|| \leq \min(L\sqrt{\frac{2d}{n_{i,t}} \log\left(\frac{2dt^2}{\delta}\right)}, 2L) = \rho_{i,t,\delta} \tag{12}$$

Proof This inequality comes from the application of the Hoeffding's inequality to each dimension separately. The min operator comes from the base hypothesis $||x_{i,t}|| \leq L$, which can be more restrictive than the Hoeffding assumption. The proof is available in appendix A.4. ■

Contrary to the bound of the deviation of the mapping parameter β which holds simultaneously for all steps of the process, the one for the profile estimators is only valid for each step separately. To obtain a bound holding for every step simultaneously, which is important for the regret analysis (see section 3.3), we use the uniform bound principle. For a given action i, we have: $\mathbb{P}(\forall t, ||\hat{x}_{i,t} - \mu_i|| \leq \rho_{i,t,\delta}) = 1 - \mathbb{P}(\exists t, ||\hat{x}_{i,t} - \mu_i|| \geq \rho_{i,t,\delta}) \geq 1 - \sum_t \mathbb{P}(||\hat{x}_{i,t} - \mu_i|| \geq \rho_{i,t,\delta}) \geq 1 - \sum_t \delta/t^2$. This justifies the introduction of the t^2 term in the bound, which allows one to define a uniform probability over all steps since we have thereby: $\mathbb{P}(\forall t, ||\hat{x}_{i,t} - \mu_i|| \leq \rho_{i,t,\delta}) \geq 1 - \delta - \sum_{t=2}^{\infty} \delta/t^2 = 1 - \delta - \delta(\pi^2/6 - 1) \geq 1 - 2\delta$.

Now that we have defined probabilistic deviation bounds for the different estimators, we can use them conjointly to define the confidence interval of the reward expectation for the setting of unknown profiles, and thus to upper bound the expected reward for each action i.

Proposition 6 *For every $i \in \mathcal{K}$ and any $t > 0$, with probability greater than $1 - \delta/t^2 - \delta$, we have:*

$$\beta^\top \mu_i \leq \hat{\beta}_{t-1}^\top (\hat{x}_{i,t} + \bar{\epsilon}_{i,t}) + \alpha_{t-1} ||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} \tag{13}$$

with: $\bar{\epsilon}_{i,t} = \dfrac{\rho_{i,t,\delta} \hat{\beta}_{t-1}}{||\hat{\beta}_{t-1}||}$ $\qquad$ $\tilde{\epsilon}_{i,t} = \dfrac{\rho_{i,t,\delta} \hat{x}_{i,t}}{\sqrt{\lambda} ||\hat{x}_{i,t}||_{V_{t-1}^{-1}}}$

Proof The proof is available in appendix A.5. ■

Compared to the bound given in proposition 4, we find the same two terms of exploitation and exploration. However in this case, profile vectors that are unknown are replaced by the estimator plus an additional term: $\bar{\epsilon}_{i,t}$ in the former part and $\tilde{\epsilon}_{i,t}$ in the latter one. These terms aim at coping with profile uncertainty, considering confidence ellipsoids for these profiles as defined in proposition 5. $\bar{\epsilon}_{i,t}$ is collinear with $\hat{\beta}_{t-1}$. It is used to translate the estimator $\hat{x}_{i,t}$ so that $\beta^\top \mu_i$ is upper-bounded. $\tilde{\epsilon}_{i,t}$ is collinear with $\hat{x}_{i,t}$. It is used to translate the estimator $\hat{x}_{i,t}$ so that the V_{t-1}^{-1}-norm $||\mu_i||_{V_{t-1}^{-1}}$ is upper-bounded. This bound enables us to derive an optimistic policy in the next section.

3.2. SampLinUCB

In this section, we detail our policy for the setting of unknown profiles, called *SampLinUCB*, which is directly derived from the bound proposed in proposition 6. Its process is detailed in algorithm 1. In words, it proceeds as follows:

1. Initialization of the shared variables V and b used to estimate the mapping parameter β in lines 1 and 2. The $d \times d$ matrix V is initialized with an identity matrix times the regularization parameter λ (the greater λ is, the more the parameter β will be constrained to have components close to zero). The vector b is initialized as a null vector of size d.

2. Initialization of the individual variables n_i, $\hat{x}_i$, R_i, N_i and S_i for every action in $\mathcal{K}$ (lines 3 to 6). The two latter are additional scalar variables which enable efficient updates for the shared variables after context observations. N_i counts the number of times an action has been selected from the beginning, S_i sums the rewards collected by selecting i from the beginning.

3. At each iteration t, for each action $i \in \mathcal{O}_t$, observation of the sample $x_{i,t}$ (line 11) and update of individual variables n_i, $\hat{x}_i$ and R_i for action i (line 12) and shared parameters V and b according to these new individual values for i (line 10 and 13). Since shared parameters are simple sums of elements, they can be simply updated by first removing old values (line 10) and then adding the new ones when updated (line 13). This is efficiently done without requiring an important memory load thanks to scalar variables N_i and S_i.

4. Computation of the selection score $s_{i,t}$ (line 21) for each action i according to equation 14 detailed below, and selection of the action associated with the highest selection score (line 23) (except in the early stages $\leq K$ where all actions are selected in turn to initialize their counts in line 16).

5. Collection of the associated reward (line 25) and update of variables N_i, S_i, V and b according to this new outcome (lines 26 to 28).

Algorithm 1: `SampLinUCB`

$V = \lambda I_{d\times d}$ (Identity matrix of size d);
$b = 0_d$ (Null vector of size d);
for $i \in \mathcal{K}$ **do**
    $N_i = 0$; $S_i = 0$;
    $n_i = 0$; $\hat{x}_i = 0_d$; $R_i = +\infty$;
end
for $t = 1..T$ **do**
    Reception of $\mathcal{O}_t$;
    for $i \in \mathcal{O}_t$ **do**
        $V = V - N_i \frac{\hat{x}_i \hat{x}_i^\top}{R_i}$; $b = b - S_i \frac{\hat{x}_i}{R_i}$;
        Observation of $x_{i,t}$;
        $n_i = n_i + 1$; $\hat{x}_i = \frac{(n_i - 1)\hat{x}_i + x_{i,t}}{n_i}$; $R_i = \sqrt{R^2 + \frac{L^2 S^2}{n_i}}$;
        $V = V + N_i \frac{\hat{x}_i \hat{x}_i^\top}{R_i}$; $b = b + S_i \frac{\hat{x}_i}{R_i}$;
    end
    if $t \leq K$ **then**
        Selection of $i_t = t$;
    end
    else
        $\hat{\beta} = V^{-1} b$;
        for $i \in \mathcal{K}$ **do**
            Computation of $s_{i,t}$ according to formula 14 ;
        end
        Selection of $i_t = \arg\max_{i \in \mathcal{K}} \quad s_{i,t}$;
    end
    Reception of $r_{i_t,t}$;
    $N_{i_t} = N_{i_t} + 1$;
    $S_{i_t} = S_{i_t} + r_{i_t,t}$;
    $V = V + \frac{\hat{x}_{i_t} \hat{x}_{i_t}^\top}{R_{i_t}}$; $b = b + r_{i_t,t} \frac{\hat{x}_{i_t}}{R_{i_t}}$;
end

The selection score $s_{i,t}$ used in our policy for each action i at any step t is directly derived from proposition 6:

$$s_{i,t} = (\hat{x}_{i,t} + \overline{\epsilon}_{i,t})^\top \hat{\beta}_{t-1} + \alpha_{t-1} ||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} \tag{14}$$

For cases 2 and 3 of the profile delivery mechanism (see at the beginning of the section), there are actions i with $n_{i,t} = 0$ in the early steps of the process. No sample has ever been observed for these actions, which is problematic for the computation of $\rho_{i,t,\delta}$, and therefore

the computation of $\bar{\epsilon}_{i,t}$ and $\tilde{\epsilon}_{i,t}$. For the case 2, where we are not active on the process for observing contexts, this can be solved by simply ignoring actions until at least one sample of profile has been observed for them. For the case 3 however, samples are only obtained by selection. Thus, we need to force the observation of a sample for every action in the first steps. In that way, for actions with $n_{i,t} = 0$ at any step t, we arbitrarily set $s_{i,t} = +\infty$ in order to make the policy favor actions without any knowledge to initialize the process. Thus, in that case, algorithm 1 selects the K actions in turn in the K first steps of the process.

The selection score defined in formula 14 corresponds to the upper-bound of the expected reward for each action, as it is done in all UCB-based policies. Intuitively, it leads the algorithm to select actions whose profile estimator is either in an area with high potential, or is sufficiently uncertain to consider still likely that the action can be potentially useful. The goal is to quickly rule out bad actions, whose confidence ellipsoid does not include any potentially useful locations w.r.t. the current estimation of β. To better analyze the algorithm, we propose below a new formulation of the selection score.

Proposition 7 *The score $s_{i,t}$ from equation 14 can be re-written in the following way:*

$$s_{i,t} = \hat{x}_{i,t}^\top \hat{\beta}_{t-1} + \alpha_{t-1} ||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + \rho_{i,t,\delta} \left(||\hat{\beta}_{t-1}|| + \frac{\alpha_{t-1}}{\sqrt{\lambda}} \right) \tag{15}$$

Proof

$$\begin{aligned}
s_{i,t} &= (\hat{x}_{i,t} + \bar{\epsilon}_{i,t})^\top \hat{\beta}_{t-1} + \alpha_{t-1} ||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} \\
&= \hat{x}_{i,t}^\top \hat{\beta}_{t-1} + \frac{\rho_{i,t,\delta} \hat{\beta}_{t-1}^\top}{||\hat{\beta}_{t-1}||} \hat{\beta}_{t-1} + \alpha_{t-1} ||\hat{x}_{i,t} + \frac{\rho_{i,t,\delta} \hat{x}_{i,t}}{\sqrt{\lambda} ||\hat{x}_{i,t}||_{V_{t-1}^{-1}}}||_{V_{t-1}^{-1}} \\
&= \hat{x}_{i,t}^\top \hat{\beta}_{t-1} + \rho_{i,t,\delta} ||\hat{\beta}_{t-1}|| + \alpha_{t-1} \left(1 + \frac{\rho_{i,t,\delta}}{\sqrt{\lambda} ||\hat{x}_{i,t}||_{V_{t-1}^{-1}}} \right) ||\hat{x}_{i,t}||_{V_{t-1}^{-1}} \\
&= \hat{x}_{i,t}^\top \hat{\beta}_{t-1} + \alpha_{t-1} ||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + \rho_{i,t,\delta} \left(||\hat{\beta}_{t-1}|| + \frac{\alpha_{t-1}}{\sqrt{\lambda}} \right)
\end{aligned}$$

■

This new formulation of the selection score allows one to take a different look at the algorithm behavior. The first part of the score $\hat{x}_{i,t}^\top \hat{\beta}_{t-1} + \alpha_{t-1} ||\hat{x}_{i,t}||_{V_{t-1}^{-1}}$ is similar to a score that would use the classical `OFUL` algorithm (although with a different construction of V_t), with an exploitation term and a classical exploration term considering the uncertainty on the estimator of β. But it exhibits an additional part $\rho_{i,t,\delta} \left(||\hat{\beta}_{t-1}|| + \frac{\alpha_{t-1}}{\sqrt{\lambda}} \right)$ which is directly proportional to the coefficient $\rho_{i,t,\delta}$ and thus enables some exploration w.r.t. the uncertainty of the profile estimators. This highlights the premium granted to less observed actions. Note that for the case 1, this additional part is the same for every action. It therefore could be removed from the score since it does not permit to discriminate some action w.r.t any other one. However, this new exploration term is particularly useful for the

case 3, where observations of samples are directly connected to the selection policy, since it prevents from moving aside some optimal actions that have unluckily provided only not promising samples in the early steps.

To demonstrate that considering uncertainty on profiles is crucial in that case, let us consider a scenario where the optimal action $i^\star$ gets a null vector as the first profile sample. Then, in a setting where all profile samples are in $[0, L]^d$ and all rewards are in $[0, +\infty]$, it suffices that a sub-optimal action i gets a non-null vector as the first profile sample and a positive value as the first reward to lead to a linear regret from a given step. Indeed, since we only get samples with all components greater or equal than 0, i will never get a null vector as a profile estimator. On the other hand, while $i^\star$ is not selected, its profile estimator cannot change from the null vector. Thus, with a naive algorithm that would not include translations w.r.t. uncertainty of profiles, we would have $s_{i^\star,t} = \hat{x}_{i^\star,t}^\top \hat{\beta}_{t-1} + \alpha ||\hat{x}_{i^\star,t}||_{V_{t-1}^{-1}} = 0$ for all t until $i_t = i^\star$. Now, the least square estimator of β approximates observed reward values from estimated profiles. Since we have at least one non-null reward associated with a non-null profile estimator, β will always output a positive expected reward at least for one action. Thus, there is always an action i' with $s_{i',t} > 0$, which prevents from selecting the optimal action until the end of the process. This shows that a naive algorithm is not well fitted here, since it is likely to stay stuck on sub-optimal actions because of wrong knowledge about profiles. The point is now to show that the proposed additional term enables to solve this problem and ensures a sub-linear pseudo regret for our profile-based bandit algorithm with unknown profiles.

3.3. Regret

The following proposition establishes an upper bound for the cumulative pseudo-regret of the `SamplinUCB` algorithm proposed above. This is a generic bound for which no assumption is done on the process generating $\mathcal{O}_t$ at each step t.

Proposition 8 (Generic bound) *By choosing $\lambda \geq \max(1, L^2/\sqrt{R^2})$, with a probability greater than $1 - 3\delta$, the cumulative pseudo-regret of the algorithm `SampLinUCB` is upper-bounded by:*

$$\begin{aligned}\hat{R}_T \leq C + 4L\left(\sqrt{\frac{d}{\lambda}\log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + 2S\right)\sqrt{2d\log\left(\frac{2dT^2}{\delta}\right)}\sum_{t=1}^{T}\frac{1}{\sqrt{n_{i_t,t}}} \\ + 2\left(\sqrt{d\log\left(\frac{1+TL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S\right) \\ \times\sqrt{Td\left(\sqrt{R^2+L^2S^2}\log\left(1+\frac{TL^2}{\lambda d}\right) + \frac{4L^2}{\lambda}\log\left(\frac{2dT}{\delta}\right)\sum_{t=1}^{T}\frac{1}{n_{i_t,t}}\right)} \quad (16)\end{aligned}$$

Proof Available in appendix A.6. ■

The study of dominant factors of the bound given above enables to obtain the following proposition for the three considered settings of the context delivery process, where we removed dependencies on L, λ, R and S to simplify the notations.

Proposition 9 (Bounds for the three profile delivery settings) *For each of the three considered settings for the profile delivery process, the upper bound of the cumulative pseudo-regret is*[2]*:*

- *For the case 1, with a probability greater than* $1-3\delta$*:*

$$\hat{R}_T = \mathcal{O}\left(d\log\left(\frac{T}{\delta}\right)\sqrt{T\log(T)}\right) \tag{17}$$

- *For the case 2, with a probability greater than* $(1-3\delta)(1-\delta)$*, and for* $T \geq 2\log(1/\delta)/p^2$*:*

$$\hat{R}_T = \mathcal{O}\left(d\log\left(\frac{T}{\delta}\right)\sqrt{T\frac{\log(T)}{p}}\right) \tag{18}$$

  where p *is the probability of profile delivery for any action at each step.*

- *For the case 3, with a probability greater than* $1-3\delta$*:*

$$\hat{R}_T = \mathcal{O}\left(d\log\left(\frac{T}{\delta}\right)\sqrt{TK\log(\frac{T}{K})}\right) \tag{19}$$

Proof The proofs for these three bounds are respectively given in appendix A.7.1, A.7.2 and A.7.3. ■

Thus, in every setting our `SamplinUCB` algorithm ensures a sub-linear upper bound for its cumulative pseudo-regret. The bound given for case 2 owns an additional dependency in p, the probability of context delivery for each action at each step. Obviously, the higher this probability is, the faster the uncertainty about profiles decreases. Note that this bound for case 2 is only valid from a given number of iterations inversely proportional to p^2, since it requires a minimal number of observations to hold. The bound for case 3 owns a dependency in the number of available actions K. This comes from the fact that only the selected action reveals its profile at each step, which re-introduces the need of considering each action a minimal number of times, as it is the case with traditional stationary approaches such as the classical `UCB` algorithm. However, as we show in our experiments below, the use of the structure of the actions, which enables some common learning of reward distributions, leads to greatly better results than existing stationary algorithms in various cases.

2. $\mathcal{O}$ renders the relation "dominated by", which means that $f = \mathcal{O}(g)$ implies that there exists a strictly positive constant C such that asymptotically we have: $|f| \leq C|g|$.

3.4. Extension to the multiple-plays setting

This short section extends our algorithm for the multiple-plays setting, where multiple actions are chosen at each step. Rather than only selecting a single action i_t at any step t, the algorithm has now to select a set $\mathcal{K}_t \subseteq \mathcal{K}$ of $k \geq 1$ actions for which it gets rewards. Algorithm 1 is therefore adapted to this setting, by simply selecting the k best actions at each step (those that get the best selection scores w.r.t. formula 14) rather than only the best one (line 23 of the algorithm). The aim is still to maximize the cumulative reward through time, where all rewards at any step t are simply summed to form the collected reward at step t (although other Lipschitz functions could have been considered for the collective reward construction from the k individual ones, such as proposed in Chen et al. (2013)).

Definition 1 *The cumulative pseudo-regret of our setting of bandit with multiple plays is defined as:*

$$\hat{R}_T = T \sum_{i \in \mathcal{K}^\star} \mu_i^\top \beta - \sum_{t=1}^{\top} \sum_{i \in \mathcal{K}_t} \mu_i^\top \beta \tag{20}$$

with $\mathcal{K}^\star$ the set of k optimal actions, i.e. the k actions with the highest values $\mu_i^\top \beta$.

Proposition 10 (Generic bound for the multiple-plays setting) *By choosing $\lambda \geq \max(1, L^2/\sqrt{R^2})$, with a probability greater than $1 - 3\delta$, the cumulative pseudo-regret for our `SampLinUCB` algorithm with multiple selections is upper bounded by:*

$$\begin{aligned}\hat{R}_T \leq\ & C + 4L\left(\sqrt{\frac{d}{\lambda}\log\left(\frac{1+TkL^2/\lambda}{\delta}\right)} + 2S\right)\sqrt{2d\log\left(\frac{2dT^2}{\delta}\right)}\sum_{t=1}^{T}\sum_{i\in\mathcal{K}_t}\frac{1}{\sqrt{n_{i,t}}} \\ & + 2\left(\sqrt{d\log\left(\frac{1+TkL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S\right) \\ & \times \sqrt{Td\left(\sqrt{R^2+L^2S^2}\log\left(1+\frac{TkL^2}{\lambda d}\right) + \frac{4L^2}{\lambda}\log\left(\frac{2dT}{\delta}\right)\sum_{t=1}^{T}\sum_{i\in\mathcal{K}_t}\frac{1}{n_{i,t}}\right)}\end{aligned} \tag{21}$$

Proof The proof is available in appendix A.8. ■

Equivalent bounds for the three cases of context delivery can be directly derived from this new generic bound by applying the same methods as in the previous section. This allows us to apply our algorithm for tasks where multiple actions can be triggered at each step, such as in the data capture task considered in our experiments in section 4.2.

4. Experiments

This section is divided in two parts. First, we propose a series of experiments on artificial data in order to observe the behavior of our approach in well-controlled environments. Then, we give results obtained on real-world data, for a task of data capture from social media.

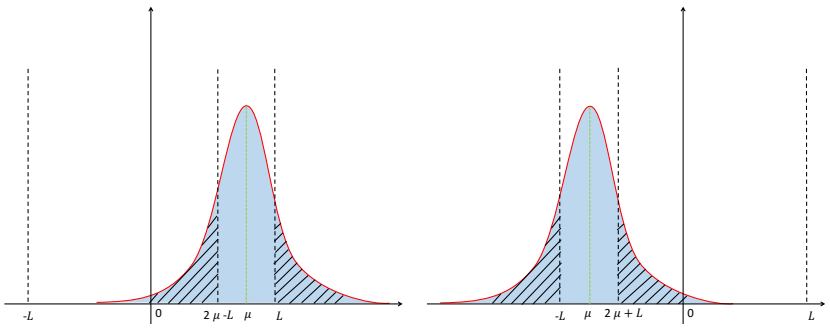
4.1. Artifical Data

4.1.1. Protocol

Data Generation: In order to assess the performances of the `SamplinUCB` algorithm, we propose to first experiment it in a context of simple selection ($k = 1$) on artificially generated data. For that purpose, we set the horizon T to 30000 iterations, the number of available actions K to 100 and the size of the profile space to $d = 5$ dimensions. Then, we sampled a mapping vector β randomly in $\left[-S/\sqrt{d}..S/\sqrt{d}\right]^d$, in order to fulfill the $||\beta|| \leq S = 1$ condition. For each arm i, we then sampled a random vector μ_i uniformly in $\left[-L/\sqrt{d}..L/\sqrt{d}\right]^d$ with $L = 1$. Finally, for each iteration $t \in \{1, ..., T\}$, we proceeded as follows to generate simulated data:

1. For each action $i \in \{1, ..., K\}$, we sampled a vector $x_{i,t}$ from the multivariate Gaussian $\mathcal{N}(\mu_i, \sigma^2 I)$. Note that, in order to assess the influence of profile samples variations on the performances of `SampLinUCB`, we tested different values for $\sigma \in \{0.5, 1.0, 2.0\}$. Moreover, in order to guarantee that $||x_{i,t}|| \leq L = 1$, while still getting sampled centered on μ_i, the Gaussian is truncated symmetrically around μ_i. This is illustrated by figure 3 for $d = 1$, where hatched areas correspond to excluded values. On the left is given the case with $\mu_i > 0$ and on the right the case with $\mu_i < 0$;

2. For each action $i \in \{1, ..., K\}$, we sampled a reward $r_{i,t}$ from a Gaussian with mean $\mu_i^\top \beta$ and variance $R^2 = 1$;



(a) Case $\mu > 0$ (b) Case $\mu < 0$

Figure 3: Profile Samples Generation Process: Truncated Gaussians

To emhasize the need of exploration on profiles, note that we set to null vectors the profile samples of the 100 first steps of each dataset. 100 datasets have been generated in such a way. The results given below correspond to averages on these artificial datasets.

Experimented Policies: We propose to compare `SampLinUCB` to the following bandit policies:

- `UCB`: the well-known `UCB` approach (Auer et al., 2002), which selects at each step the action with the best upper-confidence bound, estimated w.r.t. past rewards of actions, without any assumption about some latent structure of the actions;
- `UCBV`: the `UCBV` algorithm (Audibert et al., 2009) is an extension of `UCB`, where the variance of rewards for each action is included in the selection scores to lead the policy to ground their estimations in an higher number of observations for noisier actions.
- `UCB`$-\delta$: the `UCB`$-\delta$ algorithm (Abbasi-Yadkori et al., 2011) is a variant of `UCB` where optimism is ensured via a concentration inequality based on auto-normalised process (de la Peña et al., 2009), with a confidence level of $1 - \delta$. In our experiments, we set $\delta = 0.05$;
- `Thompson`: the `Thompson Sampling` algorithm (Thompson, 1933) introduces randomness in the exploration process by sampling reward expectations from their posterior at each time-step, following a Gaussian assumption of the rewards;
- `MOSS`: a variant of `UCB`, which usually obtains better results than the classical `UCB` but requires the knowledge about the horizon T (Audibert and Bubeck, 2009);

None of these approaches use any side information. Therefore, the only noise they have to deal with comes from the variance R^2 of the Gaussian distributions of rewards. For `SampLinUCB` an additional difficulty comes from the variations of the observed samples of profiles. The point is therefore to know whether these samples can be leveraged to exhibit some structure of the actions, that can benefit to stationary bandit tasks, despite such variations. Additionaly, the following two contextual baselines are considered in our experiments to analyze the performances of our approach:

- `LinUCB`: the very famous contextual approach that assumes a linear mapping between observed contexts and rewards (Li et al., 2010). In our case, observed profile samples correspond to the contexts that `LinUCB` takes into account in its selection policy. We consider this baseline in the interesting setting where contexts are only delivered for the selected arms (case 3 described above). In this setting, non selected arms deliver null context vectors for the next step;
- `MeanLinUCB`: this baseline corresponds to our approach but without the exploration term w.r.t. the profiles. Empirical means are considered as true profiles at each step of the process (this comes down to set $\rho_{i,t,\delta}$ to 0 for every arm and every step). As discussed above (see the last paragraph of section 3.2), such a baseline cannot guarantee a sub-linear regret since it can infinitely stay stuck on sub-optimal arms, but an empirical evaluation of its performances is useful to understand the benefits of the proposed approach.

To analyze the performances of `SampLinUCB`, we implement the three scenarios studied in previous sections. In the following, our approach is denoted `SampLinUCB_p=<p>`, where p

corresponds to the probability for any action to get a sample of its profile at every iteration. Different values for p are considered: $p \in \{0, 0.005, 0.01, 1\}$. Note that $p = 1$ corresponds to the case 1, while $p = 0$ refers to the case 3. In this latter instance, as considered in the previous sections, the samples delivery process is replaced by the ability to observe samples for the selected actions at each iteration. Note also that, for clarity and analysis purposes, in instances with $p > 0$ we do not observe samples for the selected actions (which exactly follows the cases studied in the previous section)[3]. In every instance, we set $\delta = 0.05$ for these experiments. Also, to avoid a too large exploration on profiles in the early steps, we multiplied each $\rho_{i,t,\delta}$ by a 0.01 coefficient, which still guarantees a sub-linear regret in the limit.

4.1.2. Results

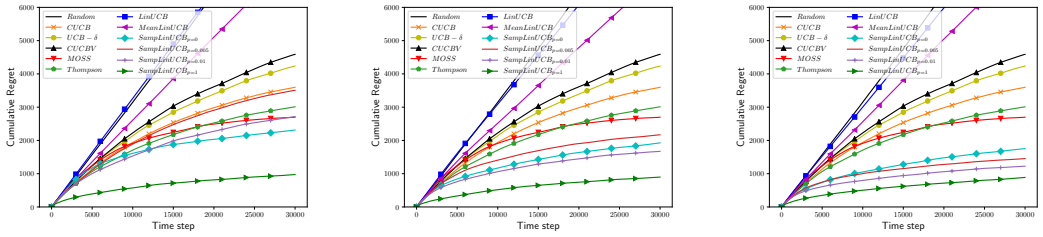
Figures 4(a), 4(b) and 4(c) report the evolution of the cumulative pseudo-regret through time for the tested policies, for σ values (variance of profile samples) respectively set to $\sigma = 2.0$, $\sigma = 1.0$ and $\sigma = 0.5$. Note that the curves of `UCB`, `UCB-`δ, `UCBV`, `Thompson` and `MOSS` are identical in every plot since their performances do not depend on the profile samples. We first notice from these plots that `UCB-`δ and `UCBV` do not provide good results on these data. It appears that these two policies over-explore during the whole bandit process. `Thompson` and `MOSS` obtain better results in average, but still far from the best contextual approach $\texttt{SampLinUCB}_{p=1}$. This confirms that using profiles of arms can be greatly advantegeous when there exist a linear correlation between these profiles and the associated rewards. In this setting (which corresponds to the case 1 studied above), the profiles are discovered step by step, but since we get a sample for every arm at each iteration, the estimators quickly converge towards the true profiles. This explains the very good results for this easy setting, and why there is nearly no differences in the results of $\texttt{SampLinUCB}_{p=1}$ for the three considered sample variances.

Let us now focus on the results provided by our `SampLinUCB` algorithm when only a subset of arms gets profile samples at each step of the process. As expected, the more the algorithm observes samples, the better it performs. However, we remark that $\texttt{SampLinUCB}_{p=0}$ obtains better results than $\texttt{SampLinUCB}_{p=0.005}$ for $\sigma = 2$ and $\sigma = 1$, and even better than $\texttt{SampLinUCB}_{p=0.01}$ when $\sigma = 2$ (while observing the same rate of samples as in this latter setting). This denotes a stronger robustness to the profile sample variance. By dynamically selecting the arms to observe, it is able to focus on the improvement of useful estimators rather than getting as many samples but for randomly selected arms (and potentially for arms that could be quickly discarded). In this interesting setting, `SampLinUCB` always outperforms non-contextual approaches for the studied sample variances, while we note a significant improvement of the results when the variance is low.

At last, we can note the very weak - near random - results obtained by `LinUCB`, which directly bases its strategy on the observed samples. More interesting are the weak results obtained by `MeanLinUCB`, which exhibits a linear regret. This emphasizes the crucial role of the profile exploration term of `SampLinUCB`: While $\texttt{SampLinUCB}_{p=0}$ is able to reconsider

3. Note that we could easily imagine tasks, which correspond to some mix of cases 2 and 3, where we both get samples from an external process and for the selected actions. For such cases, we can reasonably assume better results than those reported below for cases 2 and 3, since the process would benefit from both sample sources.

bad profiles observed in the early steps of the process, `MeanLinUCB` usually stays stuck on the first actions that provided a non-nul sample associated with a positive reward. If they are lucky, `LinUCB` and `MeanLinUCB` can exhibit good performances on some instances, but they are clearly not well fitted for the bandit setting considered in this paper.



(a) $\sigma = 2$ (b) $\sigma = 1$ (c) $\sigma = 0.5$

Figure 4: Cumulative pseudo-regret through time on artificial data with different settings for the profile samples delivery process (from very noisy samples on the left, to samples with low variance on the right).

To conclude, it appears from these first experiments that our `SampLinUCB` algorithm, while dealing with a doubled uncertainty (both on the mapping parameters and the profile estimators), is able to leverage the latent structure that it discovers through time from noisy samples of the profiles.

4.2. Real World Experiments: Social Data Capture

In this section we propose to apply our `SampLinUCB` algorithm to the task of dynamic data capture from Twitter introduced in (Gisselbrecht et al., 2015). According to a given information need, the aim is to collect relevant data from the streaming API proposed by Twitter. This API provides messages published by users on the social network in real-time. In this setting, each user account is associated with a stream of messages that can be monitored. However, for various reasons (notably w.r.t. constraints set by the social media), it is not possible to collect the whole activity of the social media. Only streams of messages published by a subset of k users can be monitored simultaneously ($k << K$). The aim is therefore to focus on users that are the most likely to publish messages that fit with the data need. The difficulty is that we do not know anything about the users beforehand, everything must be discovered during the capture process. We have thus to deal with an exploitation/exploration problem that suits well with the bandit setting studied in this paper.

Given a time period divided in T steps, the agent has to select, at each iteration $t \in \{1, ..., T\}$ of the process, a subset $\mathcal{K}_t$ of k user accounts to follow, among the whole set of possible users $\mathcal{K}$ ($\mathcal{K}_t \subseteq \mathcal{K}$). Given a relevance score $r_{i,t}$ assigned to the content posted by user $i \in \mathcal{K}_t$ during iteration t of the process (the set of tweets he posted during iteration t), the aim is to select at each iteration t the set of user accounts that maximize the sum of

collected scores:

$$\max_{(\mathcal{K}_t)_{t=1..T}} \quad \sum_{t=1}^{T} \sum_{i \in \mathcal{K}_t} r_{i,t} \tag{22}$$

4.2.1. Rewards

In our experiments, we attempt to focus on users that have a great impact on some specified thematic. The *Follow Streaming* API of Twitter provides in real-time not only tweets posted by the followed users, but also all the re-tweets and replies to these users other users post on the network. Our reward function takes all of these messages into account to provide a reward score $r_{i,t}$ for each user $i \in \mathcal{K}_t$ after each capture period t:

$$r_{i,t} = tanh\left(\sum_{\omega \in \Omega_{i,t}} g_\gamma(\omega)\right) \tag{23}$$

where $\Omega_{i,t}$ contains the original messages from i, the re-tweets of messages from i and the replies to i during the period t, and g_γ is a function returning 1 if the content of the message as argument is judged as belonging to the desired thematic γ, 0 otherwise. To build this function g_γ, we trained a SVM topic classifier on the *20 Newsgroups* dataset (with TF bag of words representations of the texts, after stemming via the Porter Stemmer algorithm). We finally focus on 4 different topics γ: *Politics*, *Religion*, *Science* and *Sport*. Four different reward functions are therefore considered in the following (one for each topic).

4.2.2. Profiles

Following the setting of our profile based bandit, we assume that each user i of the social network is associated to an unknown vector μ_i corresponding to its profile. In these experiments, we assume that the profile of a user i corresponds to the mean of its content distribution: $\mu_i = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} x_{i,t}$, where $x_{i,t}$ is a given representation of the content posted by i during step $t-1$:

$$x_{i,t+1} = f(\sum_{\omega \in \Psi_{i,t}} \omega) \tag{24}$$

where $\Psi_{i,t} \subseteq \Omega_{i,t}$ contains all messages posted by i during step t and $\omega \in \mathbb{R}^m$ (with m the size of the vocabulary) is a TF bag of words representation of a message (after stemming via the Porter Stemmer algorithm). The function f aims at reducing the dimension of the representations, since the dimension d of the profile samples is the main factor of complexity in our algorithm, due to the required $d \times d$ matrix inversions. In order to reduce the dimension of profiles, we used a *Latent Dirichlet Allocation* method specifically designed for short texts (Hong and Davison, 2010), which aims at modeling texts as a mixture of topics. We set the number of topics to $d = 30$ and learned the LDA model on a preliminary 3-days random capture from Twitter.

4.2.3. Datasets

In order to be able to test different policies and simulate a real time decision process several times, we propose to conduct our experiments on offline datasets:

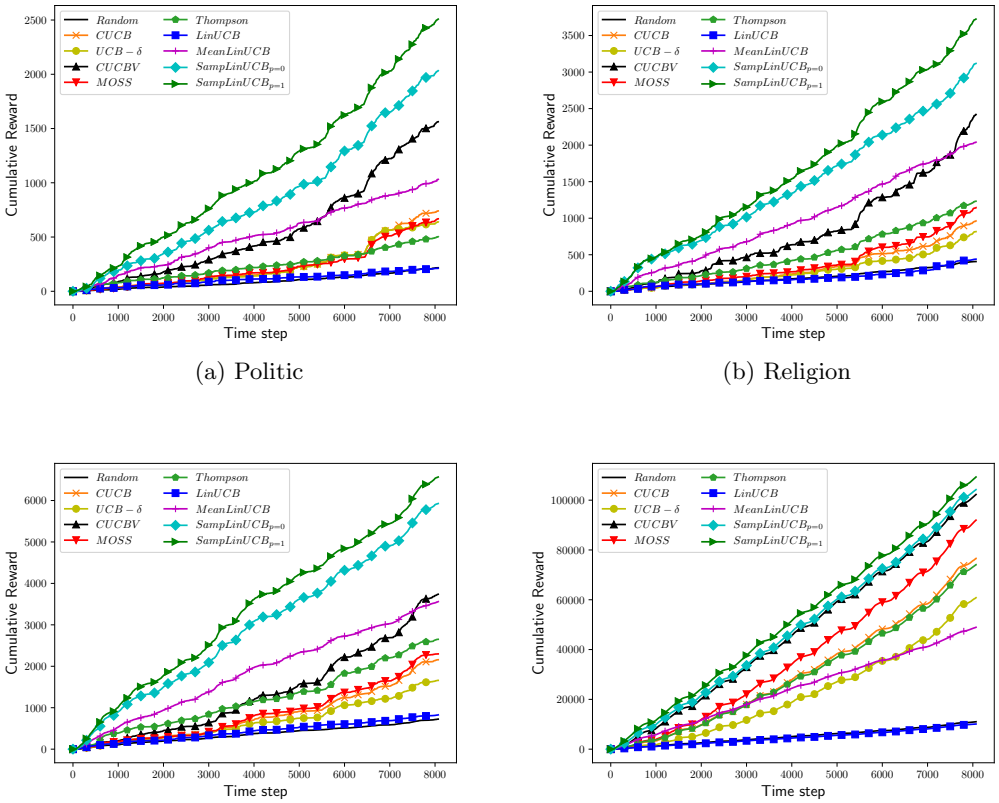
- *USElections*: dataset containing 3 587 961 messages produced by 5000 users during the ten days preceding the US presidential elections in 2012. The 5000 chosen accounts are the first ones who used either "Obama", "Romney" or "#USElections" from a preliminary random capture on Twitter.
- *OlympicGames*: dataset containing 15 010 322 messages produced by 5000 users in August 2016 during a period of three weeks covering the Olympic Games of Rio. The 5000 chosen accounts are the ones that were observed to use the most many hashtags "#Rio2016", "#Olympics", "#Olympics2016" or "#Olympicgames" within a period of preliminary random capture of three days before the Olympic Games.
- *Brexit*: dataset containing 2 118 235 messages produced by 5000 users during the first week of October 2016. The 5000 chosen accounts are the first ones who used "#Brexit" from a preliminary random capture from Twitter.

4.2.4. Results

As done in (Gisselbrecht et al., 2015), we set k, the number of listened users at each time step, to 100, and the size of an iteration to 100 seconds. In these experiments, we assume $L = S = R = 1$ and we set $\delta = 0.05$ as done with artificial data.

Figures 5, 6 and 7 give the evolution of the cumulative reward through time for the datasets *USElections*, *OlympicGames* and *Brexit* respectively. In every case, we consider the four reward functions corresponding to the four topics *Politic*, *Religion*, *Science* and *Sport*. In order to lighten the plots, we only give in these figures the results of `SamplinUCB` for $p = 0$ and $p = 1$. In every plot, our algorithm is compared to the same baselines as described in section 4.1, where the policies are extended for the multiple-plays setting (as done in (Gisselbrecht et al., 2015)).

A first important observation from these plots is that in every setting, our algorithm `SampLinUCB` obtains better results than every other policy, even `CUCBV`, the extension of `UCBV` for the multiple-plays setting. Although `CUCBV` has demonstrated good performances for the task of social data capture (Gisselbrecht et al., 2015), where a high variance can be observed in the contents posted by users, the use of profiles associated to users of the networks enables an even more efficient exploration process. Globally, same manner as with artificial data, the performances of our approach increase with p, with a maximum reached when $p = 1$. Note however that the setting $p = 0$ (the case 3 studied above) is the most realistic one, since it does not use anything but the content collected by followed users at each step, which is the case in practice when collecting data from a social media such as Twitter. Interestingly, even for this setting the results obtained are always better than those of every compared approach. The improvement w.r.t. `CUCBV` is less significant for the `Sport` reward function for which greatly more rewards exist in the datasets (greatly more messages are categorized as sport), which allows non-contextual approaches to quickly

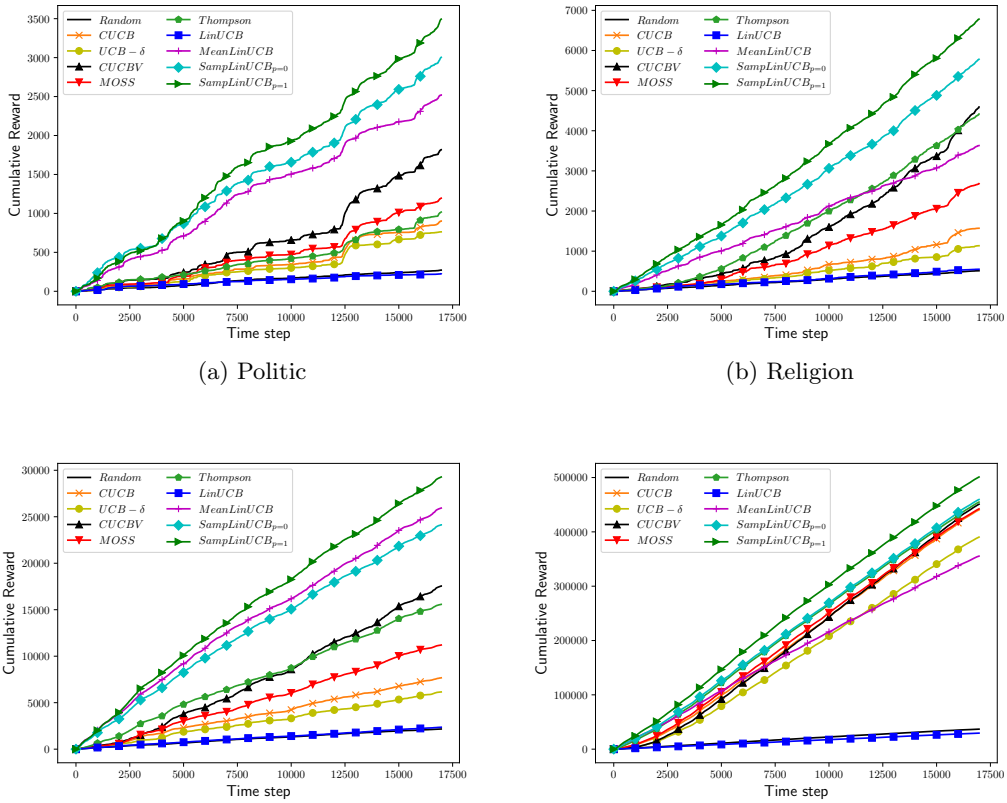


(a) Politic (b) Religion

(c) Science (d) Sport

Figure 5: Evolution of the cumulative reward through time on the *USElections* dataset, according to the three considered reward functions *Politic*, *Religion*, *Science* and *Sport*.

collect knowledge about reward expectations of users. But for every other reward function `SampLinUCB`$_{p=0}$ always obtained results comprised between 1.5 and 3 times the ones obtained with the best non-contextual approach. Note also the crucial role of the exploration term for profile discovery ρ, since `MeanLinUCB`, which considers current empirical sample means as true profiles, always obtains greatly lower results than `SampLinUCB`$_{p=0}$ (except for the `Science` reward function on the `OlympicGames` and the `Brexit` datasets, where it benefits from good rewards and profile samples observed for some useful users in the initialization steps of the process). At last, as expected, `LinUCB`, which directly bases its selection policy on profile samples observed at the current step, obtains very bad results (near random). Since obtaining nul context vectors for every user not selected at the previous step, its selection mechanism very early focuses on a given set of users without ever reconsidering the others (except in the rare cases of context samples leading to negative reward expectations according to β). All these results highlight the interest of the proposed approach, based on confidence balls of the arm profiles, for tasks where contexts are only observed when the arms are selected .

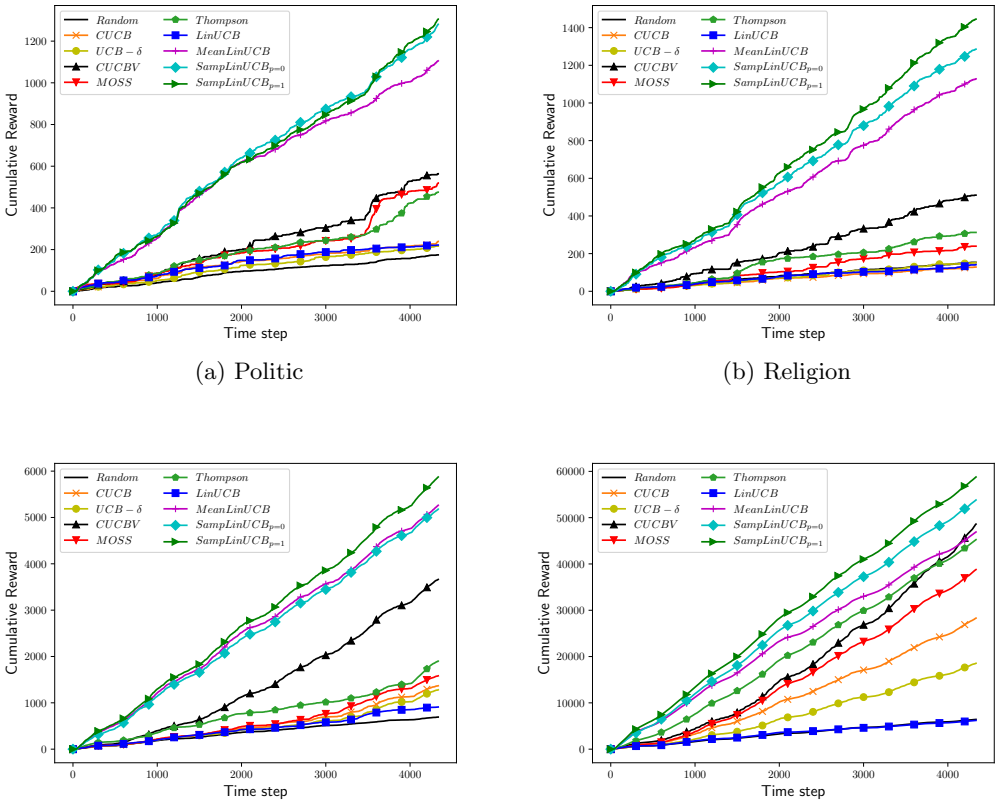


(a) Politic (b) Religion

(c) Science (d) Sport

Figure 6: Evolution of the cumulative reward through time on the *OlympicGames* dataset, according to the three considered reward functions *Politic*, *Religion*, *Science* and *Sport*.

Figures 8, 9 and 10 give the relative final cumulative rewards for different settings of the sample delivery process on the datasets *USElections*, *OlympicGames* et *Brexit* respectively (each score is normalized according to the score obtained when $p = 1$). Here we still observe that performances tend to decrease with p, for settings where $p > 0$. However it must be noticed that the setting $p = 0$ obtains results very close to other settings: it always obtains at least 80% of the final cumulative reward obtained when every user delivers a sample at each step of the process ($p = 1$). Better, in many cases $\texttt{SampLinUCB}_{p=0}$ succeeds in obtaining an higher final cumulative reward than $p = 0.01$ and $p = 0.02$. This is particularly true for the `Brexit` dataset where the dynamic selection of samples to be delivered appears very effective. On that dataset, $\texttt{SampLinUCB}_{p=0}$ even usually reaches the performances of $\texttt{SampLinUCB}_{p=0.05}$, while observing greatly less profile samples at each step (only 100 over 5000 at each iteration, which corresponds to the observation rate of the setting $p = 0.02$). While settings with $p > 0$ are greatly favored by the fact that they do not need to play an arm to get a sample of its profile, $\texttt{SampLinUCB}_{p=0}$ is not only active for the discovery of the mapping parameters, but also for the estimation of profiles. Its knowledge about profiles



(a) Politic (b) Religion

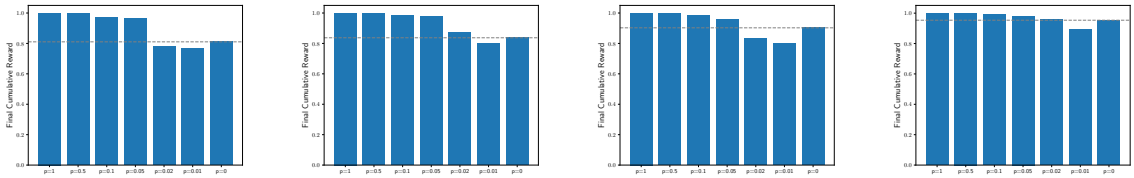
(c) Science (d) Sport

Figure 7: Evolution of the cumulative reward through time on the *Brexit* dataset, according to the three considered reward functions *Politic*, *Religion*, *Science* and *Sport*.

is directly connected to its selection strategy, with a selection score that favors promising actions with high uncertainty about their profile. This leads to an algorithm that efficiently deals with a trade-off between exploitation of good actions and exploration on both the mapping parameter and the profiles of actions.

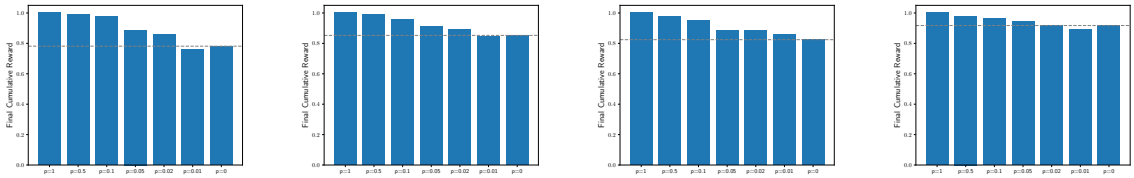
5. Conclusion

In this paper, we focused on structured stochastic bandits, where rewards depend on some constant profile associated with actions. More specifically, we introduced the case where the associated profiles are unknown beforehand, and must be discovered from samples delivered during the process. This setting implies a doubled uncertainty, both on profile estimators and on reward predictors, for which we designed a dedicated algorithm, named `SamplinUCB`, that seeks at leveraging the structure of the unknown profiles in its exploration process. Various settings for the profile samples delivery process have been considered, for which we gave theoretical convergence guarantees. Finally, experiments on both artificial data



(a) Politic (b) Religion (c) Science (d) Sport

Figure 8: Final normalized cumulative rewards for `SampLinUCB` on *USElections* with different p settings.



(a) Politic (b) Religion (c) Science (d) Sport

Figure 9: Final normalized cumulative rewards for `SampLinUCB` on *OlympicGames* with different p settings.

and a task of data capture from social networks demonstrate the very good behavior of the proposed approach. Ongoing works concern the inclusion of a non-stationary part in the selection strategy, where profiles may vary over time according to some evolving latent state of the actions.

Acknowledgments

This research work has been carried out in the framework of the Technological Research Institute SystemX, and therefore granted with public funds within the scope of the French Program "Investissements d'Avenir".

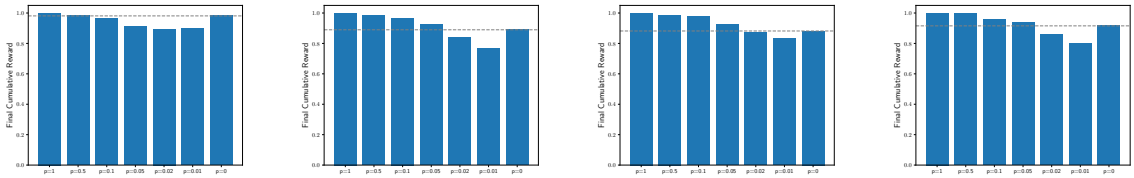
Appendix A. Appendix

A.1. Proof of proposition 1

The two following lemmas directly come from the definition of the sub-gaussian variables.

Lemma 1 *Let X be a random variable centered on 0. Then, X is said sub-gaussian with constant R if one of the two equivalent following conditions holds:*

- *Laplace Condition:* $\exists R > 0, \forall \lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda X}] \leq e^{R^2\lambda^2/2}$



(a) Politic (b) Religion (c) Science (d) Sport

Figure 10: Final normalized cumulative rewards for `SampLinUCB` on *Brexit* with different p settings.

- *Sub-Gaussian Tail:* $\exists R > 0, \forall \gamma > 0, P(|X| \geq \gamma) \leq 2e^{-\gamma^2/(2R^2)}$

Lemma 2 *Let X_1 and X_2 be two sub-gaussian variables with respective constant R_1 and R_2. Let α_1 and α_2 be two real scalars. Then the variable $\alpha_1 X_1 + \alpha_2 X_2$ is sub-gaussian too, with constant $\sqrt{\alpha_1^2 R_1^2 + \alpha_2^2 R_2^2}$.*

The lemma 3 can be deduced from the application of the lemma 1 in the context of our specific problem of profile-bandit with unknown profiles, where the deviation of the profile estimators is a sub-gaussian variable.

Lemma 3 *Let us assume that for any i all samples $x_{i,t} \in \mathbb{R}^d$ observed for i at every step $t > 0$ are iid with mean $\mu_i \in \mathbb{R}^d$. Let us also assume that $||x_{i,t}|| \leq L$ for every i and t, and that $||\beta|| \leq S$. Then, for every i, and at each step t, $\beta^\top \epsilon_{i,t}$ is sub-gaussian with constant $\frac{LS}{\sqrt{n_{i,t}}}$ (with $\epsilon_{i,t} = \mu_i - \hat{x}_{i,t}$).*

Proof

By using the Cauchy-Schwarz inequality, for all i and at each step t we have: $|x_{i,t}^\top \beta| \leq ||\beta|| ||\hat{x}_{i,t}|| \leq LS$. Then, given that for all i, all samples $x_{i,t}$ are iid and $\mathbb{E}[x_{i,t}] = \mu_i$, we can apply the Hoeffding inequality to the random variable $\beta^\top \hat{x}_{i,t}$ with mean $\mu_i^\top \beta$:

$$\forall \gamma > 0, \mathbb{P}\left(|\beta^\top \hat{x}_{i,t} - \beta^\top \mu_i| > \gamma\right) = \mathbb{P}\left(|\beta^\top \epsilon_{i,t}| > \gamma\right) \leq 2e^{-\frac{n_{i,t}\gamma^2}{2S^2L^2}}$$

which allows us to say that $\epsilon_{i,t}^\top \beta$ is sub-gaussian with constant $\frac{LS}{\sqrt{n_{i,t}}}$. ■

We finally use the lemma 2 with the sum of $\beta^\top \epsilon_{i,t}$ and $\eta_{i,s}$ to prove the proposition 1, which establishes the random variable $\beta^\top \epsilon_{i,t} + \eta_{i,s}$ is conditionally sub-gaussian with constant $R_{i,t} = \sqrt{R^2 + \frac{L^2 S^2}{n_{i,t}}}$.

A.2. Proof of the proposition 3

To lighten notations, we removed the dependence on t in A and X. We have:

$$\begin{aligned}
\hat{\beta}_{t-1} &= \arg\min_{\beta} \sum_{s=1}^{t-1} \frac{1}{R_{i_s,t}} (\beta^\top \hat{x}_{i_s,t} - r_{i_s,s})^2 + \lambda ||\beta||^2 \\
&= (X^\top A X + \lambda I)^{-1} X^\top A Y \\
&= (X^\top A X + \lambda I)^{-1} X^\top A (X\beta + \eta^{'}) \\
&= (X^\top A X + \lambda I)^{-1} X^\top A \eta^{'} + (X^\top A X + \lambda I)^{-1} (X^\top A X + \lambda I)\beta \\
&\quad - (X^\top A X + \lambda I)^{-1} \lambda I \beta \\
&= (X^\top A X + \lambda I)^{-1} X^\top A \eta^{'} + \beta - \lambda (X^\top A X + \lambda I)^{-1} \beta
\end{aligned}$$

Then, the following main arguments of this proof come from the theory of auto-normalized process (de la Peña et al., 2009). By using a similar method to the one used in (Abbasi-Yadkori et al., 2011), we get:

$$||\hat{\beta}_{t-1} - \beta||_{V_{t-1}} \leq ||X^\top A \eta^{'}||_{V_{t-1}^{-1}} + \lambda ||\beta||_{V_{t-1}^{-1}}$$

with $V_{t-1} = \lambda I + X^\top A X$, which is semi-definite positive since $\lambda > 0$. Since $||\beta|| \leq S$ and $||\beta||^2_{V_{t-1}^{-1}} \leq ||\beta||^2 / \lambda_{min}(V_{t-1}) \leq ||\beta||^2 / \lambda$, we get:

$$||\hat{\beta}_{t-1} - \beta||_{V_{t-1}} \leq ||X^\top A \eta^{'}||_{V_{t-1}^{-1}} + \sqrt{\lambda} S$$

By using the proposition 1 of (Abbasi-Yadkori et al., 2011), and since we know from proposition 1 that $\frac{\eta_s^{'}}{R_{i_s,t}}$ is sub-gaussian with constant 1, for any $\delta > 0$, with a probability of at least $1 - \delta$, for every $t \geq 0$ we have:

$$\begin{aligned}
||X^\top A \eta^{'}||_{V_{t-1}^{-1}} &= ||\sum_{s=1}^{t-1} \frac{\eta_s^{'}}{R_{i_s,t}} \hat{x}_{i_s,t}||_{V_{t-1}^{-1}} \\
&\leq \sqrt{2 \log \left(\frac{det(V_{t-1})^{1/2} det(\lambda I)^{-1/2}}{\delta} \right)} \\
&\leq \sqrt{d \log \left(\frac{1 + tL^2/\lambda}{\delta} \right)}
\end{aligned}$$

A.3. Proof of the proposition 4

Proof Let us assume that the inequality of the proposition 3 is valid. Therefore, we have for all $t > 0$ and every $i \in \mathcal{K}$:

$$\begin{aligned}\hat{\beta}_{t-1}^{\top}\mu_i + \alpha_{t-1}||\mu_i||_{V_{t-1}^{-1}} - \beta^{\top}\mu_i =& (\hat{\beta}_{t-1} - \beta)^{\top}\mu_i + \alpha_{t-1}||\mu_i||_{V_{t-1}^{-1}} \\ \geq& - ||\hat{\beta}_{t-1} - \beta||_{V_{t-1}}||\mu_i||_{V_{t-1}^{-1}} + \alpha_{t-1}||\mu_i||_{V_{t-1}^{-1}} \\ \geq& - \alpha_{t-1}||\mu_i||_{V_{t-1}^{-1}} + \alpha_{t-1}||\mu_i||_{V_{t-1}^{-1}} \\ =& 0\end{aligned}$$

■

A.4. Proof of the proposition 5

With $||x_{i,t}|| \leq L$, we know that, for any $j \in [1..d]$, $|x_{i,t}^j| \leq L$. Thus, we can apply the Hoeffding inequality to each dimension of the profile estimators:

$$\forall \gamma > 0 : \ \mathbb{P}\left(|\hat{x}_{i,t}^j - \mu_i^j| > \gamma/\sqrt{d}\right) \leq 2e^{-\frac{n_{i,t}\gamma^2}{2L^2 d}}$$

Then, by using the fact that $||\hat{x}_{i,t} - \mu_i|| \leq \frac{1}{\sqrt{d}} \sum_{i=1}^{d} |\hat{x}_{i,t}^i - \mu_i^i|$ and the uniform bound property, we get:

$$\mathbb{P}\left(||\hat{x}_{i,t} - \mu_i|| \leq \gamma\right) \geq 1 - 2de^{-\frac{n_{i,t}\gamma^2}{2L^2 d}}$$

Thus, for every $i \in \{1, \ldots, K\}$ and every step $t > 0$, with a probability of at least $1 - \delta/t^2$:

$$||\hat{x}_{i,t} - \mu_i|| \leq L\sqrt{\frac{2d}{n_{i,t}} \log\left(\frac{2dt^2}{\delta}\right)}$$

This bound for the deviation of the profile estimator can be less restrictive than the base assumption which states that for any $i \in \mathcal{K}$ and $t \geq 0$, $||x_{i,t}|| \leq L$. From this assumption we indeed know that, $||\hat{x}_{i,t}|| \leq L$, $||\mu_i|| \leq L$ and thus $||\hat{x}_{i,t} - \mu_i|| \leq 2L$.

We therefore consider the following bound that holds for any $t \geq 0$ with a probability greater than $1 - \delta/t^2$:

$$||\hat{x}_{i,t} - \mu_i|| \leq \min(L\sqrt{\frac{2d}{n_{i,t}} \log\left(\frac{2dt^2}{\delta}\right)}, 2L) = \rho_{i,t,\delta}$$

A.5. Proof of the proposition 6

Proof Let us assume that the inequality of the proposition 5 is valid. Therefore, we have:

- $||\mu_i||_{V_{t-1}^{-1}} - ||\hat{x}_{i,t}||_{V_{t-1}^{-1}} \leq ||\mu_i - \hat{x}_{i,t}||_{V_{t-1}^{-1}} \leq ||\mu_i - \hat{x}_{i,t}||/\sqrt{\lambda} \leq \rho_{i,t,\delta}/\sqrt{\lambda}$. Thus: $||\mu_i||_{V_{t-1}^{-1}} \leq ||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + \rho_{i,t,\delta}/\sqrt{\lambda} = ||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}}$, with $\tilde{\epsilon}_{i,t} = \rho_{i,t,\delta}\hat{x}_{i,t}/(\sqrt{\lambda}||\hat{x}_{i,t}||_{V_{t-1}^{-1}})$.

- $|\hat{\beta}_{t-1}^{\top}(\hat{x}_{i,t} - \mu_i)| \leq ||\hat{\beta}_{t-1}||||(\hat{x}_{i,t} - \mu_i)|| \leq \hat{\beta}_{t-1}^{\top}\bar{\epsilon}_{i,t}$, with $\bar{\epsilon}_{i,t} = \rho_{i,t,\delta}\hat{\beta}_{t-1}/||\hat{\beta}_{t-1}||$.

By using these two results and the uniform bound property, we can proof the proposition:

$$
\begin{aligned}
&\hat{\beta}_{t-1}^{\top}(\hat{x}_{i,t} + \bar{\epsilon}_{i,t}) + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} - \beta^{\top}\mu_i \\
=&(\hat{\beta}_{t-1} - \beta)^{\top}\mu_i + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} - \hat{\beta}_{t-1}^{\top}(\mu_i - \hat{x}_{i,t}) + \hat{\beta}_{t-1}^{\top}\bar{\epsilon}_{i,t} \\
\geq& - ||\hat{\beta}_{t-1} - \beta||_{V_{t-1}}||\mu_i||_{V_{t-1}^{-1}} + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} + \hat{\beta}_{t-1}^{\top}(\hat{x}_{i,t} - \mu_i + \bar{\epsilon}_{i,t}) \\
\geq& - \alpha_{t-1}||\mu_i||_{V_{t-1}^{-1}} + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} + \hat{\beta}_{t-1}^{\top}(\hat{x}_{i,t} - \mu_i + \bar{\epsilon}_{i,t}) \\
\geq& - \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} + \hat{\beta}_{t-1}^{\top}(\hat{x}_{i,t} - \mu_i + \bar{\epsilon}_{i,t}) \\
\geq& 0
\end{aligned}
$$

■

A.6. Proof of the proposition 8

Lemma 4 *For every $i \in \mathcal{K}$ and $t > 0$, with a probability of at least $1 - \delta/t^2 - \delta$, we have:*

$$
\begin{aligned}
&\hat{\beta}_{t-1}^{\top}(\hat{x}_{i,t} + \bar{\epsilon}_{i,t}) + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} - \beta^{\top}\mu_i \\
&\qquad\qquad \leq 2\alpha_t||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + 4\sqrt{d}(\alpha_{t-1}/\sqrt{\lambda} + S)\rho_{i,t,\delta}
\end{aligned}
$$

Proof

As for proposition 6, we assume that the inequality of proposition 5 holds. Then, noting that $||\bar{\epsilon}_{i,t}||_{V_{t-1}^{-1}} \leq ||\bar{\epsilon}_{i,t}||/\sqrt{\lambda} = \rho_{i,t,\delta}/\sqrt{\lambda}$ and $||\tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} = \rho_{i,t,\delta}/\sqrt{\lambda}$, we have:

$$
\begin{aligned}
&\hat{\beta}_{t-1}^{\top}(\hat{x}_{i,t} + \bar{\epsilon}_{i,t}) + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} - \beta^{\top}\mu_i \\
=&(\hat{\beta}_{t-1} - \beta)^{\top}\mu_i + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} - \hat{\beta}_{t-1}^{\top}(\mu_i - \hat{x}_{i,t}) + \hat{\beta}_{t-1}^{\top}\bar{\epsilon}_{i,t} \\
\leq&||\hat{\beta}_{t-1} - \beta||_{V_{t-1}}||\mu_i||_{V_{t-1}^{-1}} + \alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} + \hat{\beta}_{t-1}^{\top}(\hat{x}_{i,t} - \mu_i + \bar{\epsilon}_{i,t}) \\
\leq&2\alpha_{t-1}||\hat{x}_{i,t} + \tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} + 2||\hat{\beta}_{t-1}||_{V_{t-1}}||\bar{\epsilon}_{i,t}||_{V_{t-1}^{-1}} \\
\leq&2\alpha_{t-1}||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + 2\alpha_{t-1}||\tilde{\epsilon}_{i,t}||_{V_{t-1}^{-1}} + 2(\alpha_{t-1} + S\sqrt{\lambda})||\bar{\epsilon}_{i,t}||_{V_{t-1}^{-1}} \\
\leq&2\alpha_{t-1}||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + 4(\alpha_{t-1}/\sqrt{\lambda} + S)\rho_{i,t,\delta}
\end{aligned}
$$

■

Lemma 5 *For every t, with a probability of at least $1-\delta/t^2-\delta$, the instantaneous pseudo-regret of the algorithm* ***SampLinUCB****, noted $reg_t = \beta^\top \mu_{i^\star} - \beta^\top \mu_{i_t}$, is upper-bounded as:*

$$reg_t \leq \underbrace{2\alpha_{t-1}||\hat{x}_{i_t,t}||_{V_{t-1}^{-1}}}_{reg_t^{(1)}} + \underbrace{4(\alpha_{t-1}/\sqrt{\lambda}+S)\rho_{i_t,t,\delta}}_{reg_t^{(2)}}$$

Proof The previous lemma allows us to say that for all t:

$s_{i_t,t} \leq \beta^\top \mu_{i_t} + 2\alpha_{t-1}||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + 4(\alpha_{t-1}/\sqrt{\lambda}+S)\rho_{i,t,\delta}$

Now, given the selection policy of `SampLinUCB` and the proposition 6, we get for all t: $s_{i_t,t} \geq s_{i^\star,t} \geq \beta^\top \mu_{i^\star}$. Thus:

$reg_t \leq s_{i_t,t} - \beta^\top \mu_{i_t} \leq 2\alpha_{t-1}||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + 4(\alpha_{t-1}/\sqrt{\lambda}+S)\rho_{i,t,\delta}$

■

Next, we use the uniform bound property, the fact that $\sum_{t=2}^{\infty} \frac{\delta}{t^2} = \delta(\pi^2/6-1) \leq \delta$ and the fact that in the proposition 3 the bound is uniform (i.e., it holds for all step t simultaneously) to say that, with a probability of at least $1-2\delta$:

$$\begin{aligned}
\sum_{t=1}^{T} reg_t^{(2)} &\leq C + \sum_{t=2}^{T} 4(\alpha_{t-1}/\sqrt{\lambda}+S)\rho_{i_t,t,\delta} \\
&\leq C + \sum_{t=2}^{T} 4(\alpha_{t-1}/\sqrt{\lambda}+S)L\sqrt{\frac{2d}{n_{i,t}}\log\left(\frac{2dt^2}{\delta}\right)} \\
&\leq C + 4L(\alpha_T/\sqrt{\lambda}+S)\sqrt{2d\log\left(\frac{2dT^2}{\delta}\right)}\sum_{t=2}^{T}\frac{1}{\sqrt{n_{i_t,t}}}
\end{aligned}$$

On another hand, we have:

$$\begin{aligned}
\sum_{t=1}^{T} reg_t^{(1)} &\leq \sum_{t=1}^{T} 2\alpha_{t-1}||\hat{x}_{i_t,t}||^2_{V_{t-1}^{-1}} \\
&\leq \sqrt{T\sum_{t=1}^{T} 4\alpha_{t-1}^2||\hat{x}_{i_t,t}||^2_{V_{t-1}^{-1}}} \\
&\leq 2\alpha_T\sqrt{T\sum_{t=1}^{T}||\hat{x}_{i_t,t}||^2_{V_{t-1}^{-1}}}
\end{aligned}$$

Now, it remains to upper-bound the term $\sum_{t=1}^{T}||\hat{x}_{i_t,t}||^2_{V_{t-1}^{-1}}$.

For that purpose, we introduce the following notation: $\nu_{i,t,\delta} = L\sqrt{2d/(n_{i,t})\log(2dT/\delta)}$. By using again the Hoeffding inequality, with a probability of at least $1-\delta/T$, we get for all $s \leq t-1$:

$$||\hat{x}_{i_s,t}|| \leq ||\mu_{i_s}|| + \nu_{i_s,t,\delta}$$

With $\check{\epsilon}_{i,t} = min(\nu_{i,t,\delta}, ||\mu_i||)\mu_i/||\mu_i||$, we get, for all $s \leq t-1$:

$$1/\sqrt{R_{i_s,s}}||\mu_{i_s} - \check{\epsilon}_{i_s,s}|| \leq 1/\sqrt{R_{i_s,t}}||\hat{x}_{i_s,t}||$$

Then, we arrive to:

$$V_{t-1} = \lambda I + \sum_{s=1}^{t-1} \frac{1}{R_{i_s,t}} \hat{x}_{i_s,t}\hat{x}_{i_s,t}^\top \geq \lambda I + \sum_{s=1}^{t-1} \frac{1}{R_{i_s,s}} (\mu_{i_s} - \check{\epsilon}_{i_s,s})(\mu_{i_s} - \check{\epsilon}_{i_s,s})^\top = W_{t-1}$$

Which means that for every vector x: $||x||_{V_{t-1}^{-1}} \leq ||x||_{W_{t-1}^{-1}}$.
Let us now define $\hat{\epsilon}_{i,t} = \nu_{i,t,\delta}\mu_i/(\sqrt{\lambda}||\mu_i||_{W_{t-1}^{-1}})$, such that for all $s \leq t-1$:

$$||\hat{x}_{i_s,t}||_{W_{t-1}^{-1}} \leq ||\mu_{i_s} + \hat{\epsilon}_{i_s,s}||_{W_{t-1}^{-1}}$$

and

$$||\hat{\epsilon}_{i_s,s}||_{W_{t-1}^{-1}} = \nu_{i_s,s,\delta}/\sqrt{\lambda}$$

Finally, by using the uniform bound property and the fact that $\sum_{t=1}^{T} \frac{\delta}{T} = \delta$, with a probability of at least $1-\delta$:

$$\begin{aligned}
\sum_{t=1}^{T} ||\hat{x}_{i_t,t}||^2_{V_{t-1}^{-1}} &\leq \sum_{t=1}^{T} ||\hat{x}_{i_t,t}||^2_{W_{t-1}^{-1}} \\
&\leq \sum_{t=1}^{T} ||\mu_{i_t} + \hat{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} \leq \sum_{t=1}^{T} ||\mu_{i_t} + \hat{\epsilon}_{i_t,t} - \check{\epsilon}_{i_t,t} + \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} \\
&\leq \sum_{t=1}^{T} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} + \sum_{t=1}^{T} ||\hat{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} + \sum_{t=1}^{T} ||\check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} \\
&\leq \sum_{t=1}^{T} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} + \frac{2}{\lambda} \sum_{t=1}^{T} \nu^2_{i_t,t,\delta} \\
&\leq \sum_{t=1}^{T} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} + \frac{4L^2 d}{\lambda} \log\left(\frac{2dT}{\delta}\right) \sum_{t=1}^{T} \frac{1}{n_{i_t,t}}
\end{aligned}$$

On another hand, we have:

$$\begin{aligned}
det(W_T) &= det(W_{T-1} + \frac{1}{R_{i_T,T}}(\mu_{i_T} - \check{\epsilon}_{i_T,T})(\mu_{a_T} - \check{\epsilon}_{i_T,T})^\top) \\
&= det(W_{T-1}) det(I + \frac{1}{R_{i_T,T}} W_{T-1}^{-1/2}(\mu_{i_T} - \check{\epsilon}_{i_T,T})(W_{T-1}^{-1/2}(\mu_{i_T} - \check{\epsilon}_{i_T,T}))^\top) \\
&= det(W_{T-1})(1 + \frac{1}{R_{i_T,T}} ||\mu_{i_T} - \check{\epsilon}_{i_T,T}||^2_{W_{T-1}^{-1}}) \\
&= det(\lambda I) \prod_{t=1}^{T} (1 + \frac{1}{R_{i_t,t}} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}})
\end{aligned}$$

Where we used the fact that all eigenvalues of $I + xx^\top$ equal 1 except one that is associated to the eigenvector x and thus equals $1 + ||x||^2$.

Since by assumption $\lambda > \max(1, L^2/\sqrt{R^2})$, we have:

$$\frac{1}{R_{i_t,t}} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} \leq ||\hat{x}_{i_t,t}||^2/\sqrt{R^2}\lambda \leq L^2/\sqrt{R^2}\lambda \leq 1$$

Thus, by using the fact that $x \leq 2\log(1+x)$ when $0 \leq x \leq 1$, we get:

$$\begin{aligned}
2\log\left(\frac{det(W_T)}{det(\lambda I)}\right) &\geq \sum_{t=1}^{T} \frac{1}{R_{i_t,t}} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} \\
&\geq \min_{t=1..T}\left(\frac{1}{R_{i_t,t}}\right) \sum_{t=1}^{T} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} \\
&\geq 1/\sqrt{R^2 + L^2S^2} \sum_{t=1}^{T} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}}
\end{aligned}$$

As in the lemma 11 of (Abbasi-Yadkori et al., 2011), we also have:

$$\log\left(\frac{det(W_T)}{det(\lambda I)}\right) \leq d\log\left(1 + \frac{TL^2}{\lambda d}\right)$$

Which leads us to:

$$\sum_{t=1}^{T} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} \leq \sqrt{R^2 + L^2S^2} d\log\left(1 + \frac{TL^2}{\lambda d}\right)$$

Finally, same manner as in the lemma 10 of Abbasi-Yadkori et al. (2011), the trace-determinant inequality gives:

$$\alpha_T \leq \sqrt{d\log\left(\frac{1 + TL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S$$

Gathering all these results together allows us to prove the announced result.

A.7. Proof of the proposition 9

Lemma 6 *When removing dependencies on L, λ, R and S, the bound from proposition 16 for the cumulative regret $\hat{R}_T$ can written as follows (when $T > d$):*

$$\hat{R}_T \leq C + C_1 d \log\left(\frac{T}{\delta}\right) \sum_{t=1}^{T} \frac{1}{\sqrt{n_{i_t,t}}} + C_2 d \log\left(\frac{T}{\delta}\right) \sqrt{T \sum_{t=1}^{T} \frac{1}{n_{i_t,t}}}$$

with C,C_1 and C_2 three constants.

Proof From proposition 16, we have:

$$\hat{R}_T \leq C + \hat{R}_{T,1} + \hat{R}_{T,2}$$

where:

$$\begin{aligned}
\hat{R}_{T,1} &= 4L \left(\sqrt{\frac{d}{\lambda} \log\left(\frac{1 + TL^2/\lambda}{\delta}\right)} + 2S \right) \sqrt{2d \log\left(\frac{2dT^2}{\delta}\right)} \sum_{t=1}^{T} \frac{1}{\sqrt{n_{i_t,t}}} \\
&\leq Constant \times d \sqrt{\log\left(\frac{T}{\delta}\right) \log\left(\frac{dT^2}{\delta}\right)} \sum_{t=1}^{T} \frac{1}{\sqrt{n_{i_t,t}}} \\
&\leq Constant \times d \log\left(\frac{T}{\delta}\right) \sum_{t=1}^{T} \frac{1}{\sqrt{n_{i_t,t}}} \qquad \text{(when } T > d > 0\text{)}
\end{aligned}$$

And

$$\begin{aligned}
\hat{R}_{T,2} &= 2 \left(\sqrt{d \log\left(\frac{1 + TL^2/\lambda}{\delta}\right)} + \sqrt{\lambda} S \right) \\
&\quad \times \sqrt{Td \left(\sqrt{R^2 + L^2 S^2} \log\left(1 + \frac{TL^2}{\lambda d}\right) + \frac{4L^2}{\lambda} \log\left(\frac{2dT}{\delta}\right) \sum_{t=1}^{T} \frac{1}{n_{i_t,t}} \right)} \\
&\leq Constant \times \sqrt{d \log\left(\frac{T}{\delta}\right)} \sqrt{Td \left(\log(T) + \log\left(\frac{dT}{\delta}\right) \sum_{t=1}^{T} \frac{1}{n_{i_t,t}} \right)} \\
&\leq Constant \times \sqrt{d \log\left(\frac{T}{\delta}\right)} \sqrt{Td \log\left(\frac{dT}{\delta}\right) \sum_{t=1}^{T} \frac{1}{n_{i_t,t}}} \\
&\leq Constant \times d \log\left(\frac{T}{\delta}\right) \sqrt{T \sum_{t=1}^{T} \frac{1}{n_{i_t,t}}} \qquad \text{(when } T > d > 0\text{)}
\end{aligned}$$

■

Thanks to this lemma, we are ready to derive specific bounds for the three considered profile delivery settings.

A.7.1. Case 1:

On one hand, we have:

$$\sum_{t=1}^{T} \frac{1}{n_{i_t,t}} = \sum_{t=1}^{T} \frac{1}{t} \leq 1 + \log(T)$$

On the other hand, we have:

$$\sum_{t=1}^{T} \frac{1}{\sqrt{n_{i_t,t}}} = \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq \int_0^T \frac{1}{\sqrt{t}} \, \mathrm{d}t \leq 2\sqrt{T}$$

Finally, we get from lemma 6 that, when $T > d > 0$:

$$\hat{R}_{T,1} \leq C + C_1 d \log\left(\frac{T}{\delta}\right)\sqrt{T} + C_2 d \log\left(\frac{T}{\delta}\right)\sqrt{T\log(T)}$$

with C,C_1 and C_2 three constants. Since the last term is clearly the greatest, we get the announced result.

A.7.2. Case 2:

Lemma 7 *$\forall i, \forall t \geq \lceil 2\log(1/\delta)/p^2 \rceil$, with a probability of at least $1-\delta$:*

$$n_{i,t} \geq \frac{tp}{2}$$

Proof By the Hoeffding inequality, for all $\epsilon > 0$:

$$\mathbb{P}(n_{i,t} \geq tp - \epsilon) \geq 1 - e^{-2\epsilon^2/t}$$

By taking $\epsilon = tp/2$, we get:

$$\mathbb{P}(n_{i,t} \geq tp/2) \geq 1 - e^{-tp^2/2}$$

If $t \geq 2\log(1/\delta)/p^2$, then $1 - e^{-tp^2/2} \geq 1 - \delta$, which proves the lemma. ■

Let us note $u = ceil(2\log(1/\delta)/p^2)$. Thus, following lemma 7, with a probability of at least $1-\delta$, we have:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{1}{n_{i_t,t}} &= \sum_{t=1}^{u} \frac{1}{n_{i_t,t}} + \sum_{t=u+1}^{T} \frac{1}{n_{i_t,t}} \\
&\leq u + \frac{2}{p} \sum_{t=u+1}^{T} \frac{1}{t} \\
&\leq u + \frac{2}{p} \int_u^T \frac{1}{t} \, \mathrm{d}t \\
&\leq u + \frac{2\log(T)}{p}
\end{aligned}$$

From another hand, still thanks to the lemma 7, with a probability of at least $1-\delta$:

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{\sqrt{n_{i_t,t}}} &= \sum_{t=1}^{u}\frac{1}{\sqrt{n_{i_t,t}}} + \sum_{t=u+1}^{T}\frac{1}{\sqrt{n_{i_t,t}}} \\ &\leq u + \sqrt{\frac{2}{p}}\sum_{t=u+1}^{T}\frac{1}{\sqrt{t}} \\ &\leq u + \sqrt{\frac{2}{p}}\int_{u}^{T}\frac{1}{\sqrt{t}}\,\mathrm{d}t \\ &\leq u + 2\sqrt{\frac{2T}{p}}\end{aligned}$$

Finally, we get from lemma 6 that, when $T > d > 0$:

$$\hat{R}_{T,1} \leq C + C_1 d\log\left(\frac{T}{\delta}\right)\sqrt{\frac{T}{p}} + C_2 d\log\left(\frac{T}{\delta}\right)\sqrt{T\frac{\log(T)}{p}}$$

with C,C_1 and C_2 three constants. Since the last term is clearly the greatest, we get the announced result.

A.7.3. Case 3:

First note that the sum $\sum_{t=1}^{T}\frac{1}{n_{i_t,t}}$ is maximized when each action has delivered exactly $\lfloor T/K \rfloor$ samples in the $\lfloor T/K \rfloor K$ first iterations (i.e., every action has been played as many times). Thus:

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{n_{i_t,t}} &\leq \sum_{i=1}^{K}\sum_{t=1}^{\lfloor T/K \rfloor+1}\frac{1}{t} \\ &\leq K\sum_{t=1}^{\lceil T/K \rceil}\frac{1}{t} \\ &\leq K(1+\log(\lceil T/K \rceil))\end{aligned}$$

With the same argument, we also get:

$$\begin{aligned}\sum_{t=1}^{T}\frac{1}{\sqrt{n_{i_t,t}}} &\leq K\sum_{t=1}^{\lceil T/K \rceil}\frac{1}{\sqrt{t}} \\ &\leq 2K\sqrt{\lceil T/K \rceil}\end{aligned}$$

Finally, by noting that $K\log(\lceil T/K \rceil) \sim K\log(T/K)$, that $K\sqrt{\lceil T/K \rceil} \sim \sqrt{KT}$ and by using the generic bound from the proposition 9, we get the announced result. Finally, since

$K\log(\lceil T/K\rceil) \sim K\log(T/K)$ and $K\sqrt{\lceil T/K\rceil} \sim \sqrt{KT}$, we get from lemma 6 that, when $T > d > 0$:

$$\hat{R}_{T,1} \leq C + C_1 d\log\left(\frac{T}{\delta}\right)\sqrt{KT} + C_2 d\log\left(\frac{T}{\delta}\right)\sqrt{TK\log(T/K)}$$

with C,C_1 and C_2 three constants. Since the last term is clearly the greatest, we get the announced result.

A.8. Proof of the proposition 10

We follow a similar method to the one presented in Qin et al. (2014) for the specific case of sums of individual rewards: Since we consider that the reward obtained by playing a set of actions at a given step t is the sum of rewards observed for every single played action at t, we have:

$$reg_t = \sum_{i\in\mathcal{K}^\star} \mu_i^\top \beta - \sum_{i\in\mathcal{K}_t} \mu_i^\top \beta$$

with $\mathcal{K}^\star$ the set of k optimal actions, i.e., those with greatest expectations $\mu_i^\top\beta$. Then, we use the fact that for every step t:

$$\sum_{i\in\mathcal{K}_t} s_{i,t} \geq \sum_{i\in\mathcal{K}_\star} s_{i^\star,t}$$

This leads to:

$$reg_t \leq \sum_{i\in\mathcal{K}_t} 2\alpha_{t-1}||\hat{x}_{i,t}||_{V_{t-1}^{-1}} + 4\sqrt{d}(\alpha_{t-1}/\sqrt{\lambda} + S)\rho_{i,t,\delta}$$

Where the matrix V_{t-1} is defined by considering the k played actions at each step: $V_{t-1} = \lambda I + \sum_{s=1}^{t-1}\sum_{i\in\mathcal{K}_s} \frac{1}{R_{i,t}}\hat{x}_{i,t}\hat{x}_{i,t}^\top$. The fact that we add k terms to the matrix V at each iteration implies two distinct things:

- First on the confidence ellipsoid of β. For k actions played at each step, we have:
  $\alpha_T \leq \sqrt{d\log\left(\frac{1+kTL^2/\lambda}{\delta}\right)} + \sqrt{\lambda}S$;
- On another hand on the upper-bounding of $\sum_{t=1}^{T}\sum_{i_t\in\mathcal{K}_t} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}}$. We have:
  $\sum_{t=1}^{T}\sum_{i_t\in\mathcal{K}_t} ||\mu_{i_t} - \check{\epsilon}_{i_t,t}||^2_{W_{t-1}^{-1}} \leq \sqrt{R^2+L^2S^2}d\log\left(1+\frac{TkL^2}{\lambda d}\right)$

Finally, we can use the same methods as in the previous proofs for deriving specific bounds from the generic one, where the selection of k actions at each step appears explicitly in the terms $\sum_{t=1}^{T}\sum_{i\in\mathcal{K}_t}\frac{1}{n_{i,t}}$ and $\sum_{t=1}^{T}\sum_{i\in\mathcal{K}_t}\frac{1}{\sqrt{n_{i,t}}}$.

A.9. Table of the main notations

T	number of steps of the process
$\mathcal{K}$	set of available actions
K	number of available arms
k	number of simultaneaous plays at each step
d	dimension of the arms' profiles
i_t	arm selected at step t
i^*	arm with the highest final cumulative reward
r_i	reward obtained by arm i at step t
$x_{i,t}$	profile sample vector observed for arm i at step t
$\hat{x}_{i,t}$	average of profile sample vectors observed for arm i until step t
L	upper-bound for the profiles' norm
$\mathcal{O}_t$	set of arms delivering a profile context at step t
$n_{i,t}$	number of samples observed for arm i until step t
μ_i	profile vector of arm i
β	mapping parameter between profiles and rewards
$\hat{\beta}_t$	estimator of β at step t
S	upper-bound for the β parameter norm
λ	l2-regularization constant of the β estimator
$\eta_{i,t}$	sub-gaussian noise of the reward of arm i at step t
R	sub-gaussian constant of the rewards distribution
$\epsilon_{i,t}$	deviation between the true profile of arm i and its estimator at step t ($\epsilon_{i,t} = \mu_i - \hat{x}_{i,t}$)
η'_{t-1}	vector of reward deviations of the first $t-1$ selected arms from their expectation at step t: $\eta'_{t-1} = (\eta_{i_s,s} + \epsilon_{i_s,t}^\top \beta)_{s=1..t-1}^\top$
$R_{i,t}$	sub-gaussian constant of the noise of the reward of i at step t w.r.t. $\hat{x}_{i,t}^\top \beta$
X_{t-1}	$(t-1)\times d$ matrix containing the empirical means of the selected actions, where the s-th row corresponds to the estimator at step t of the action selected at step s: $X_{t-1} = (\hat{x}_{i_s,t}^\top)_{s=1..t-1}$
Y_{t-1}	rewards vector of size $t-1$: $Y_{t-1} = (r_{i_s,s})_{s=1..t-1}^\top$
A_{t-1}	diagonal $(t-1)\times(t-1)$ matrix, where the s-th diagonal element equals $1/R_{i_s,t}$: $A_{t-1} = diag(1/R_{i_s,t})_{s=1..t-1}$
δ	parameter controling the confidence level of the regret bound
V_t^{-1}	variance-covariance matrix of the posterior distribution of β at step t
$\rho_{i,t,\delta}$	quantity used to bound the deviation of the estimators of profiles: $\rho_{i,t,\delta} = \min(L\sqrt{\frac{2d}{n_{i,t}} \log\left(\frac{2dt^2}{\delta}\right)}, 2L)$
$s_{i,t}$	selection score for arm i at step t
α_t	exploration coefficient at step t w.r.t. the confidence of the *beta* estimator
$\bar{\epsilon}_t$	quantity used to bound $\mu_i^\top \hat{\beta}_t$
$\tilde{\epsilon}_t$	quantity used to bound $\mu_i^\top (\beta - \hat{\beta}_t)$

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 39.1–39.26, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 127–135, 2013.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration-exploitation tradeoff using variance estimates in multi-armed bandits. *Theor. Comput. Sci.*, 410(19):1876–1902, April 2009. ISSN 0304-3975.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 2003.

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

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Proceedings of the 25th Conference on Neural Information Processing Systems 2011, Granada, Spain*, pages 2249–2257, 2011.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 151–159, 2013.

Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pages 208–214, 2011.

Varsha Dani, Thomas P. Hayes, and Sham M. Kakade. Stochastic linear optimization under bandit feedback. In *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, pages 355–366, 2008.

V. H. de la Peña, T. L. Lai, and Q. M. Shao. *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer Series in Probability and its Applications. Springer, 2009.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6-9 December 2010, Vancouver, British Columbia, Canada.*, pages 586–594, 2010.

Aurlien Garivier. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, 2011.

Thibault Gisselbrecht, Ludovic Denoyer, Patrick Gallinari, and Sylvain Lamprier. Whichstreams: A dynamic approach for focused data capture from large social media. In *Proceedings of the Ninth International Conference on Web and Social Media, ICWSM 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, pages 130–139, 2015.

Liangjie Hong and Brian D. Davison. Empirical study of topic modeling in twitter. In *ECIR*, 2010.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On bayesian upper confidence bounds for bandit problems. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012*, pages 592–600, 2012a.

Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*, pages 199–213, 2012b.

T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4 – 22, 1985. ISSN 0196-8858.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 661–670, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8.

Benedict C. May, Nathan Korda, Anthony Lee, and David S. Leslie. Optimistic bayesian sampling in contextual-bandit problems. *J. Mach. Learn. Res.*, 13:2069–2106, 2012. ISSN 1532-4435.

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, Philadelphia, Pennsylvania, USA, April 24-26, 2014*, pages 461–469, 2014.

Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, 2010.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Bulletin of the American Mathematics Society*, 25:285–294, 1933.