JMLR: Workshop and Conference Proceedings 83:1–24, 2018 29th Conference on Algorithmic Learning Theory

Corrupt Bandits for Preserving Local Privacy

Pratik Gajane PRATIK.GAJANE@UNILEOBEN.AC.AT
Montanuniversität Leoben

Tanguy Urvoy TANGUY.URVOY@ORANGE.COM
Orange labs

Emilie Kaufmann EMILIE.KAUFMANN@UNIV-LILLE1.FR
CNRS & Univ. Lille, UMR 9189 (CRIStAL), Inria Lille Nord-Europe, SequeL team

Abstract

We study a variant of the stochastic multi-armed bandit (MAB) problem in which the rewards are corrupted. In this framework, motivated by privacy preservation in online recommender systems, the goal is to maximize the sum of the (unobserved) rewards, based on the observation of transformation of these rewards through a stochastic corruption process with known parameters. We provide a lower bound on the expected regret of any bandit algorithm in this corrupted setting. We devise a frequentist algorithm, KLUCB-CF, and a Bayesian algorithm, TS-CF and give upper bounds on their regret. We also provide the appropriate corruption parameters to guarantee a desired level of local privacy and analyze how this impacts the regret. Finally, we present some experimental results that confirm our analysis.

Keywords: Sequential learning, multi-armed bandits, incomplete feedback, local privacy

1. Introduction

The classical multi-armed bandits (MAB) problem is the formulation of the exploration-exploitation dilemma inherent to reinforcement learning (see Bubeck and Cesa-Bianchi, 2012, for a survey). In this setup, a learner has access to a number of available actions, also called "arms" in reference to the arm of a slot machine or a one-armed bandit. They have to repeatedly select (or "draw") one of these arms, which yields a reward generated from an unknown reward process, with the aim to maximize the sum of the gathered rewards. After each arm selection, a *feedback* is provided to the learner, that shall influence their arm selection strategy in the next rounds. In the classical MAB problem, the feedback is the observation of the reward itself. However, this assumption does not hold true for some practical scenarios. For example, in adaptive routing, positive feedback means the corresponding path is usable but no feedback could either mean that the corresponding path is unusable or the feedback was dropped due to extraneous issues. In the literature, such an *asymmetric feedback* is called *Positive and Unlabeled (PUN)* feedback. See Zhang and Zuo (2008) for a survey.

On-purpose feedback corruption is an effective way to protect the respondent's individual privacy in online recommender systems or survey systems. For instance, Warner (1965) proposed the *randomized response method* as a survey technique to reduce potential bias due to non-response and social desirability when asking questions about sensitive behaviors and beliefs. This method asks the respondents to employ randomization, say with a coin flip,

the outcome of which is not available to the interviewer. By introducing random noise, the method conceals the individual responses and protects respondent privacy. This method could also be applied within a recommender system, that would thus receive corrupted version of the user's original feedback about the items presented. Contrary to most previous works which apply privacy at the recommender level, this privacy mechanism, called *local privacy*, can be deployed at the user level. The challenge for the recommender is then to present good items to the users (in terms of their "true" feedback), based only on the received corrupted feedback. Moreover, users may be willing to tune the level of corruption in order to balance between their privacy and the utility of the recommendation they obtain.

The corrupted feedback we consider is a particular type of an incomplete feedback. Therefore, the natural framework to deal with this situation appears to be *Partial Monitoring (PM)* (Piccolboni and Schindelhauer (2001); Bartók et al. (2014)), which is a general framework for sequential decision making problems with incomplete feedback. The partial monitoring problem may be either *trivial* with a minimax regret of 0, *easy* with a minimax regret $\tilde{\Theta}(\sqrt{T})$ at time T, *hard* with a minimax regret $\tilde{\Theta}(T^{2/3})$, or *hopeless* with a linear minimax regret. The MAB problem with corrupted feedback however does not fit directly in the PM setting as defined in Bartók et al. (2014) since it requires additional constraints on the environment. Specifically, the infinite observation usually space assumed in the literature of finite stochastic PM precludes the PM results from being applicable to our setting. In this work, exploiting the specificity of the corrupted MAB problem, we aim for the best problem-dependent regret, that scales with $\log(T)$.

The article is structured as follows. In Section 2, we formally define the corrupted MAB prob lem and the relevant parameters. In Section 3, a lower bound on the regret of any corrupt bandit algorithm is given. In Section 4, the algorithms kl-UCB-CF and TS-CF are introduced and we provide upper bounds on their regret. In Section 5, we describe how corrupted feedback can be used to enforce privacy. The proof sketches for the lower and the upper bounds are given in Section 7, while the complete proofs are postponed to the appendices. The penultimate section, Section 6, gives an overview of our experiments on the proposed algorithms.

2. The Corrupt Bandit Problem

A (stochastic) corrupt bandit problem $\boldsymbol{\nu}$ is formally characterized by a set of arms $A = \{1, \dots, K\}$ on which are indexed a list of unknown sub-Gaussian reward distributions $\{\nu_a\}_{a\in A}$, a list of unknown sub-Gaussian feedback distributions $\{\varsigma_a\}_{a\in A}$, and a list of known *mean-corruption functions* $\{g_a\}_{a\in A}$.

If the learner pulls an arm $a \in A$ at time t, they receive a reward R_t drawn from the distribution ν_a with mean $\mu_a^{\boldsymbol{\nu}}$ and observe a feedback F_t drawn from the distribution ς_a with mean $\lambda_a^{\boldsymbol{\nu}}$. We assume that, for each arm, there exists a loose link between the reward and the feedback through a known mean-corruption function (or simply, corruption function) g_a which maps the mean of the reward distribution to the mean of the feedback distribution : $g_a(\mu_a^{\boldsymbol{\nu}}) = \lambda_a^{\boldsymbol{\nu}}, \forall a \in A$. Note that these g_a functions may be completely different from one arm to another. For Bernoulli distributions, μ_a and λ_a are in $[0, 1]$ for all $a \in A$ and we assume all the corruption functions $\{g_a\}_{a\in A}$ to be continuous aleast in this interval.

Let $a_*(\boldsymbol{\nu}) \in \arg\max \mu_a^{\boldsymbol{\nu}}$ be the optimal arm in the corrupt bandit model $\boldsymbol{\nu}$[1]. Without loss of generality, we assume when presenting the results that arm 1 is the optimal arm for the rest of this article, unless otherwise specified. The objective is to design a strategy, which chooses an arm $\hat{a}_t$ to be pulled at time t based only on the previously observed feedback, $F_1, \dots, F_{t-1}$, in order to maximize the expected sum of rewards, or equivalently to minimize the regret: $\mathrm{Regret}_T(\boldsymbol{\nu}) \coloneqq \mathbb{E}_{\boldsymbol{\nu}}\left[\mu_1 \cdot T - \sum_{t=1}^T R_t\right] = \sum_{a=2}^K \Delta_a \cdot \mathbb{E}_{\boldsymbol{\nu}}[N_a(T)]$ where $N_a(T) \coloneqq \sum_{t=1}^T \mathbb{1}_{(\hat{a}_t=a)}$ denotes the number of pulls of arm a up to time T and $\Delta_a \coloneqq \mu_1 - \mu_a$ i.e. the gap between the optimal mean reward and the mean reward of arm a.

Another way to define the link between the reward and the feedback is to provide a *corruption scheme* operator $\tilde{g}_a$ which maps the reward outcomes into feedback distributions. If the mean is a sufficient statistic of the reward distribution, then the learner can build their own corruption function from the corruption scheme and the two definitions are equivalent. This equivalence is true for Bernoulli distributions where most of our results apply.

Randomized response. Randomized response (Warner (1965)), described in the introduction, can be simulated by a Bernoulli corrupt bandit and the corresponding corruption scheme $\tilde{g}_a$ can be encoded by the matrix:

$$\mathbb{M}_a \coloneqq \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{bmatrix} p_{00}(a) & 1-p_{11}(a) \\ 1-p_{00}(a) & p_{11}(a) \end{bmatrix} \end{array} \tag{1}$$

where $\mathbb{M}_a(y,x) \coloneqq \mathbb{P}(\text{Feedback from arm } a=y \,|\, \text{Reward from arm } a=x)$. The corresponding linear corruption function is $g_a(x) = 1 - p_{00}(a) + [p_{00}(a) + p_{11}(a) - 1] \cdot x$.

3. Lower Bound on the Regret for MAB with Corrupted Feedback

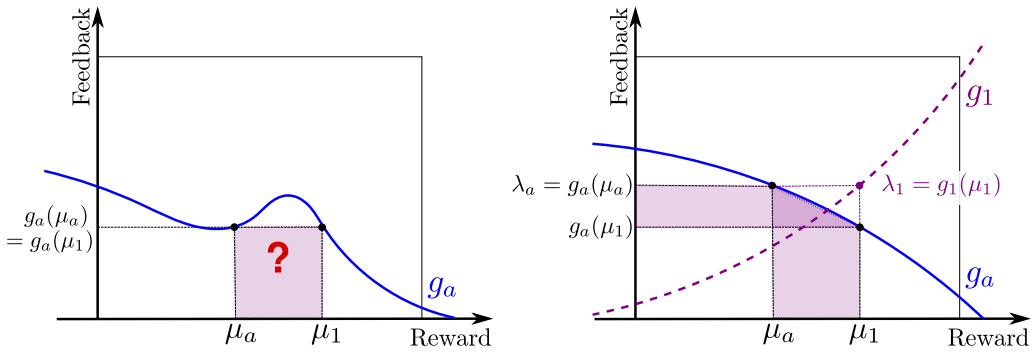
Following a definition by Lai and Robbins (1985) for the classical MAB, we define a *uniformly efficient* algorithm for the corrupt bandit problem as an algorithm which, for any problem instance $\boldsymbol{\nu}$, has $\mathrm{Regret}_T(\boldsymbol{\nu}) = o(T^\alpha)$ for all $\alpha \in]0,1[$. Theorem 1 provides a lower bound on the regret of a uniformly efficient algorithm, in terms of the Kullback-Leibler (KL) divergence between some distributions. We denote by $d(x,y)$ the KL-divergence between the Bernoulli distribution of mean x and that of mean y.

Theorem 1 *Given continuous corruption functions $\{g_a\}_{a\in A}$, any uniformly efficient algorithm for a Bernoulli corrupt bandit problem satisfies,* $\liminf_{T\to\infty} \frac{\mathrm{Regret}_T}{\log(T)} \geq \sum_{a=2}^K \frac{\Delta_a}{d(\lambda_a, g_a(\mu_1))}$.

The proof of Theorem 1 can be found in Section 7.1.

The lower bound reveals that the divergence between the mean feedback from $a \in A$ and the image of the optimal reward μ_1 with g_a plays a crucial role in distinguishing arm a from the optimal arm. The shape of the g_a function in the neighborhood of both a and 1 has a great impact on the information the learner can extract from the received feedback. Particularly, if the g_a function is non-monotonic, as shown in Figure 1a, it might be impossible to distinguish between arm a and the optimal arm. To circumvent

[1] When the associated model is clear from the context, we drop the symbol $\boldsymbol{\nu}$.



(a) Uninformative g_a function (b) Informative g_a function

Figure 1: In Figure 1a, g_a such that $\lambda_a = g_a(\mu_1)$ thereby making it impossible to discern arm a from the optimal arm given the mean feedback. In Figure 1b, a steep monotonic g_a leads the reward gap $\Delta_a = \mu_1 - \mu_a$ into a clear gap between λ and $g_a(\mu_1)$.

this problem, we assume the corruption functions $\{g_a\}_{a\in A}$ to be strictly monotonic in our algorithms and we denote its corresponding inverse function by g_a^{-1}. Such an informative corruption function is shown in Figure 1b. To clarify that the gap between λ_a and λ_1 is not relevant here, we also plot in Figure 1b, a corruption function g_1 which differs from g_a and causes fortuitously the two arms to have the same mean feedback with different interpretations in terms of mean rewards.

4. Algorithms for MAB with Corrupted Feedback

There are two popular approaches to solve the MAB problem in its many variations: the frequentist approach and the Bayesian approach. In this article, we propose both a frequentist and a Bayesian algorithm for the problem at hand.

4.1. kl-UCB for MAB with Corrupted Feedback (kl-UCB-CF)

We propose in Algorithm 1 an adaptation of the kl-UCB algorithm of Cappé et al. (2013). $\text{Index}_a(t)$ is an upper-confidence bound on μ_a built from a confidence interval on λ_a based on the KL-divergence. The quantity $\hat{\lambda}_a(t)$ in the algorithm denotes the empirical mean of the feedback observed from arm a until time t: $\hat{\lambda}_a(t) \coloneqq \frac{1}{N_a(t)} \sum_{s=1}^{t} F_s \cdot \mathbb{1}_{(\hat{a}_s=a)}$.

Theorem 2 gives an upper bound on the regret of kl-UCB-CF, showing that it matches the lower bound given in Theorem 1. A more explicit finite-time bound is proved in Appendix I.

Theorem 2 kl-UCB-CF *using* $f(t) \coloneqq \log(t) + 3\log(\log(t))$ *on a* K*-armed Bernoulli corrupt bandit with strictly monotonic and continuous corruption functions* $\{g_a\}_{a\in A}$ *satisfies at time* T*,*

$$\text{Regret}_T \leq \sum_{a=2}^{K} \frac{\Delta_a \log(T)}{d\left(\lambda_a, g_a(\mu_1)\right)} + O(\sqrt{\log(T)}).$$

Algorithm 1 KLUCB for MAB with corrupted feedback (kl-UCB-CF)

Input: A bandit model with a set of arms $A \coloneqq \{1, \dots, K\}$ with unknown mean rewards $\mu_1, \dots, \mu_K$ and unknown mean feedbacks $\lambda_1, \dots, \lambda_K$ and monotonic and continuous corruption functions $g_1, \dots, g_k$.
Parameters: A non-decreasing (exploration) function $f : \mathbb{N} \to \mathbb{R}$,
$d(x, y) \coloneqq \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$, Time horizon T.

1. **Initialization:** Pull each arm once.
2. **for** time $t = K, \dots, T-1$ **do**
   (a) Compute for each arm a in A the quantity

$$\mathrm{Index}_a(t) \coloneqq \max\left\{q : \ N_a(t) \cdot d(\hat{\lambda}_a(t), g_a(q)) \leq f(t)\right\}$$

   (b) Pull arm $\hat{a}_{t+1} \coloneqq \underset{a}{\mathrm{argmax}}\ \mathrm{Index}_a(t)$ and observe the feedback F_{t+1}.
   end for

The UCB1 algorithm (Auer et al. (2002)) can also be updated to UCB-CF to deal with the corrupted feedback by modifying the index to

$$\mathrm{Index}_a(t) \coloneqq \begin{cases} g_a^{-1}\left(\hat{\lambda}_a(t) + \sqrt{\frac{f(t)}{2N_a(t)}}\right) & \text{if increasing } g_a \\ g_a^{-1}\left(\hat{\lambda}_a(t) - \sqrt{\frac{f(t)}{2N_a(t)}}\right) & \text{if decreasing } g_a \end{cases}$$

Corollary 1 *With $f(t) \coloneqq \log(t) + 3\log(\log(t))$, the regret of* UCB-CF *at time T on a K-armed Bernoulli corrupt bandit with strictly monotonic and continuous corruption functions $\{g_a\}_{a \in A}$ is in $O\Big(\sum_{a=2}^{K} \frac{\Delta_a \log(T)}{(\lambda_a - g_a(\mu_1))^2}\Big)$.*

The proof of this corollary follows the proof of Theorem 2, using the quadratic divergence $2(x-y)^2$ in place of $d(x, y)$ through Pinsker's inequality. UCB-CF is only order optimal with respect to the bound of Theorem 1, but its index is simpler to compute.

4.2. Thompson Sampling for MAB with Corrupted Feedback (TS-CF)

TS-CF maintains a Beta posterior distribution on the mean feedback of each arm. At time $t+1$, for each arm a, it draws a sample $\theta_a(t)$ from the posterior distribution on λ_a and pulls the arm which maximizes $g_a^{-1}(\theta_a(t))$. This mechanism ensures that at each time, the probability that arm a is played is the posterior probability of this arm to be optimal, as in classical Thompson Sampling (TS) (Thompson (1933)).

Theorem 3 *When* TS-CF *is run on a K-armed Bernoulli corrupt bandit with strictly monotonic and continuous corruption functions $\{g_a\}_{a \in A}$, for all $\psi > 0$, there exists a constant $C_\psi \coloneqq C(\psi, \{\mu_a\}_{a \in A}, \{g_a\}_{a \in A})$ such that at time T,*

$$\mathrm{Regret}_T \leq (1+\psi) \sum_{a=2}^{K} \frac{\Delta_a \log(T)}{d(\lambda_a, g_a(\mu_1))} + C_\psi.$$

Algorithm 2 Thompson sampling for MAB with corrupted feedback (TS-CF)

Input: A bandit model with a set of arms $A \coloneqq \{1, \dots, K\}$ arms with unknown reward means $\mu_1, \dots, \mu_K$ and unknown feedback means $\lambda_1, \dots, \lambda_K$ and monotonic and continuous corruption functions $g_1, \dots, g_K$.
Parameters: Time horizon T.

1. **Initialization:** For each arm a in A, set $\text{success}_a = 0$ and $\text{fail}_a = 0$
2. **for** time $t = 0, \dots, T-1$ **do**
   (a) For each arm a in A, sample $\theta_a(t)$ from $\text{Beta}(\text{success}_a + 1, \text{fail}_a + 1)$.
   (b) Pull arm $\hat{a}_{t+1} \coloneqq \arg\max_a g_a^{-1}(\theta_a(t))$ and observe the feedback F_{t+1}.
   (c) **if** $F_{t+1} = 1$ **then**
   i. $\text{success}_{\hat{a}_{t+1}} = \text{success}_{\hat{a}_{t+1}} + 1$
   else
   ii. $\text{fail}_{\hat{a}_{t+1}} = \text{fail}_{\hat{a}_{t+1}} + 1$
   end if
   end for

This theorem also yields the asymptotic optimality of TS-CF with respect to the lower bound given in Theorem 1. We give a sketch of its proof in Section 7.3.

We can use the above algorithms on a MAB problem with randomized response. The following corollary bounds their regret.

Corollary 2 *The regret of* kl-UCB-CF *and* TS-CF *for a K-armed Bernoulli MAB problem with randomized response using corruption matrices $\{\mathbb{M}\}_{a \in A}$ at time T is* $\sum_{a=2}^{K} \frac{2\log(T)}{\Delta_a(p_{00}(a)+p_{11}(a)-1)^2} + O(\sqrt{\log(T)})$.

This corollary follows from Theorem 2 and 3 together with Pinsker's inequality: $d(x, y) > 2(x-y)^2$. The term $(p_{00}(a) + p_{11}(a) - 1)$ is the slope of the corruption function for arm a.

5. Corrupted Feedback to Preserve Local Differential Privacy

Differential privacy (DP), introduced by Dwork et al. (2006), is one of the usual approaches for the privacy concerns. Dwork and Roth (2014) present a comprehensive overview. Jain et al. (2012); Thakurta and Smith (2013); Mishra and Thakurta (2015); Tossou and Dimitrakakis (2016) have observed the importance of privacy to MAB applications. Recently, the notion of differential privacy has been extended to *local differential privacy* by Duchi et al. (2014) in which data remains private even from the learner.

Definition 1 *(Locally differentially private mechanism) Any randomized mechanism $\mathcal{M}$ is ϵ-locally differentially private for $\epsilon \geq 0$ if for all $d_1, d_2 \in Domain(\mathcal{M})$ and for all $S \subset Range(\mathcal{M})$,*

$$\mathbb{P}[\mathcal{M}(d_1) \in S] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{M}(d_2) \in S].$$

In both global and local contexts, differential privacy is achieved by the addition of noise. The main difference between global and local differential privacy is whether privacy is to be maintained from the algorithm or the (possibly unintended) recipient of the output of

the algorithm. In global differential privacy, noise is added by the algorithm so the output does not reveal private information about the input. In local differential privacy, noise is added to the input of the algorithm so that privacy is maintained even from the algorithm. To the best of our knowledge, hitherto all the previous work combining differential privacy and bandits has used global differential privacy, either within a stochastic (Mishra and Thakurta (2015), Tossou and Dimitrakakis (2016)) or an adversarial (Thakurta and Smith (2013), Tossou and Dimitrakakis (2017)) bandit problem.

In this article, we consider local differential privacy. To understand the motivation for local differential privacy, let us consider these settings in the context of Internet advertising, which is one of the major applications of bandit algorithms. An advertising system receives, as input, feedback from the users which may reveal private information about them. The advertising system employs a suitable bandit algorithm and selects the ads for the users tailored to the feedback given by them. These selected ads are then given to the advertisers as the output[2]. While using global differential privacy, privacy is maintained from the advertisers by ensuring that the output of the bandit algorithms does not reveal information about the input (i.e. user information). Typically, advertising systems are established by leading social networks, web browsers and other popular websites. Korolova (2010), Kosinski et al. (2013) show that it is possible to accurately predict a range of highly sensitive personal attributes including age, sexual orientation, relationship status, political and religious affiliation, presence or absence of a particular interest, as well as exact birthday using the the feedback available to the advertising systems. Such possible breach of privacy necessitates us to protect personal user information not only from the advertisers but also from the advertising systems. Local differential privacy is able to achieve this goal unlike global differential privacy.

Recently, Wang et al. (2016) addressed a similar scenario in data collection. They used randomized response to perturb sensitive information before being collected by an untrusted server so as to limit the server's ability to learn the sensitive information with confidence. We too shall use the corruption process as a mechanism to provide local differential privacy.

Definition 2 *(ϵ-locally differentially private bandit feedback corruption scheme) A bandit feedback corruption scheme $\tilde{g}$ is ϵ-locally differentially private for $\epsilon \geq 0$ if for all reward sequences $R_{t1}, \ldots, R_{t2}$ and $R'_{t1} \ldots, R'_{t2}$, and for all $\mathcal{S} \subset Range(\tilde{g})$,*

$$\mathbb{P}[\tilde{g}(R_{t1}, \ldots, R_{t2}) \in \mathcal{S}] \leq e^{\epsilon} \cdot \mathbb{P}[\tilde{g}(R'_{t1}, \ldots, R'_{t2}) \in \mathcal{S}].$$

In the case where corruption is done by randomized response, local differential privacy requires that $\max_{1 \leq a \leq K} \left(\frac{p_{00}(a)}{1-p_{11}(a)}, \frac{p_{11}(a)}{1-p_{00}(a)} \right) \leq e^{\epsilon}$. By ensuring the appropriate values for the parameters of randomized response, users can send differentially private feedback to the learner. The learner can then employ kl-UCB-CF or TS-CF to learn from such feedback. From Corollary 2, we can see that to achieve lower regret, $p_{00}(a) + p_{11}(\mathrm{a})$ is to be maximized for all $a \in A$. Using Result 1 from Wang et al. (2016), we can state that, in order to achieve

[2]This description does not express our belief of how real-life Internet advertising systems work. We use it for the purpose of illustration only.

ϵ-local differential privacy while maximizing $p_{00}(a) + p_{11}(a)$,

$$\mathbb{M}_a = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} \frac{e^\epsilon}{1+e^\epsilon} & \frac{1}{1+e^\epsilon} \\ \frac{1}{1+e^\epsilon} & \frac{e^\epsilon}{1+e^\epsilon} \end{bmatrix} \end{array}. \tag{2}$$

As it turns out, this is equivalent to the *staircase* mechanism for local privacy given in Kairouz et al. (2016, Eq. (15)) for binary rewards and feedbacks. Moreover, this is the optimal local differential privacy mechanism for *low privacy regime* (Kairouz et al., 2016, Theorem 14). In low privacy regime, the noise added to the data is small and the aim of the privacy mechanism is to send as much information about data as allowed, but no more (Kairouz et al., 2014). This is in alignment with our dual goal of using privacy with bandit algorithms: learn from the data while respecting the privacy as much as possible. The trade-off between utility and privacy is controlled by ϵ. At one extreme, for $\epsilon = 0$, feedbacks are independent of rewards and learning about rewards from feedbacks is not possible. On the other extreme, for $\epsilon = \infty$, feedbacks can be made equal to rewards.

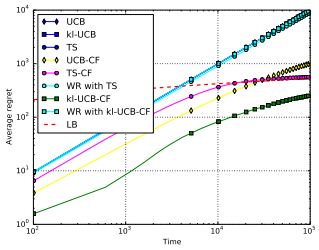
Using the corruption parameters from Eq. (2) with Corollary 2, we arrive at the following upper bound.

Corollary 3 *The regret of* kl-UCB-CF *or* TS-CF *at time T with ϵ-locally differentially private bandit feedback corruption scheme is* $\text{Regret}_T \leq \sum_{a=2}^{K} \frac{2\log(T)}{\Delta_a \left(\frac{e^\epsilon-1}{e^\epsilon+1}\right)^2} + O(\sqrt{\log(T)})$.

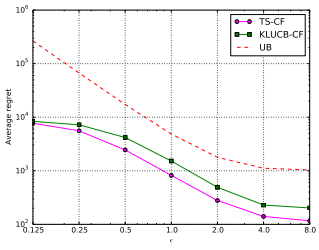
The term $\left(\frac{e^\epsilon-1}{e^\epsilon+1}\right)^2$ in the above expression conveys the relationship of the regret with the level of local differential privacy symbolized by ϵ. For low values of ϵ, $\left(\frac{e^\epsilon-1}{e^\epsilon+1}\right) \approx \epsilon/2$. This is in-line with the regret of the stochastic bandit algorithms providing global DP given by Mishra and Thakurta (2015, Theorem 4 and 8) which have a multiplicative factor of $\mathrm{O}(\epsilon^{-1})$ or $\mathrm{O}(\epsilon^{-2})$. Tossou and Dimitrakakis (2016, Corollary 3.2 and Theorem 3.5) provided a regret bound for a stochastic bandit algorithm achieving global DP with an additive factor of $\mathrm{O}(\epsilon^{-1})$. Our lower bound, given in Theorem 1, shows that such an improvement is not expected for local differential privacy as parameters of the corruption mechanism are featured in the (asymptotic) multiplicative factor of $\log(T)$. It is also worthwhile to recall that local differential privacy comes at a higher price for the user : as local DP is a more stringent privacy notion than global DP, it is justifiable that the regret of the algorithms providing the latter is lower than that of the algorithms providing the former.

6. Empirical Evaluation

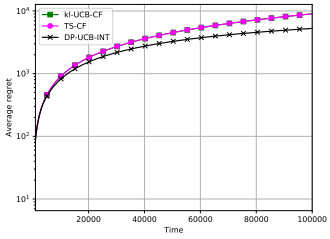
Before delving into the empirical evaluation, we first describe a naive algorithm called, Wrapper to be used as a baseline. This algorithm simply applies the appropriate inverse corruption function to the received feedback values and uses the result as a substitute for empirical reward. It then treats the corrupt bandit problem as a classical MAB problem and solves it using any classical MAB algorithm as a black-box. It is easy to see that this naive algorithm won't work for the corruptions functions in which $\mathbb{E}(g_\cdot^{-1}(y)) \neq g_\cdot^{-1}(\mathbb{E}(y))$. Even while using linear corruption functions, this algorithm gives worse performance than the algorithms provided in this article, as can be verified below. The inferior performance



(a) Regret plots for kl-UCB-CF and TS-CF with others for varying horizons up to 10^5



(b) Regret plots for kl-UCB-CF and TS-CF with $T = 10^5$, $\epsilon = \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$



(c) Regret plots for $\epsilon = 1$ for kl-UCB-CF and TS-CF , DP-UCB-Int

Figure 2: Regret curves

is because this naive algorithm doesn't take into account the variance of the sequence generated by applying inverse corruption functions to the received feedback values.

We provide here the evaluation of the algorithms on a 10-armed Bernoulli corrupt bandit problem. The reward means of the arms were set as follows:

$$\mu_1 = 0.9 \qquad \mu_2 = \mu_3 = \cdots = \mu_{10} = 0.8$$

Further experiments can be found in Appendix III.

6.1. Regret over a period of time

In this experiment, we aim to see the effect of time on the regret of kl-UCB-CF and TS-CF. Randomized response was employed to corrupt the feedback with $p_{00} = p_{11} = 0.6$ for the optimal arm, while for all the other arms, both p_{00} and p_{11} were set to 0.9. The time horizon was varied to 10^5 and each experiment was repeated 1000 times. As a baseline, we plot the regret curves for two instances of the Wrapper algorithm (denoted as WR) with kl-UCB and TS used as the black-box subroutine respectively. To demonstrate the inability of the traditional MAB algorithms to solve the corrupt bandit problem, we also include the regret curves for kl-UCB ,UCB1 and TS (treating feedback as reward). The regret curves for all the considered algorithms are given in Figure 2a. LB denotes the lower bound given by Theorem 1. The performance superiority of the proposed algorithms for corrupt bandits is more pronounced as the time increases.

6.2. Regret with varying level of local differential privacy

In this experiment, we vary the local differential privacy parameter and examine the effect on the regret of kl-UCB-CF and TS-CF. We chose ϵ from the set $\{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$. The corruption parameters are set by substituting the values of ϵ in Eq. (2). The time horizon was fixed to 10^5 and the experiment was repeated 1000 times. The corresponding curves for average regret can be seen in Figure 2b. UB indicates the upper bound given by Corollary 3. The regret for both the algorithms decreases with increasing ϵ. This behavior is expected

since, lower the value of ϵ, more stringent is the level of differential privacy. Towards both the end points of the range ($\epsilon < 1/4$ and $\epsilon > 4$), the regret tends to plateau as a change in ϵ causes an infinitesimal change in the required level of differential privacy.

6.3. Regret for local and global differential privacy

For comparison, we plot the regret of kl-UCB-CF and TS-CF against the recent stochastic bandit algorithm for global DP, DP-UCB-Int (Tossou and Dimitrakakis (2016)) for $\epsilon = 1$ in Figure 2c. The comparison aims to convey how much utility, in terms of regret, is lost by opting for local DP instead of global DP. As already mentioned, lower regret for achieving global DP is to be expected as local DP is a much stronger notion of privacy than global DP. For DP-UCB-Int, we chose the same values of the algorithm parameters ($\delta = e^{-10}$ and $v = 1.1$) as in the experiments given in Tossou and Dimitrakakis (2016, Section 4).

7. Elements of proofs

We denote by $\hat{\lambda}_a(t)$ the empirical mean of the feedback obtained from arm a until time t. Letting $F_{a,s}$ being the successive feedbacks of arm a and $\hat{\lambda}_{a,s} := \frac{1}{s}\sum_{\ell=1}^{s} F_{a,\ell}$, one has $\hat{\lambda}_a(t) = \hat{\lambda}_{a,N_a(t)}$ when $N_a(t) > 0$.

7.1. Proof of Theorem 1

To obtain a lower bound on the regret, we use a *change-of-distribution* argument. Let $\boldsymbol{\nu}$ and $\boldsymbol{\nu'}$ be K-armed corrupted bandit models with different optimal arms i.e. $a_*(\boldsymbol{\nu}) \neq a_*(\boldsymbol{\nu'})$. For the ease of readability, let's assume without loss of generality that $a_*(\boldsymbol{\nu}) = 1$.

The log-likelihood ratio of the observations up to time T under $\boldsymbol{\nu}$ and $\boldsymbol{\nu}$, $L_T(\boldsymbol{\nu}, \boldsymbol{\nu'})$, can be written as $L_T(\boldsymbol{\nu}, \boldsymbol{\nu'}) = \sum_{a=1}^{K}\sum_{s=1}^{N_a(T)} \log \frac{f_{\lambda_a^{\boldsymbol{\nu}}}(F_{a,s})}{f_{\lambda_a^{\boldsymbol{\nu'}}}(F_{a,s})}$ where $f_x(\cdot)$ denotes the Bernoulli density of mean x. By Wald's lemma, $\mathbb{E}_{\boldsymbol{\nu}}\left[L_T(\boldsymbol{\nu}, \boldsymbol{\nu'})\right] = \sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}[N_a(T)] \cdot d(\lambda_a^{\boldsymbol{\nu}}, \lambda_a^{\boldsymbol{\nu'}})$.

The following lemma can be extracted from Garivier et al. (2016).

Lemma 1 *Let $\boldsymbol{\nu}$ and $\boldsymbol{\nu}'$ be two bandit models with K arms and and $T \in \{0\} \cup \mathbb{N}$, then: $\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}[N_a(T)] \cdot KL(\lambda_a^{\boldsymbol{\nu}}, \lambda_a^{\boldsymbol{\nu}'}) \geq d\left(\mathbb{E}_{\boldsymbol{\nu}}(Z), \mathbb{E}_{\boldsymbol{\nu}'}(Z)\right)$, where $d(x, y)$ is the binary relative entropy and $Z \in [0, 1]$ is a random variable measurable from the past-observations filtration.*

Using Lemma 1 with $Z := \frac{N_1(T)}{T}$, one obtains

$$\sum_{a=1}^{K} \mathbb{E}_{\nu}(N_a(T)) \cdot d\left(\lambda_a^{\nu}, \lambda_a^{\nu'}\right) \geq d\Big(\frac{\mathbb{E}_{\nu}(N_1(T))}{T}, \frac{\mathbb{E}_{\nu'}(N_1(T))}{T}\Big) \tag{3}$$

Using the inequality $d(p, q) \geq p\log(1/q) - \log(2)$ (see Garivier et al. (2016)) yields

$$d\Big(\frac{\mathbb{E}_{\boldsymbol{\nu}}(N_1(T))}{T}, \frac{\mathbb{E}_{\boldsymbol{\nu'}}(N_1(T))}{T}\Big) \geq \frac{\mathbb{E}_{\boldsymbol{\nu}}(N_1(T))}{T} \log\Big(\frac{T}{\mathbb{E}_{\boldsymbol{\nu'}}(N_1(T))}\Big) - \log(2)$$

Since $a_*(\boldsymbol{\nu}) = 1$, and $a_*(\boldsymbol{\nu'}) \neq 1$, $\mathbb{E}_{\boldsymbol{\nu}}(N_1(T)) \sim T$ and $\mathbb{E}_{\boldsymbol{\nu'}}(N_1(T)) = o(T^{\alpha})$ for all $\alpha \in]0, 1]$. Hence one can show that, $\frac{\mathbb{E}_{\boldsymbol{\nu}}(N_1(T))}{T} \sim 1$ and $\log\Big(\frac{T}{\mathbb{E}_{\nu'}[N_1(T)]}\Big) \sim \log(T)$. Equation (3)

yields

$$\liminf_{T\to\infty} \frac{\sum_{a=1}^{K} \mathbb{E}_{\boldsymbol{\nu}}(N_a(T)) \cdot d\left(\lambda_a^{\boldsymbol{\nu}}, \lambda_a^{\boldsymbol{\nu}'}\right)}{\log T} \geq 1. \tag{4}$$

To obtain a lower bound on $\mathbb{E}_{\boldsymbol{\nu}}[N_a(T)]$ for each $a \in \{2, \ldots, K\}$, one can choose $\boldsymbol{\nu}'$ such that, for some $\epsilon > 0$,

$$\mu_b^{\boldsymbol{\nu}'} = \begin{cases} \mu_1^{\boldsymbol{\nu}} + \epsilon, & \text{if } b = a \\ \mu_b^{\boldsymbol{\nu}} & \text{otherwise} \end{cases}$$

This translates to the following change in feedback, $\lambda_b^{\boldsymbol{\nu}'} = \begin{cases} g_b(\mu_1^{\boldsymbol{\nu}} + \epsilon) & \text{if } b = a, \\ g_b(\mu_b^{\boldsymbol{\nu}}) = \lambda_b^{\boldsymbol{\nu}} & \text{otherwise.} \end{cases}$

As $d\left(\lambda_b^{\boldsymbol{\nu}}, \lambda_b^{\boldsymbol{\nu}'}\right) = 0$ for $b \neq a$, using equation (4) we get

$$\liminf_{T\to\infty} \frac{\mathbb{E}_{\boldsymbol{\nu}}(N_a(T))}{\log T} \geq \frac{1}{d\left(\lambda_a^{\boldsymbol{\nu}}, g_a(\mu_1 + \epsilon)\right)}$$

Letting ϵ go to zero for each $a \in \{2, \ldots, K\}$ (and assuming $\{g_a\}_{a\in A}$ are continuous), one obtains, $\liminf_{T\to\infty} \frac{\text{Regret}_T(\boldsymbol{\nu})}{\log(T)} \geq \sum_{a=2}^{K} \frac{\Delta_a^{\boldsymbol{\nu}}}{d\left(\lambda_a^{\boldsymbol{\nu}}, g_a(\mu_1^{\boldsymbol{\nu}})\right)}$.

7.2. Proof outline for Theorem 2

We defer the complete proof of Theorem 2 to Appendix I. In this subsection, we describe the road-map for the proof. We arrive at a upper bound on the regret of kl-UCB-CF by first bounding the number of times any suboptimal arm a is pulled by the algorithm till horizon T, $\mathbb{E}[N_a(T)]$. Recall that, at any time kl-UCB-CF pulls an arm maximizing a index defined as

$$\text{Index}_a(t) := \max\left\{q : \ N_a(t) \cdot d(\hat{\lambda}_a(t), g_a(q)) \leq f(t)\right\} = \max g_a^{-1}\left(\{q : N_a(t) \cdot d(\hat{\lambda}_a(t), q) \leq f(t)\}\right)$$

For the purpose of this proof, we further decompose the index computation as follows:

$$\text{Index}_a(t) := \begin{cases} g_a^{-1}(\ell_a(t)) & \text{with } \ell_a(t) := \min\{q : N_a(t) \cdot d(\hat{\lambda}_a(t), q) \leq f(t)\} \quad \text{if } g_a \text{ is decreasing,} \\ g_a^{-1}(u_a(t)) & \text{with } u_a(t) := \max\{q : N_a(t) \cdot d(\hat{\lambda}_a(t), q) \leq f(t)\} \quad \text{if } g_a \text{ is increasing,} \end{cases}$$

The interval $[\ell_a(t), u_a(t)]$ is a KL-based confidence interval on the mean feedback λ_a of arm a. This is in contrast to kl-UCB (Cappé et al. (2013)) where a confidence interval is placed on the mean reward (refer Figure 3 in Appendix I for a depiction).

In our analysis, we use the fact that when arm a is picked at time $t+1$ by kl-UCB-CF, one of the following is true. Either the mean feedback of the optimal arm 1 is outside its confidence interval (i.e. $g_1(\mu_1) < \ell_1(t)$ or $g_1(\mu_1) > u_1(t)$), which is unlikely, or the mean feedback of the optimal arm is where it should be, and then the fact that arm a is selected indicates that the confidence interval on λ_a cannot be too small as either $(u_a(t) \geq g_a(\mu_1))$ or $(\ell_a(t) \leq g_a(\mu_1))$. We then need to control the two terms in the decomposition of the expected number of draws of arm a. The term regarding the "unlikely" event, is easily bounded using the same technique as in the kl-UCB analysis, and is of order $o(\log(T))$. To control the second term, depending on the monotonicity of the corruption functions g_a and g_1, we need to adapt the arguments in Cappé et al. (2013) to control the number of draws of arm a, as can be seen in Appendix I.

7.3. Proof outline for Theorem 3

Our proof follows the analysis of Agrawal and Goyal (2013) for classical Thompson Sampling. We proceed by controlling the number of draws of each suboptimal arm a. For this purpose, we introduce two thresholds u_a and w_a that satisfy $\lambda_a < u_a < w_a < g_a(\mu_1)$ if g_a is increasing and $\lambda_a > u_a > w_a > g_a(\mu_1)$ if g_a is decreasing. We introduce $E_a^\lambda(t)$ as the event $\{g_a^{-1}(\hat{\lambda}_a(t)) \leq g_a^{-1}(u_a)\}$ and $E_a^\theta(t)$ as the event $\{g_a^{-1}(\theta_a(t)) \leq g_a^{-1}(w_a)\}$. We then upper bound $\mathbb{E}[N_a(T)]$ by the sum of the three terms as,

$$\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), \overline{E_a^\theta(t)}) + \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), E_a^\theta(t)) + \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \overline{E_a^\lambda(t)}).$$

Using arguments similar to Agrawal and Goyal (2013), with some adaptations, we then show that the last two terms are of order $o(\log(T))$. To control the first term, we prove the following, which requires some extra technicalities compared to the original proof, as shall be seen in Appendix II, where the full proof of Theorem 3 is given.

Lemma 2 *When g_a is increasing (resp. decreasing), for any $u'_a \in (u_a, w_a)$ (resp. (w_a, u_a)),*

$$\sum_{t=0}^{T-1} \mathbb{P}\Big(\hat{a}_{t+1} = a, \overline{E_a^\theta(t)}, E_a^\lambda(t)\Big) \leq \frac{\log(T)}{d(u'_a, w_a)} + 1 \quad \textit{when } T \textit{ is large enough.}$$

8. Conclusion

Both the algorithms introduced in this article, kl-UCB-CF and TS-CF provide suitable solutions to the MAB problem with corrupted feedback, as they are proved to asymptotically attain the best possible (problem-dependent) regret. Our experiments confirm the theoretical analysis by demonstrating the superior performance of kl-UCB-CF and TS-CF. Furthermore, we exhibit appropriate corruption matrices that achieve a desired level of local differential privacy, and quantify their impact on the regret. These algorithms are thus good candidates to be used in recommender systems which apply a randomized response mechanism to protect the user privacy.

This work can be extended in many ways. In our setting, although the feedback is corrupted, it is available at all times. In some situations however, the feedback is simply lost. As future work, we plan to extend our problem setting to incorporate such scenarios by making appropriate changes to the corruption process. An adversarial corruption of the feedback can be considered too. Another possible extension is to incorporate contextual information in the learning process. We conjecture that the invertibility condition on the corruption functions can be relaxed for kl-UCB-CF as long as $\lambda_a \neq g_a(\mu_1)$ for all suboptimal arms but it remains to be proven.

Acknowledgments

This work has been supported by the Austrian Science Fund (FWF): I 3437-N33 in the framework of the CHIST-ERA ERA-NET (DELTA project) and the French L'Agence Nationale de la Recherche (ANR), under grant ANR-16-CE40-0002, project BADASS (BAnDits Against non-Stationarity and Structure).

References

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, pages 99–107, 2013. URL http://jmlr.org/proceedings/papers/v31/agrawal13a.html.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352.

Gábor Bartók, Dean P. Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring - classification, regret bounds, and algorithms. *Math. Oper. Res.*, 39(4):967–997, 2014. doi: 10.1287/moor.2014.0663. URL http://dx.doi.org/10.1287/moor.2014.0663.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024. URL http://dx.doi.org/10.1561/2200000024.

O. Cappé, A. Garivier, O-A. Maillard, R. Munos, and G. Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Privacy aware learning. *J. ACM*, 61(6):38:1–38:57, December 2014. ISSN 0004-5411. doi: 10.1145/2666468. URL http://doi.acm.org/10.1145/2666468.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9:211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://dx.doi.org/10.1561/0400000042.

Cynthia Dwork, Frank Mcsherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *In Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz. Explore First, Exploit Next: The True Shape of Regret in Bandit Problems. working paper or preprint, June 2016. URL https://hal.archives-ouvertes.fr/hal-01276324.

Prateek Jain, Pravesh Kothari, and Abhradeep Thakurta. Differentially private online learning. In *COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland*, pages 24.1–24.34, 2012.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2879–2887. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5392-extremal-mechanisms-for-local-differential-privacy.pdf.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. *Journal of Machine Learning Research*, 17(17):1–51, 2016. URL http://jmlr.org/papers/v17/15-135.html.

Aleksandra Korolova. Privacy violations using microtargeted ads: A case study. In *ICDMW 2010, The 10th IEEE International Conference on Data Mining Workshops, Sydney, Australia, 13 December 2010*, pages 474–482, 2010. doi: 10.1109/ICDMW.2010.137. URL http://dx.doi.org/10.1109/ICDMW.2010.137.

Michal Kosinski, David Stillwell, and Thore Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110 (15):5802–5805, 2013.

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

Nikita Mishra and Abhradeep Thakurta. (nearly) optimal differentially private stochastic multi-arm bandits. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*, pages 592–601, 2015.

Antonio Piccolboni and Christian Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *COLT/EuroCOLT*, volume 2111 of *LNCS*, pages 208–223. Springer, 2001.

Abhradeep Guha Thakurta and Adam D. Smith. (nearly) optimal algorithms for private online learning in full-information and bandit settings. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2733–2741, 2013.

W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4):285–294, 1933.

Aristide C. Y. Tossou and Christos Dimitrakakis. Algorithms for differentially private multi-armed bandits. In *13th International Conference on Artificial Intelligence (AAAI 2016)*, 2016.

Aristide C. Y. Tossou and Christos Dimitrakakis. Achieving privacy in the adversarial multi-armed bandit. In *14th International Conference on Artificial Intelligence (AAAI 2017)*, 2017. URL https://arxiv.org/abs/1701.04222.

Yue Wang, Xintao Wu, and Donghui Hu. Using randomized response for differential privacy preserving data collection. In *Proceedings of the Workshops of the EDBT/ICDT 2016 Joint Conference, EDBT/ICDT Workshops 2016, Bordeaux, France, March 15, 2016.*, 2016. URL http://ceur-ws.org/Vol-1558/paper35.pdf.

Stanley L. Warner. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309):63+, March 1965. URL http://dx.doi.org/10.2307/2283137.

Bangzuo Zhang and Wanli Zuo. Learning from Positive and Unlabeled Examples: A Survey. In *2008 International Symposiums on Information Processing*, volume 0, pages 650–654, May 2008. URL http://dx.doi.org/10.1109/isip.2008.79.

Notations. For the proofs, we recall that $\hat{\lambda}_a(t)$ is the empirical mean of the feedback obtained from arm a until time t. Letting $F_{a,s}$ being the successive feedbacks of arm a and $\hat{\lambda}_{a,s} := \frac{1}{s}\sum_{\ell=1}^{s} F_{a,\ell}$, one has $\hat{\lambda}_a(t) = \hat{\lambda}_{a,N_a(t)}$ when $N_a(t) > 0$.

Appendix I. Proof of Theorem 2

Proof. The index is defined by

$$\text{Index}_a(t) := \max\left\{q : \ N_a(t)\cdot d(\hat{\lambda}_a(t), g_a(q)) \le f(t)\right\} = \max g_a^{-1}\left(\{q : N_a(t)\cdot d(\hat{\lambda}_a(t), q) \le f(t)\}\right)$$

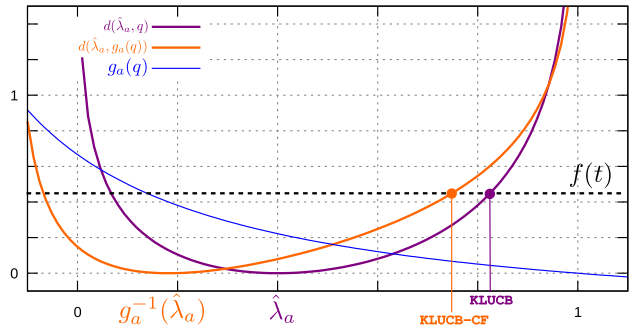


Figure 3: KL indices calculation.

For the purpose of this proof, we further decompose the computation of index as follows,

$$\text{Index}_a(t) := \begin{cases} g_a^{-1}(\ell_a(t)) & \text{if } g_a \text{ is decreasing,} \\ g_a^{-1}(u_a(t)) & \text{if } g_a \text{ is increasing} \end{cases}$$

where,

$$\ell_a(t) := \min\{q : N_a(t)\cdot d(\hat{\lambda}_a(t), q) \le f(t)\} \text{ and } u_a(t) := \max\{q : N_a(t)\cdot d(\hat{\lambda}_a(t), q) \le f(t)\}$$

To get an upper bound on the regret of this algorithm, we first bound $\mathbb{E}[N_a(t)]$ for all the non-optimal arms a. Note that, we assume 1 to be the optimal arm.

$$\mathbb{E}(N_a(T)) = 1 + \sum_{t=K}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a)$$

Depending upon if g_a and g_1 are increasing or decreasing there are four possible sub-cases:

- Both g_1 and g_a are increasing.

$$\begin{aligned} &(\hat{a}_{t+1} = a) \\ &\subseteq (u_1(t) < g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, u_1(t) \ge g_1(\mu_1)) \\ &= (u_1(t) < g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, g_1^{-1}(u_1(t)) \ge \mu_1) && \text{since } g_1 \text{ is increasing} \\ &= (u_1(t) < g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, g_a^{-1}(u_a(t)) \ge \mu_1) && \text{since Index}_a > \text{Index}_1 \end{aligned}$$

$$= (u_1(t) < g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, u_a(t) \geq g_a(\mu_1)) \qquad \text{since } g_a \text{ is increasing}$$

$$\therefore \mathbb{E}(N_a(T)) \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(u_1(t) < g_1(\mu_1)) + \sum_{t=K}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, u_a(t) \geq g_a(\mu_1)) \qquad (5)$$

- g_1 is decreasing and g_a is increasing.

$$\begin{aligned}
&(\hat{a}_{t+1} = a) \\
&\subseteq (\ell_1(t) > g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, \ell_1(t) \leq g_1(\mu_1)) \\
&= (\ell_1(t) > g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, g_1^{-1}(\ell_1(t)) \geq \mu_1) \qquad \text{since } g_1 \text{ is decreasing} \\
&= (\ell_1(t) > g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, g_a^{-1}(u_a(t)) \geq \mu_1) \qquad \text{since } \text{Index}_a > \text{Index}_1 \\
&= (\ell_1(t) > g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, u_a(t) \geq g_a(\mu_1)) \qquad \text{since } g_a \text{ is increasing}
\end{aligned}$$

$$\therefore \mathbb{E}(N_a(T)) \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(\ell_1(t) > g_1(\mu_1)) + \sum_{t=K}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, u_a(t) \geq g_a(\mu_1)) \qquad (6)$$

- g_1 is increasing and g_a is decreasing.

$$\begin{aligned}
&(\hat{a}_{t+1} = a) \\
&\subseteq (u_1(t) < g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, u_1(t) \geq g_1(\mu_1)) \\
&= (u_1(t) < g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, g_1^{-1}(u_1(t)) \geq \mu_1) \qquad \text{since } g_1 \text{ is increasing} \\
&= (u_1(t) < g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, g_a^{-1}(\ell_a(t)) \geq \mu_1) \qquad \text{since } \text{Index}_a > \text{Index}_1 \\
&= (u_1(t) < g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, \ell_a(t) \leq g_a(\mu_1)) \qquad \text{since } g_a \text{ is decreasing}
\end{aligned}$$

$$\therefore \mathbb{E}(N_a(T)) \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(u_1(t) < g_1(\mu_1)) + \sum_{t=K}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \ell_a(t) \leq g_a(\mu_1)) \qquad (7)$$

- g_1 is decreasing and g_a is decreasing.

$$\begin{aligned}
&(\hat{a}_{t+1} = a) \\
&\subseteq (\ell_1(t) > g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, \ell_1(t) \leq g_1(\mu_1)) \\
&= (\ell_1(t) > g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, g_1^{-1}(\ell_1(t)) \geq \mu_1) \qquad \text{since } g_1 \text{ is decreasing} \\
&= (\ell_1(t) > g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, g_a^{-1}(\ell_a(t)) \geq \mu_1) \qquad \text{since } \text{Index}_a > \text{Index}_1 \\
&= (\ell_1(t) > g_1(\mu_1)) \cup (\hat{a}_{t+1} = a, \ell_a(t) \leq g_a(\mu_1)) \qquad \text{since } g_a \text{ is decreasing}
\end{aligned}$$

$$\therefore \mathbb{E}(N_a(T)) \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(\ell_1(t) > g_1(\mu_1)) + \sum_{t=K}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \ell_a(t) \leq g_a(\mu_1)) \qquad (8)$$

We first upper bound the two sums

$$\sum_{t=K}^{T-1} \mathbb{P}(u_1(t) < g_1(\mu_1)) \quad \text{and} \quad \sum_{t=K}^{T-1} \mathbb{P}(\ell_1(t) > g_1(\mu_1)) \qquad (9)$$

using that $\ell_1(t)$ and $u_1(t)$ are respectively lower and upper confidence bound on $g_1(\mu_1)$. Indeed,

$$\begin{aligned}\mathbb{P}(u_1(t) < g_1(\mu_1)) &\leq \mathbb{P}\Big(g_1(\mu_1) > \hat{\lambda}_1(t) \text{ and } N_1(t)d(\hat{\lambda}_1(t), g_1(\mu_1)) \geq f(t)\Big) \\ &\leq \mathbb{P}\Big(\exists s \in \{1,\dots,t\} : g_1(\mu_1) > \hat{\lambda}_{1,s} \text{ and } sd(\hat{\lambda}_{1,s}, g_1(\mu_1)) \geq f(t)\Big) \\ &\leq min\{1, e\lceil f(t)\log t\rceil e^{-f(t)}\},\end{aligned}$$

where the upper bound follows from Lemma 2 in Cappé et al. (2013), and the fact that $\hat{\lambda}_{1,s}$ is the empirical mean of s Bernoulli samples with mean $g_1(\mu_1)$. Similarly, one has

$$\mathbb{P}(\ell_1(t) > g_1(\mu_1)) \leq \min\{1, e\lceil f(t)\log t\rceil e^{-f(t)}\}.$$

As $f(t) \coloneqq \log t + 3(\log\log t)$ for $t \geq 3$,

$$e\lceil f(t)\log t\rceil \leq 4e\log^2 t,$$

the two quantities in (9) can be upper bounded by

$$\begin{aligned}1 + \sum_{t=3}^{T-1} e\lceil f(t)\log t\rceil e^{-f(t)} &\leq 1 + \sum_{t=3}^{T-1} 4e\cdot\log^2 t\cdot e^{-f(t)} \\ &= 1 + 4e\sum_{t=3}^{T-1}\frac{1}{t\log t} \\ &\leq 4e\Big(\frac{1}{3\log 3} + \int_3^{T-1}\frac{1}{t\log t}dt\Big) \\ &\leq 4e\Big(\frac{1}{3\log 3} + \log(\log(T-1)) - \log(\log 3)\Big) \\ &\leq 3 + 4e\log(\log T).\end{aligned}$$

This proves that

$$\sum_{t=K}^{T-1}\mathbb{P}(u_1(t) < g_1(\mu_1)) \leq 3 + 4e\log(\log T) \in o(\log T) \tag{10}$$

$$\sum_{t=K}^{T-1}\mathbb{P}(\ell_1(t) > g_1(\mu_1)) \leq 3 + 4e\log(\log T) \in o(\log T) \tag{11}$$

We now turn our attention to the other two sums involved in the upper bound we gave for $\mathbb{E}(N_a(t))$. We introduce the notation $d^+(x,y) = d(x,y)\cdot\mathbb{1}_{(x<y)}$ and $d^-(x,y) = d(x,y)\cdot\mathbb{1}_{(x>y)}$. So we can write, when g_a is increasing,

$$\begin{aligned}&\sum_{t=K}^{T-1}\mathbb{P}(\hat{a}_{t+1} = a, u_a(t) \geq g_a(\mu_1)) \\ &= \mathbb{E}\left[\sum_{t=K}^{T-1}\mathbb{1}_{\hat{a}_{t+1}=a}\cdot\mathbb{1}_{N_a(t)\cdot d^+(\hat{\lambda}_{i,N_a(t)}, g_a(\mu_1))\leq f(t)}\right]\end{aligned}$$

$$\leq \mathbb{E}\left[\sum_{t=K}^{T-1}\sum_{s=1}^{t} \mathbb{1}_{\hat{a}_{t+1}=a} \cdot \mathbb{1}_{N_a(t)=s} \cdot \mathbb{1}_{s\cdot d^+(\hat{\lambda}_{a,s},g_a(\mu_1))\leq f(T)}\right]$$
$$= \mathbb{E}\Big[\sum_{s=1}^{T-1} \mathbb{1}_{s\cdot d^+(\hat{\lambda}_{a,s},g_a(\mu_1))\leq f(T)} \underbrace{\sum_{s=1}^{T-1} \mathbb{1}_{\hat{a}_{t+1}=a} \cdot \mathbb{1}_{N_a(t)=s}}_{\leq 1}\Big].$$

One obtains, when g_a is increasing,

$$\sum_{t=K}^{T-1} \mathbb{P}(\hat{a}_{t+1}=a, u_a(t) \geq g_a(\mu_1)) \leq \sum_{s=1}^{T-1} \mathbb{P}\left(s \cdot d^+(\hat{\lambda}_{a,s}, g_a(\mu_1)) \leq f(T)\right). \tag{12}$$

Using similar arguments, one can show that when g_a is decreasing,

$$\sum_{t=K}^{T-1} \mathbb{P}(\hat{a}_{t+1}=a, \ell_a(t) \leq g_a(\mu_1)) \leq \sum_{s=1}^{T-1} \mathbb{P}\left(s \cdot d^-(\hat{\lambda}_{a,s}, g_a(\mu_1)) \leq f(T)\right). \tag{13}$$

The quantity in the right-hand side of (12) is upper bounded in Appendix A.2. of Cappé et al. (2013) by

$$\frac{f(T)}{d(\lambda_a, g_a(\mu_1)} + \sqrt{2\pi}\sqrt{\frac{(d'(\lambda_a, g_a(\mu_1)))^2}{(d(\lambda_a, g_a(\mu_1)))^3}}\sqrt{f(T)} + 2\left(\frac{d'(\lambda_a, g_a(\mu_1))}{d(\lambda_a, g_a(\mu_1))}\right)^2 + 1. \tag{14}$$

For the second term, noting that $d^-(x,y) = d^+(1-x, 1-y)$, one has

$$\begin{aligned}\mathbb{P}\left(s \cdot d^-(\hat{\lambda}_{a,s}, g_a(\mu_1)) \leq f(T)\right) &= \mathbb{P}\left(s \cdot d^+(1-\hat{\lambda}_{a,s}, 1-g_a(\mu_1)) \leq f(T)\right) \\ &= \mathbb{P}\left(s \cdot d^+(\hat{\mu}_{a,s}, 1-g_a(\mu_1)) \leq f(T)\right),\end{aligned}$$

where $\hat{\mu}_{a,s} \coloneqq 1-\hat{\lambda}_{a,s}$, is the empirical mean of s observations of a Bernoulli random variable with mean $1-\lambda_a < 1-g_a(\mu_1)$. Hence, the analysis of Cappé et al. (2013) can be applied, and using that $d(1-\lambda_a, 1-g_a(\mu_1)) = d(\lambda_a, g_a(\mu_1))$ and $d'(1-\lambda_a, 1-g_a(\mu_1)) = -d'(\lambda_a, g_a(\mu_1))$, the left hand side of (13) can also be upper bound by (14).

Combining inequalities (10), (11) and (12),(13), (14) with the initial decomposition of $\mathbb{E}[N_a(T)]$ yields in all cases,

$$\begin{aligned}\mathbb{E}[N_a(T)] \leq\ & \frac{\log(T)}{d(\lambda_a, g_a(\mu_1))} + \sqrt{2\pi}\sqrt{\frac{d'(\lambda_a, g_a(\mu_1))^2}{d(\lambda_a, g_a(\mu_1))^3}}\sqrt{\log(T) + 3\log\log(T)} \\ &+ \left(4e + \frac{3}{d(\lambda_a, g_a(\mu_1)}\right)\log\log(T) + 2\left(\frac{d'(\lambda_a, g_a(\mu_1)}{d(\lambda_a, g_a(\mu_1)}\right)^2 + 5.\end{aligned}$$

Hence the regret of kl-UCB-CF is upper bounded by

$$\sum_{a=2}^{K} \Delta_a\Big[\frac{\log(T)}{D_a} + \sqrt{2\pi}\sqrt{\frac{(D'_a)^2}{D_a^3}}\sqrt{\log(T) + 3\log\log(T)} + \left(4e + \frac{3}{D_a}\right)\log\log(T) + 2\left(\frac{D'_a}{D_a}\right)^2 + 5\Big]$$

where $D_a \coloneqq d(\lambda_a, g_a(\mu_1))$ and $D'_a \coloneqq d'(\lambda_a, g_a(\mu_1))$, which concludes the proof.

Appendix II. Proof of Theorem 3

Proof. Assume 1 to be the optimal arm. For each arm non-optimal arm a, choose two thresholds u_a and w_a such that $\lambda_a < u_a < w_a < g_a(\mu_1)$ if g_a is increasing and $\lambda_a > u_a > w_a > g_a(\mu_1)$ if g_a is decreasing. Define $E_a^\lambda(t)$ as the event $\{g_a^{-1}(\hat{\lambda}_a(t)) \leq g_a^{-1}(u_a)\}$ and $E_a^\theta(t)$ as the event $\{g_a^{-1}(\theta_a(t)) \leq g_a^{-1}(w_a)\}$. Define $\mathcal{F}_t$ as the history of arm selections and received feedbacks including time t and recall that TS-CF selects the arm as follows,

$$\hat{a}_{t+1} = \arg\max_a \ \theta_a(t)$$

, where $\theta_a(t)$ is a sample from the posterior distribution on arm a after t observations. Define $p_{a,t} := \mathbb{P}(g_1^{-1}(\theta_1(t)) > g_a^{-1}(w_a) \mid \mathcal{F}_t)$.

We start from the following decomposition.

$$\begin{aligned}\mathbb{E}[N_a(T)] = &\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), E_a^\theta(t)) + \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), \overline{E_a^\theta(t)}) \\ &+ \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \overline{E_a^\lambda(t)})\end{aligned}$$

Below are the lemmas that permit us to bound these three terms. These results generalize to the corrupted setting the main steps of the analysis of Thompson Sampling by Agrawal and Goyal (2013). The proofs for these lemmas follow that of the corresponding lemmas in the aforementioned article, with some technicalities that arise from the fact that g_1 and g_a may be either increasing or decreasing.

Lemma 3 $\mathbb{P}(\hat{a}_{t+1} = a, E_a^\theta(t), E_a^\lambda(t) \mid \mathcal{F}_t) \leq \frac{(1-p_{a,t})}{p_{a,t}} \mathbb{P}(\hat{a}_{t+1} = 1, E_a^\theta(t), E_a^\lambda(t) \mid \mathcal{F}_t)$

Proof. Assume that $E_a^\lambda(t)$ is true (otherwise the lemma holds trivially because the left hand size is 0). Hence, it is sufficient to prove that,

$$\mathbb{P}(\hat{a}_{t+1} = a \mid E_a^\theta(t), \mathcal{F}_t) \leq \frac{(1-p_{a,t})}{p_{a,t}} \mathbb{P}(\hat{a}_{t+1} = 1 \mid E_a^\theta(t), \mathcal{F}_t)) \tag{15}$$

Define $M_a(t)$ the event in which the index of arm a at time t is the largest among those of all suboptimal arms: $M_a(t) := \left\{ g_a^{-1}(\theta_a(t)) \geq g_j^{-1}(\theta_j(t)), \forall j \neq 1 \right\}$.

$$\begin{aligned}&\mathbb{P}(\hat{a}_{t+1} = 1 \mid E_a^\theta(t), \mathcal{F}_t)) \\ &\geq \mathbb{P}(\hat{a}_{t+1} = 1, M_a(t) \mid E_a^\theta(t), \mathcal{F}_t)) \\ &= \mathbb{P}(M_a(t) \mid E_a^\theta(t), \mathcal{F}_t)) \cdot \mathbb{P}(\hat{a}_{t+1} = 1 \mid M_a(t), E_a^\theta(t), \mathcal{F}_t))\end{aligned} \tag{16}$$

Now, given $M_a(t)$ and $E_a^\theta(t)$ hold,

$$g_j^{-1}(\theta_j(t)) \leq g_a^{-1}(\theta_a(t)) \leq g_a^{-1}(w_a) \quad \forall j \neq a, j \neq 1$$

So,

$$\begin{aligned}\mathbb{P}(\hat{a}_{t+1}=1 \mid M_a(t), E_a^\theta(t), \mathcal{F}_t) &\geq \mathbb{P}(g_1^{-1}(\theta_1(t)) > g_a^{-1}(w_a) \mid M_a(t), E_a^\theta(t), \mathcal{F}_t) \\ &= \mathbb{P}(g_1^{-1}(\theta_1(t)) > g_a^{-1}(w_a) \mid \mathcal{F}_t) \\ &= p_{a,t} \end{aligned} \tag{17}$$

From inequalities (16) and (17),

$$\mathbb{P}(\hat{a}_{t+1}=1 \mid E_a^\theta(t), \mathcal{F}_t) \geq p_{a,t} \cdot \mathbb{P}(M_a(t) \mid E_a^\theta(t), \mathcal{F}_t) \tag{18}$$

Now, let's consider the left hand side of the inequality (15). The fact that $E_a^\theta(t)$ holds and $\hat{a}_{t+1}=a$ implies that $g_1^{-1}(\theta_1(t)) < g_a^{-1}(\theta_a(t)) < g_a^{-1}(w_a)$. Hence

$$\begin{aligned}&\mathbb{P}(\hat{a}_{t+1}=a \mid E_a^\theta(t), \mathcal{F}_t) \\ &\leq \mathbb{P}\Big(g_1^{-1}(\theta_1(t)) \leq g_a^{-1}(w_a), g_a^{-1}(\theta_a(t)) \geq g_j^{-1}(\theta_j(t)), \forall j \neq 1 \mid E_a^\theta(t), \mathcal{F}_t\Big) \\ &= \mathbb{P}\Big(g_1^{-1}(\theta_1(t)) \leq g_a^{-1}(w_a) \mid \mathcal{F}_{t-1}\Big) \cdot \mathbb{P}\Big(g_a^{-1}(\theta_a(t)) \geq g_j^{-1}(\theta_j(t)), \forall j \neq 1 \mid E_a^\theta(t), \mathcal{F}_t\Big) \\ &= (1-p_{a,t}) \cdot \mathbb{P}(M_a(t) \mid E_a^\theta(t), \mathcal{F}_t) \end{aligned} \tag{19}$$

From inequalities (18) and (19),

$$\mathbb{P}(\hat{a}_{t+1}=a \mid E_a^\theta(t), \mathcal{F}_t) \leq \frac{(1-p_{a,t})}{p_{a,t}} \mathbb{P}(\hat{a}_{t+1}=1 \mid E_a^\theta(t), \mathcal{F}_t)$$

Lemma 4 *When g_a is increasing (resp. decreasing), for any $x'_a \in]x_a, y_a[$ (resp. $]y_a, x_a[$), when T is large enough,*

$$\sum_{t=0}^{T-1} \mathbb{P}\Big(\hat{a}_{t+1}=a, \overline{E_a^\theta(t)}, E_a^\lambda(t)\Big) \leq \frac{\log(T)}{d(u'_a, w_a)} + 1.$$

Proof. When g_a is increasing, the application of Lemma 3 in Agrawal and Goyal (2013) directly yields

$$\sum_{t=0}^{T-1} \mathbb{P}\left(\hat{a}_{t+1}=a, \overline{E_a^\theta(t)}, E_a^\lambda(t)\right) \leq \frac{\log T}{d(u_a, w_a)} + 1.$$

The proof is based on the use of deviation inequalities and a link between the Beta and Binomial c.d.f. that shall also be useful in the decreasing case, that we handle now (using slightly different arguments).

Fact 1

$$F_{\alpha,\beta}^{beta}(w) = 1 - F_{\alpha+\beta-1,w}^{B}(\alpha-1)$$

Note that for decreasing g_a, one has $\overline{E_a^\theta(t)} = \{\theta_a(t) \leq w_a\}$ and $E_a^\lambda(t) = \{\hat{\lambda}_a(t) > u_a\}$. Fix u'_a such that $w_a < u'_a < u_a$ and let $L'_a(T) = \frac{\log(T)}{d(u'_a, w_a)}$.

$$\sum_{t=0}^{T-1} \mathbb{P}\left(\hat{a}_{t+1}=a, \hat{\lambda}_a(t) > u_a, \theta_a(t) \leq w_a\right)$$

$$
\begin{aligned}
&\leq \frac{\log(T)}{d(u'_a,w_a)} + \sum_{t=0}^{T-1} \mathbb{P}\left(\hat{a}_{t+1}=a, N_a(t) \leq L'_a(T), \theta_a(t) \leq w_a, \hat{\lambda}_a(t) > u_a\right) \\
&\leq \frac{\log(T)}{d(u'_a,w_a)} + \mathbb{E}\sum_{t=0}^{T-1}\sum_{s=L'_a(T)}^{t} \mathbb{1}_{(\hat{a}_{t+1}=a, N_a(t)=s, \theta_a(t)\leq w_a, \hat{\lambda}_a(t)>u_a)} \\
&= \frac{\log(T)}{d(u'_a,w_a)} + \mathbb{E}\sum_{t=0}^{T-1}\sum_{s=L'_a(T)}^{t} \mathbb{1}_{\left(\hat{a}_{t+1}=a, N_a(t)=s, \hat{\lambda}_a(t)>u_a\right)} \mathbb{P}\left(\theta_a(t) \leq w_a \mid \mathcal{F}_t\right) \\
&= \frac{\log(T)}{d(u'_a,w_a)} + \mathbb{E}\sum_{t=0}^{T-1}\sum_{s=L'_a(T)}^{t} \mathbb{1}_{\left(\hat{a}_{t+1}=a, N_a(t)=s, \hat{\lambda}_a(t)>u_a\right)} F^{beta}_{(s\hat{\lambda}_a(t)+1, s-s\hat{\lambda}_a(t)+1)}(w_a) \\
&= \frac{\log(T)}{d(u'_a,w_a)} + \mathbb{E}\sum_{t=0}^{T-1}\sum_{s=L'_a(T)}^{t} \mathbb{1}_{\left(\hat{a}_{t+1}=a, N_a(t)=s, \hat{\lambda}_a(t)>u_a\right)} \left(1 - F^{B}_{(s+1,w_a)}(s\hat{\lambda}_a(t))\right) \\
&\leq \frac{\log(T)}{d(u'_a,w_a)} + \mathbb{E}\sum_{t=0}^{T-1}\sum_{s=L'_a(T)}^{t} \mathbb{1}_{\left(\hat{a}_{t+1}=a, N_a(t)=s, \hat{\lambda}_a(t)>u_a\right)} \underbrace{\left(1 - F^{B}_{(s+1,w_a)}(su_a)\right)}_{A_s}
\end{aligned}
$$

Introducing (X_k) an i.i.d. sequence drawn from Bernoulli of mean w_a, term A_s can be written, for any s, .

$$
\begin{aligned}
A_s &= \mathbb{P}\left(\sum_{k=1}^{s+1} X_k \geq u_a s\right) \leq \mathbb{P}\left(\sum_{k=1}^{s} X_k \geq u_a s - 1\right) = \mathbb{P}\left(\frac{1}{s}\sum_{k=1}^{s} X_k \geq u_a - \frac{1}{s}\right) \\
&\leq \exp\left(-sd\left(u_a - 1/s, w_a\right)\right) \leq \exp\left(-\log(T)\frac{d(u_a - 1/s, w_a)}{d(u'_a, w_a)}\right) \leq \frac{1}{T},
\end{aligned}
$$

for large enough T, and s larger than $L'_a(T)$ (as it holds that $d(u_a - 1/s, w_a) \geq d(u'_a, w_a)$). Finally, for T large enough,

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{P}\left(\hat{a}_t = a, \hat{\lambda}_a(t) \geq u_a, \theta_a(t) \leq w_a\right) \\
&\leq \frac{\log(T)}{d(u'_a,w_a)} + \sum_{s=0}^{T-1} \frac{1}{T} \mathbb{E} \underbrace{\sum_{t=s}^{T} \mathbb{1}_{(\hat{a}_{t+1}=a, N_a(t)=s)}}_{\leq 1} \\
&\leq \frac{\log(T)}{d(u'_a,w_a)} + \sum_{t=1}^{T} \frac{1}{T} = \frac{\log(T)}{d(u'_a,w_a)} + 1.
\end{aligned}
$$

Lemma 5 $\sum_{t=0}^{T-1} \mathbb{P}\left(\hat{a}_{t+1}=a, \overline{E_a^{\lambda}(t)}\right) \leq 1 + \frac{1}{d(u_a,\lambda_a)}$.

Proof. This result follows from the application of Chernoff bound for the concentration of $\hat{\lambda}_a(t)$. When g_a is increasing, it follows directly from the application of Lemma 2 in

Agrawal and Goyal (2013), hence we write the proof in the decreasing case only, where we shall justify that for $u_a < \lambda_a$,

$$\sum_{t=0}^{T-1} \mathbb{P}\left(\hat{a}_{t+1} = a, \hat{\lambda}_a(t) < u_a\right) \leq \frac{1}{d(u_a, \lambda_a)} + 1.$$

Using $\hat{\lambda}_{a,s}$ to denote the empirical mean of the s first observations from the feedback of arm a,

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbb{P}\left(\hat{a}_{t+1} = a, \hat{\lambda}_a(t) < u_a\right) &= \mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{s=0}^{t} \mathbb{1}_{(\hat{a}_{t+1}=a, N_a(t)=s)} \mathbb{1}_{(\hat{\lambda}_{a,s}<u_a)}\right] \\
&= \mathbb{E}\Big[\sum_{s=0}^{T} \mathbb{1}_{(\hat{\lambda}_{a,s}<u_a)} \underbrace{\sum_{t=s}^{T} \mathbb{1}_{(\hat{a}_{t+1}=a, N_a(t)=s)}}_{\leq 1}\Big] \\
&\leq 1 + \sum_{s=1}^{T-1} \mathbb{P}\left(\hat{\lambda}_{a,s} < u_a\right) \leq 1 + \sum_{s=1}^{T-1} \exp(-sd(u_a, \lambda_a)) \\
&\leq 1 + \frac{1}{d(u_a, \lambda_a)},
\end{aligned}$$

where the last but one inequality follows from Chernoff inequality (as $u_a < \lambda_a$).

Lemma 6 *Let τ_s be the instant of the s-th play of arm 1. Then there exists a function $f(s) = f(s, \lambda_1, g_1(g_a^{-1}(\mu_1)))$ satisfying $\sum_{s=1}^{\infty} f(s) < \infty$ such that for all s,*

$$\mathbb{E}\left[\frac{1}{p_{a,\tau_s+1}}\right] \leq 1 + f(s).$$

Proof. Let $\tilde{w}_a \coloneqq g_1(g_a^{-1}(w_a))$. Examining all possibilities, one can easily show that

- if g_1 is increasing and g_a is increasing, $p_{a,t} = \mathbb{P}\left(\theta_1(t) > \tilde{w}_a\right)$, with $\tilde{w}_a < \lambda_1$,
- if g_1 is increasing and g_a decreasing, $p_{a,t} = \mathbb{P}\left(\theta_1(t) > \tilde{w}_a\right)$, with $\tilde{w}_a < \lambda_1$,
- if g_1 is decreasing and g_a is increasing, $p_{a,t} = \mathbb{P}\left(\theta_1(t) < \tilde{w}_a\right)$, with $\tilde{w}_a > \lambda_1$,
- if g_1 is decreasing and g_a is decreasing, $p_{a,t} = \mathbb{P}\left(\theta_1(t) < \tilde{w}_a\right)$, with $\tilde{w}_a > \lambda_1$.

When g_1 is increasing, $\tilde{w}_a < \lambda_1$ and

$$p_{a,\tau_s+1} = 1 - F^{beta}_{(S_1(\tau_s)+1, s-S_1(\tau_s)+1)}(\tilde{w}_a) = F^{B}_{(s+1,\tilde{w}_a)}(S_1(\tau_s)).$$

Using that $S_1(\tau_s)$ has a binomial distribution with parameters (s, λ_1) yields

$$\mathbb{E}\left[\frac{1}{p_{a,\tau_s+1}}\right] = \sum_{j=0}^{s} \frac{f^{B}_{(s,\lambda_1)}(j)}{F^{B}_{(s+1,\tilde{w}_a)}(j)}. \tag{20}$$

When g_1 is decreasing, recall $\tilde{w}_a > \lambda_1$ and one has

$$p_{a,\tau_s+1} = F^{beta}_{(S_1(\tau_s)+1,s-S_1(\tau_s)+1)}(\tilde{w}_a) = 1 - F^B_{(s+1,\tilde{w}_a)}(S_1(\tau_s)).$$

Using again the distribution of $S_1(\tau_s)$ yields

$$\mathbb{E}\left[\frac{1}{p_{a,\tau_s+1}}\right] = \sum_{j=0}^{s} \frac{f^B_{(s,\lambda_1)}(j)}{1 - F^B_{(s+1,\tilde{w}_a)}(j)}$$

Note here two simple properties of Binomial distributions: for all $t \in \mathbb{N}^*$ and $c \in [0,1]$, for all $j \in \{0,\dots,t\}$,

- $f^B_{(t,c)}(j) = f_{(t,1-c)}(s-j)$
- $F^B_{(t,c)}(j) = 1 - F^B_{(t,1-c)}(t-j-1)$

It follows that

$$\mathbb{E}\left[\frac{1}{p_{a,\tau_s+1}}\right] = \sum_{j=0}^{s} \frac{f^B_{(s,1-\lambda_1)}(s-j)}{F^B_{(s+1,1-\tilde{w}_a)}(s-j)} = \sum_{j=0}^{s} \frac{f^B_{(s,1-\lambda_1)}(j)}{F^B_{(s+1,1-\tilde{w}_a)}(j)}, \tag{21}$$

with $1-\lambda_1 > 1-\tilde{w}_a$.

The proof for Lemma 4 given in Agrawal and Goyal (2013) provides an upper bound on the quantity

$$\sum_{j=0}^{s} \frac{f^B_{(s,c)}(j)}{F^B_{(s+1,c)}(j)}$$

whenever c is larger that d. Using this result one can bound (20) and (21) by the same quantity:

$$\mathbb{E}\left[\frac{1}{p_{a,\tau_s+1}}\right] \leq \begin{cases} 1 + \frac{3}{\Delta'_a}, & \text{if } s < \frac{8}{\Delta'_a} \\ 1 + \Theta\left(\exp\left(-{\Delta'_a}^2 s/2\right) + \frac{1}{(s+1){\Delta'_a}^2}\exp\left(-D_a s\right) + \frac{1}{\exp\left({\Delta'_a}^2 s/4\right)-1}\right), & \text{if } s \geq \frac{8}{\Delta'_a} \end{cases}$$

where $\Delta'_a := \lambda_1 - \tilde{w}_a$ and $D_a := \tilde{w}_a \log\frac{\tilde{w}_a}{\lambda_1} + (1-\tilde{w}_a)\log\frac{1-\tilde{w}_a}{1-\lambda_1}$. Hence, Lemma 6 follows with

$$f(s) := \begin{cases} \frac{3}{\Delta'_a}, & \text{if } s < \frac{8}{\Delta'_a} \\ \Theta\left(\exp\left(-{\Delta'_a}^2 s/2\right) + \frac{1}{(s+1){\Delta'_a}^2}\exp\left(-D_a s\right) + \frac{1}{\exp\left({\Delta'_a}^2 s/4\right)-1}\right), & \text{if } s \geq \frac{8}{\Delta'_a} \end{cases},$$

that satisfies $\sum_{s=0}^{\infty} f(s) < \infty$.

One can now complete the proof of Theorem 3.

$$\mathbb{E}[N_a(T)] = \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a)$$

$$
\begin{aligned}
&= \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^{\lambda}(t), E_a^{\theta}(t)) + \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^{\lambda}(t), \overline{E_a^{\theta}(t)}) \\
&+ \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \overline{E_a^{\lambda}(t)}) \\
&\leq \sum_{t=0}^{T-1} \mathbb{E}\left[\frac{(1-p_{a,t})}{p_{a,t}} \mathbb{1}_{(\hat{a}_{t+1}=1, E_a^{\theta}(t), E_a^{\lambda}(t))}\right] + \frac{\log T}{d(u'_a, w_a)} + 1 + \frac{1}{d(u_a, \lambda_a)} + 1 \\
&\leq \sum_{s=0}^{T-1} \mathbb{E}\left[\frac{(1-p_{a,\tau_s+1})}{p_{a,\tau_s+1}} \sum_{t=\tau_s}^{\tau_{s+1}-1} \mathbb{1}_{(\hat{a}_{t+1}=1)}\right] + \frac{\log T}{d(u'_a, w_a)} + 1 + \frac{1}{d(u_a, \lambda_a)} + 1 \\
&= \sum_{s=0}^{T-1} \mathbb{E}\left[\frac{1}{p_{a,\tau_s+1}} - 1\right] + \frac{\log T}{d(u'_a, w_a)} + 1 + \frac{1}{d(u_a, \lambda_a)} + 1 \\
&\leq \frac{\log T}{d(u'_a, w_a)} + \sum_{s=0}^{T-1} f(s) + \frac{1}{d(u_a, \lambda_a)} + 2.
\end{aligned}
$$

Fix $\psi > 0$. Using the monotonicity properties of the divergence function d, there exists $u_a < u'_a < w_a$ in the increasing case and $u_a > u'_a > w_a$ in the decreasing case such that $d(u'_a, w_a) \geq d(\lambda_a, g_a(\mu_1))/(1+\psi)$. For this particular choice, one obtains

$$
\mathbb{E}[N_a(T)] \leq (1+\psi)\frac{\log(T)}{d(\lambda_a, g_a(\mu_a))} + R(u_a, u'_a, w_a),
$$

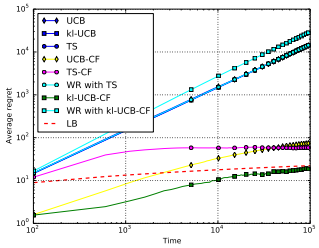
where $R(u_a, u'_a, w_a)$ is a rest term that depends on ψ, μ_1, μ_a, g_1 and g_a. The result follows using that $\text{Regret}_T = \sum_{a=2}^{K} \Delta_a \mathbb{E}[N_a(T)]$.

Appendix III. Additional Empirical Evaluation

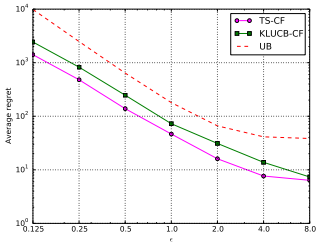
We ran the experiments mentioned in Section 6.1, 6.2 and 6.3 on 4 additional Bernoulli corrupt bandit problems. These problems are succinctly described by the mean rewards of their arms given in Table 1. Recall that in the experiment to compare the performance of the algorithms over a period of time, randomized response was employed to corrupt the feedback with $p_{00} = p_{11} = 0.6$ for the optimal arm, while for all the other arms, both p_{00} and p_{11} were set to 0.9. The time horizon was varied to 10^5 and each experiment was repeated 1000 times. Figures 4a, 5a, 6a and 7a show the average regret of the considered algorithms. In the second experiment aiming to see the effect of various levels of differential privacy on the regret, we chose ϵ from the set $\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$. The corruption parameters are set by substituting the values of ϵ in Equation (2). The horizon was fixed to 10^5 and the experiment was repeated 1000 times. The corresponding curve for the average regret are given in Figures 4b, 5b, 6b and 7b. The third experiment compares the regret of kl-UCB-CF and TS-CF with DP-UCB-Int for $\epsilon = 1$ and its results are given in Figures 4c, 5c, 6c and 7c .

Table 1: Bernoulli mean arm rewards for experimental scenarios

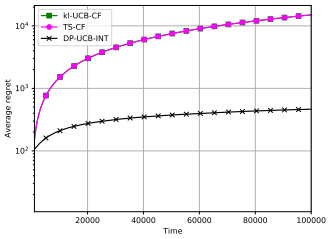
Scenario	Arms									
	1	2	3	4	5	6	7	8	9	10
1	0.9	0.6								
2	0.9	0.8								
3	0.9	0.8	0.8	0.8	0.7	0.7	0.7	0.6	0.6	0.6
4	0.9	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6



(*a*) Regret plots for kl-UCB-CF and TS-CF with others for varying horizons up to 10^5

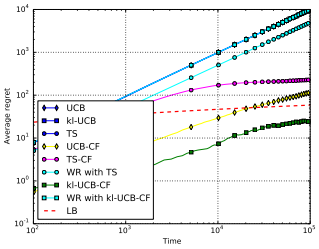


(*b*) Regret plots for kl-UCB-CF and TS-CF with $T = 10^5$, $\epsilon = \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$

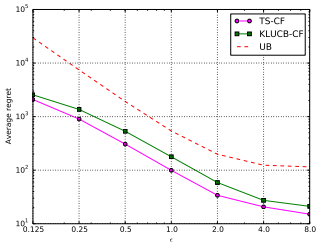


(*c*) Regret plots for $\epsilon = 1$ for kl-UCB-CF and TS-CF , DP-UCB-Int

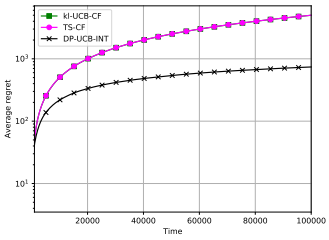
Figure 4: Regret plots for scenario 1



(*a*) Regret plots for kl-UCB-CF and TS-CF with others for varying horizons up to 10^5

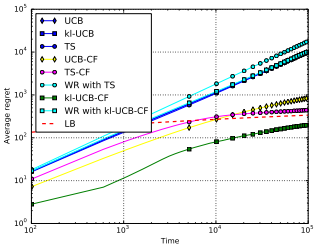


(*b*) Regret plots for kl-UCB-CF and TS-CF with $T = 10^5$, $\epsilon = \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$

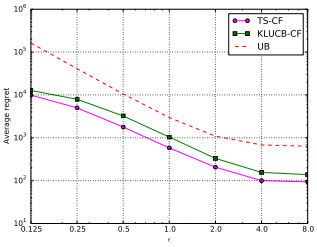


(*c*) Regret plots for $\epsilon = 1$ for kl-UCB-CF and TS-CF , DP-UCB-Int

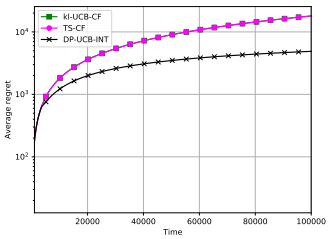
Figure 5: Regret plots scenario 2



(*a*) Regret plots for kl-UCB-CF and TS-CF with others for varying horizons up to 10^5

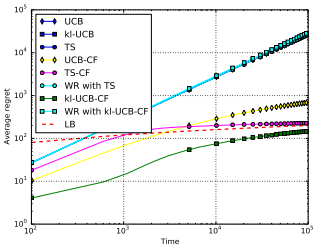


(*b*) Regret plots for kl-UCB-CF and TS-CF with $T = 10^5$, $\epsilon = \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$

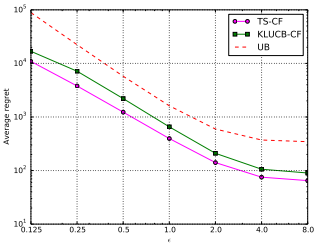


(*c*) Regret plots for $\epsilon = 1$ for kl-UCB-CF and TS-CF , DP-UCB-Int

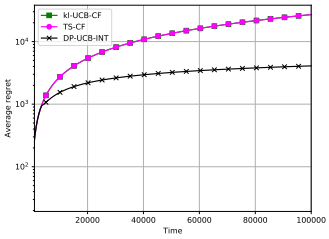
Figure 6: Regret plots scenario 3



(*a*) Regret plots for kl-UCB-CF and TS-CF with others for varying horizons up to 10^5



(*b*) Regret plots for kl-UCB-CF and TS-CF with $T = 10^5$, $\epsilon = \{\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8\}$



(*c*) Regret plots for $\epsilon = 1$ for kl-UCB-CF and TS-CF , DP-UCB-Int

Figure 7: Regret plots scenario 4