Corrupt Bandits

Pratik Gajane PRATIK.GAJANE@INRIA.FR
Orange labs/INRIA SequeL
Tanguy Urvoy TANGUY.URVOY@ORANGE.COM
Orange labs
Emilie Kaufmann EMILIE.KAUFMANN@INRIA.FR
INRIA SequeL

Editor:

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 preserving 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, and devise two upper confidence bound algorithms, UCB-CF and KLUCB-CF. We provide upper bounds on their regret and present some experimental results which confirm our analysis.

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

1. Introduction

The classical multi-armed bandits (MAB) problem is the formulation of the exploitation-exploration dilemma inherent to reinforcement learning (Bubeck and Cesa-Bianchi (2012)). In this problem, the learner has access to a number of available actions (symbolized by arms) and she has to select one of them (symbolized by pulling an arm) over a period of time to maximize his cumulative reward. The learner obtains information about the actions through the available feedback. In the classical MAB problem, the feedback is equal to the reward.

However, this assumption does not hold true for some practical scenarios. In online advertising, the feedback is usually given only when it's positive since propagating negative feedback as well is costly in terms of network load. The reception of a click feedback can be safely interpreted as a positive reward, but the absence of such a click (timeout) might either be a consequence of a negative reward (the user did not like the ad) or the consequence of a bug or a packet loss. In adaptive routing, positive feedback means the corresponding route is good but no feedback could either mean that the corresponding route was bad 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.

Feedback corruption is also an effective way to protect the respondent's individual privacy in online recommendation or survey systems. For instance, Warner (1965) proposed the *randomized response method (RR)* 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 respondents to employ randomization say with a coin flip, whose outcome is not available to the interviewer. By introducing random noise, the method conceals individual responses and protects respondent privacy.

The corrupted feedback we consider is a kind of incomplete feedback. Hence 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. Partial monitoring allows the learner, when it is possible, to retrieve the expected value of actions through an analysis of the feedback matrix and the reward matrix, both of which are assumed to be known to the learner. According to these matrices, the partial monitoring game may be either *trivial* with a minimax regret of 0, *easy* with minimax regret $\tilde{\Theta}(\sqrt{T})$ at time T, *hard* with minimax regret $\tilde{\Theta}(T^{2/3})$, or *hopeless* with a linear minimax regret.

Corrupt bandits is a particular instance of stochastic partial monitoring game with constrained environment, described below, and we aim for the best problem-dependent regret, that scales with $\log(T)$.

2. Problem setting

A corrupt bandit problem, denoted by ν, is characterized by K arms with means $\mu_1^\nu, \ldots, \mu_K^\nu$. If the learner pulls an arm a at round t, she receives a reward R_t drawn from a Bernoulli distribution with mean μ_a^ν and observes a feedback F_t drawn from a Bernoulli distribution with mean λ_a^ν. Let $a_*(\nu) \in \arg\max \mu_a^\nu$ be the optimal arm in the bandit model ν.[1] Without loss of generality, we assume 1 to be the optimal arm, unless mentioned otherwise, for the rest of this article. We assume that there exists a known function g_a which maps mean reward μ_a to mean feedback λ_a (such that the model ν is indeed characterized by μ_a^ν). We assume the function g_a to be (strictly) monotonic and continuous and we denote its corresponding inverse function by g_a^{-1}. The objective is to design a strategy, which chooses an arm A_t to be pulled at time t based only on the previously observed feedback, $F_1, \ldots, F_{t-1}$, in order to maximize the expected sum of rewards, or equivalently to minimize the expected regret (used interchangeably with regret in this article):

$$\mathrm{Regret}_T(\nu) = \mathbb{E}\left[\mu_1 - \sum_{t=1}^{T} R_t\right] = \sum_{a=2}^{K} \Delta_a \mathbb{E}[N_a(T)]$$

where $N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{(A_t=a)}$ denotes the number of pulls of arm a up to round T and $\Delta_a = \mu_1 - \mu_a$ i.e. the gap between the optimal reward and reward of arm a.

Randomized response. Randomized response (Warner (1965)), described in the introduction, can be simulated by having mean corruption function $g_a : \lambda_a = p_{10}(a) + (p_{11}(a) -$

1. When the associated model is clear from the context, we drop the symbol ν.

$p_{10}(aa))\mu_a$. The corresponding corruption scheme $\tilde{g}_a$ can be encoded by the matrix:

$$\mathbb{M}_a = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} p_{00}(a) & p_{01}(a) \\ p_{10}(a) & p_{11}(a) \end{bmatrix} \end{array}$$

The matrix $\mathbb{M}_a$ contains elements denoting the probability with which the learner sees, for arm a, the feedback given by the row index on receiving the reward given by the column index i.e. $\mathbb{P}(\text{feedback} = x \mid \text{reward} = y) = \mathbb{M}_a(x, y)$

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 bandit model ν has $\text{Regret}_T(\nu) = o(T^\alpha)$ for all $\alpha \in]0, 1[$. Theorem 1 provides a lower bound on the regret of a uniformly efficient algorithm. Its proof is given in Appendix A.

Theorem 1 *Fix $\{g_a\}_{a=1}^K$ the corruption functions. Any uniformly efficient algorithm satisfies, for a corrupt bandit problem ν,*

$$\liminf_{T\to\infty} \frac{\text{Regret}_T(\nu)}{\log(T)} \geq \sum_{a=2}^{K} \frac{\Delta_a}{d\left(\lambda_a, g_a(\mu_1)\right)} \qquad \textit{where } d(x, y) = \text{KL}(\mathcal{B}(x), \mathcal{B}(y))$$

Note the presence of $g_a(\mu_1)$ instead of λ_1; the latter would mean just a mirror image of the classical MAB problem in which rewards are replaced by feedbacks. Depending upon g_1 and g_a, λ_a can be greater or lower than $g_a(\mu_1)$ even though $\mu_1 > \mu_a$, which presents technical difficulties in this proof (and the subsequents proofs), unlike the classical MAB proofs.

4. Algorithms for MAB with corrupted feedback

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

4.1 KLUCB for MAB with corrupted feedback

We propose below an adaptation of the KL-UCB algorithm of Cappé et al. (2013): $\text{Index}_a(t)$ is an upper-confidence bounds on μ_a^ν built from a KL-based confidence interval on λ_a^ν. Theorem 2 gives an upper bound on the regret of KLUCB-CF, showing that it matches the lower bound of Theorem 1. A more explicit finite-time bound is proved in Appendix B.

Theorem 2 *The expected regret of KLUCB-CF using $f(t) = \log(t) + 3\log(\log(t))$ on a K-armed corrupted bandit with corruption functions $\{g_a\}_{a=1}^K$ is upper bounded by*

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

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

Input: A bandit model having K arms
Parameters: $\{g\}_{a=1}^{K}$, a non-decreasing (exploration) function $f : \mathbb{N} \to \mathbb{R}$, $d(x, y) = KL(\mathcal{B}(x), \mathcal{B}(y))$.
Initialization: Pull each arm once.
At time $t \geq K + 1$, do

Compute for each arm a one of the following quantities:

$$\begin{aligned} \ell_a(t) &= \min \{q : N_a(t) \cdot d(\hat{\lambda}_a(t), q) \leq f(t)\} \qquad \text{if } g_a \text{ is decreasing} \\ u_a(t) &= \max \{q : N_a(t) \cdot d(\hat{\lambda}_a(t), q) \leq f(t)\} \qquad \text{if } g_a \text{ is increasing} \end{aligned}$$

Pull arm $\arg\max_a \text{Index}_a(t)$ where

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

4.2 UCB for MAB with corrupted feedback

UCB1 (Auer et al. (2002)) can be adapted as well to corrupted feedback by modifying the index to the following:

$$\text{Index}_a(t) = \begin{cases} g_a^{-1}\Big(\hat{\lambda}_a(t) + \sqrt{\frac{\log t}{2N_a(t)}}\Big), & \text{if } \ g_a \text{ is increasing} \\ g_a^{-1}\Big(\hat{\lambda}_a(t) - \sqrt{\frac{\log t}{2N_a(t)}}\Big), & \text{if } g_a \text{ is decreasing} \end{cases}$$

Theorem 3 *The expected regret of UCB-CF using $f(t) = \log(t) + 3\log(\log(t))$ on a K-armed corrupted bandit with corruption functions $\{g_a\}_{a=1}^{K}$ is in $\mathcal{O}\Big(\sum_{a=2}^{K} \frac{\Delta_a \log(T)}{(g_a(\mu_a) - g_a(\mu^*))^2}\Big)$*

The proof of this theorem follows the proof of Theorem 2 given in Appendix B using the quadratic divergence $2(x - y)^2$ in place of $d(x, y)$. The UCB-CF algorithm is only order optimal with respect to the bound of Theorem 1, but its index is simpler to compute. The following corollary follows from Theorem 3 for UCB-CF and Theorem 2 and Pinsker's inequality ($d(x, y) \geq 2(x - y)^2$) for KL-UCB-CF.

Corollary 3.1 *The expected regret of UCB-CF and KLUCB-CF in a MAB problem with randomized response using corruption matrices $\mathbb{M}_a$ is in $\mathcal{O}\Big(\sum_{a=2}^{K} \frac{\log(T)}{\Delta_a(p_{00}(a)+p_{11}(a)-1)^2}\Big)$*

4.3 Thompson sampling for MAB with corrupted feedback (Thompson Sampling-CF)

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

5. Corrupted feedback to enforce differential Privacy

Jain et al. (2012), Thakurta and Smith (2013), Mishra and Thakurta (2015) have observed the importance of privacy to MAB applications. There are two scenarios in the MAB privacy protection. In the first scenario, privacy is to be protected by the way of making the output of the algorithm (i.e. the chosen actions) not reveal private information (i.e. the individual rewards) to an outside observer. We shall call this as *privacy preserving output*. While the second scenario is *privacy preserving input*, where privacy protection is done by making input to the algorithm (i.e. the feedback) unintelligible to an outside observer.

Differential privacy is a way which could solve the privacy concerns in MAB applications. Differential privacy was introduced by Dwork et al. (2006). For a comprehensive overview, refer to Dwork and Roth (2014). Thakurta and Smith (2013) present a differentially private algorithm for the adversarial MAB problem, while Mishra and Thakurta (2015) provide a differentially private algorithm for the stochastic MAB problem. Both of these approaches address the scenario of privacy preserving output. As stated in Tossou and Dimitrakakis (2016), a differentially private bandit algorithm, for all reward sequences $R_{1:t-1}$ and $R'_{1:t-1}$ that differ in at most one place and for all $S \subseteq K$, satisfies

$$\mathbb{P}(A_t \in S \mid A_{1:t-1}, R_{1:t-1}) \leq \mathbb{P}(A_t \in S \mid A_{1:t-1}, R'_{1:t-1})e^{\epsilon} + \delta$$

In this article, we consider the scenario of privacy preserving input. 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 with confidence the sensitive information. We, on the other hand, corrupt the rewards using randomized response in such a way that an untrustworthy observer can not, with confidence, recover rewards from feedback being propagated over an unreliable connection.

Definition 4 *((ϵ, δ)-differentially private bandit feedback corruption scheme) A bandit feedback corruption scheme $\tilde{g}$ is (ϵ, δ)-differentially private if for all reward sequences $R_{t1}, \ldots, R_{t2}$ and $R'_{t1} \ldots, R'_{t2}$ that differ in at most one reward, and for all $\mathcal{S} \subseteq Range(g)$*

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

If $\delta = 0$, then $\tilde{g}$ is said to be ϵ-differentially private.

In the case, where corruption is done by randomized response, differential privacy requires that

$$\max_{a \in K} \Big(\frac{p_{00}(a)}{p_{11}(a)}, \frac{p_{11}(a)}{p_{10}(a)} \Big) \leq e^{\epsilon} + \delta$$

From Corollary 3.1, we can see that to achieve lower expected regret, $p_{00}(a) + p_{11}(\mathrm{a})$ is to be maximized for all $a \in K$. Using result 1 from Wang et al. (2016, p. 3), we can state that, in order to achieve (ϵ, δ)-differential privacy while maximizing $p_{00}(a) + p_{11}(a)$,

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

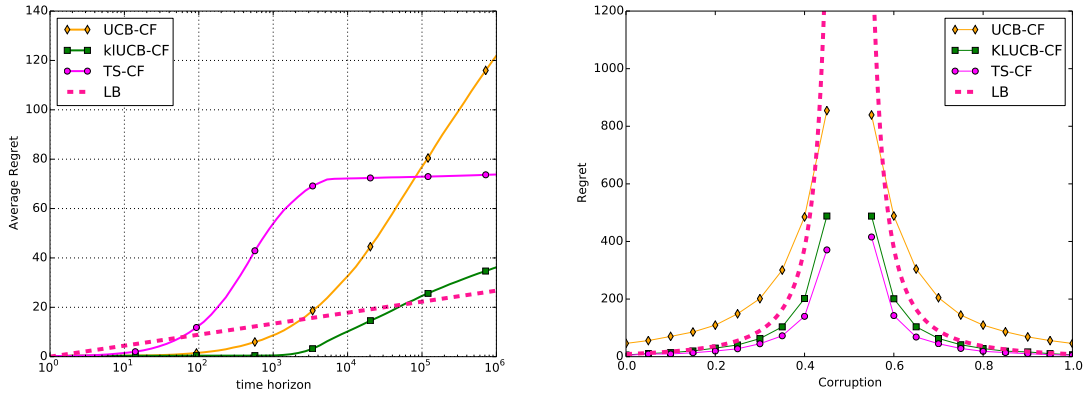
Thus, from Corollary 3.1, the expected regret of UCB-CF or KL-UCB-CF with (ϵ, δ)-differentially private bandit feedback corruption scheme is $\mathcal{O}\Big(\sum_{a=2}^{K} \Big(\frac{e^{\epsilon}+\delta+1}{e^{\epsilon}+\delta-1} \Big)^2 \frac{\log(T)}{\Delta_a} \Big)$.

6. Experiments

Randomized response was employed to corrupt the feedback with different degrees of corruption and each experiment was repeated 400 times.

$$\text{Scenario 1 mean rewards :} \qquad 0.9 \quad 0.6$$

We give plots for scenario 1 in Figure 1. Subfigure 1a shows average regret for $p_{00} = p_{11} = 0.6$ for the optimal arm, while for all the other arms, both p_{00} and p_{11} were set to 0.9, while Subfigure 1b shows the performance of the algorithms for varying values of $p = p_{00} = p_{11}$ kept same for all the arms. Additional plots for the rest of scenarios are given in Appendix C. KLUCB-CF outperforms all the competitors in these scenarios. The performance superiority of KLUCB-CF is more pronounced in the scenarios in which the corruption causes the best arm to be switched (or at least when to be no longer unique).



(a) (b)

Figure 1: Regret plots for scenario 1

7. Conclusion

As shown by experiments, UCB-CF, KLUCB-CF, and Thompson Sampling-CF provide suitable solutions to the MAB problem with corrupted feedback. KLUCB-CF is shown to be the best solution: it is proved to be asymptotically optimal as it matches the asymptotic lower bound on the expected regret of any uniformly efficient algorithm, but it also outperforms all its competitors in our numerical experiments. It is thus a good candidate to be used in recommender systems to apply a randomized response mechanism to protect the users privacy. Furthermore, we exhibit appropriate corruption matrices that achieve a desired level of differential privacy, and quantify their impact on the regret.

This work can be extended in many ways. In some situations, the feedback is simply lost. It is possible to extend our problem setting to incorporate such scenarios by making appropriate changes to the corruption process. Another possible extension is adversarial corruption of the feedback. It is also possible to extend present partial monitoring algorithms to deal with this problem setting.

References

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.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):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.

Aurlien Garivier, Gilles Stoltz, and Pierre Mnard. Explore first, exploit next: The true shape of regret in bandit problems. 2016.

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. URL `http://www.jmlr.org/proceedings/papers/v23/jain12/jain12.pdf`.

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. ISBN 3-540-42343-5. doi: 10.1007/3-540-44581-1_14. URL `http://dblp.uni-trier.de/db/conf/colt/colt2001.html#PiccolboniS01`.

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. *Bulletin of the AMS*, 25:285–294, 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.

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`.

Appendix

Appendix A. Proof for the lower bound on the regret

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

Given two corrupted bandit models ν and ν', the log-likelihood ratio of the observations up to time T under ν and ν', that we denote by $L_T(\nu, \nu')$ can be written

$$L_T(\nu, \nu') = \sum_{a=1}^{K} \sum_{s=1}^{N_a(T)} \log \frac{f_{\lambda_a^\nu}(F_{a,s})}{f_{\lambda_a^{\nu'}}(F_{a,s})},$$

where $f_x(\cdot)$ denotes the Bernoulli density of mean x and $F_{a,s}$ are the successive observed *feedback* from arm a. By Wald's lemma

$$\mathbb{E}_\nu \left[L_T(\nu, \nu') \right] = \sum_{a=1}^{K} \mathbb{E}_\nu[N_a(T)] d(\lambda_a^\nu, \lambda_a^{\nu'}),$$

where $d(x, y) = \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y)) = x \log(x/y) + (1-x) \log((1-x)/(1-y))$, with the convention that $d(0, 0) = d(1, 1) = 0$. The following lemma can thus be extracted from (Garivier et al., 2016).

Lemma 5 *Let ν and ν' be two corrupted bandit models and $T \in \mathbb{N}^*$. For any random variable $Z \in [0, 1]$ that is $\mathcal{F}_T$-measurable, one has*

$$\sum_{a=1}^{K} \mathbb{E}_\nu[N_a(T)] \cdot d\left(\lambda_a^\nu, \lambda_a^{\nu'}\right) \geq d\left(\mathbb{E}_\nu(Z), \mathbb{E}_{\nu'}(Z)\right).$$

Using Lemma 5 with $Z = \frac{N_1(T)}{T}$, we obtain

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

Using the inequality

$$d(p, q) \geq p \log(1/q) - \log(2)$$

(see Garivier et al. (2016)) yields

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

Since $a^*(v) = 1$, and $a^*(v') \neq 1$, $\mathbb{E}_\nu(N_1(T)) \sim T$ and $\mathbb{E}_{\nu'}(N_1(T)) = o(T^\alpha)$ for all $\alpha \in]0, 1]$. Hence one can show that

$$\frac{\mathbb{E}_\nu(N_1(T))}{T} \sim 1 \quad \text{and} \quad \log\left(\frac{T}{\mathbb{E}_{\nu'}[N_1(T)]}\right) \sim \log(T).$$

Using equation (1), one obtains

$$\liminf_{T\to\infty} \frac{\sum_{a=1}^{K} \mathbb{E}_\nu(N_a(T)) d\left(\lambda_a^\nu, \lambda_a^{\nu'}\right)}{\log T} \geq 1 \tag{2}$$

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

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

This translates to the following change in feedback,

$$\lambda_b^{\nu'} = \begin{cases} g_b(\mu_1^\nu + \epsilon) & \text{if } b = a, \\ g_b(\mu_b^\nu) = \lambda_b^\nu & \text{otherwise.} \end{cases}$$

As $d\left(\lambda_b^\nu, \lambda_b^{\nu'}\right) = 0$ for $b \neq a$, using equation 2 we get

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

Since the regret of the algorithm at time T is

$$\text{Regret}_T = \sum_{a=2}^{K} \mathbb{E}(N_a(T))(\mu_*^\nu - \mu_a^\nu),$$

one obtains, letting ϵ go to zero for each $a = 2, \ldots, K$ (and using that the g_a are continuous),

$$\liminf_{T\to\infty} \frac{\text{Regret}_T(\nu)}{\log(T)} \geq \sum_{a=2}^{K} \frac{\mu_*^\nu - \mu_i^\nu}{d\left(\lambda_a^v, g_a(\mu_*^\nu)\right)}.$$

Appendix B. Proof for the upper bound on the regret of KLUCB-CF

To get an upper bound on the expected regret of this algorithm, we first bound $\mathbb{E}[N_a(t)]$ for all the non-optimal arms a. A_t = arm pulled by the algorithm at time t. Again for the sake of readability, we assume 1 to be the optimal arm.

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

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.

$$\begin{aligned}
(A_{t+1} = a) &\subseteq (u_1(t) < g_1(\mu_1^\nu)) \cup (A_{t+1} = a, u_1(t) \geq g_1(\mu_1^\nu)) \\
&= (u_1(t) < g_1(\mu_1^\nu)) \cup (A_{t+1} = a, g_1^{-1}(u_1(t)) \geq \mu_1^\nu) && \text{since } g_1^{-1} \text{ is increasing} \\
&= (u_1(t) < g_1(\mu_1^\nu)) \cup (A_{t+1} = a, g_i^{-1}(u_a(t)) \geq \mu_1^\nu) && \text{since Index}_i > \text{Index}_1 \\
&= (u_1(t) < g_1(\mu_1^\nu)) \cup (A_{t+1} = a, u_a(t) \geq g_a(\mu_1^\nu)) && \text{since } g_a \text{ is increasing}
\end{aligned}$$

$$\therefore \mathbb{E}(N_a(T)) \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(u_1(t) < g_1(\mu_1^\nu)) + \sum_{t=K}^{T-1} \mathbb{P}(A_{t+1} = a, u_a(t) \geq g_a(\mu_1^\nu)) \quad (3)$$

- g_1 is decreasing and g_a is increasing.

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

$$\therefore \mathbb{E}(N_a(T)) \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(\ell_1(t) > g_1(\mu_1^\nu)) + \sum_{t=K}^{T-1} \mathbb{P}(A_{t+1} = a, u_a(t) \geq g_a(\mu_1^\nu)) \quad (4)$$

- g_1 is increasing and g_a is decreasing.

$$\begin{aligned}
(A_{t+1} = a) &\subseteq (u_1(t) < g_1(\mu_1^\nu)) \cup (A_{t+1} = a, u_1(t) \geq g_1(\mu_1^\nu)) \\
&= (u_1(t) < g_1(\mu_1^\nu)) \cup (A_{t+1} = a, g_1^{-1}(u_1(t)) \geq \mu_1^\nu) && \text{since } g_1 \text{ is increasing} \\
&= (u_1(t) < g_1(\mu_1^\nu)) \cup (A_{t+1} = a, g_a^{-1}(\ell_a(t)) \geq \mu_1^\nu) && \text{since } A_{t+1} = a \\
&= (u_1(t) < g_1(\mu_1^\nu)) \cup (A_{t+1} = a, \ell_a(t) \leq g_a(\mu_1^\nu)) && \text{since } g_a \text{ is decreasing}
\end{aligned}$$

$$\therefore \mathbb{E}(N_a(T)) \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(u_1(t) < g_1(\mu_1^\nu)) + \sum_{t=K}^{T-1} \mathbb{P}(A_{t+1} = a, \ell_a(t) \leq g_a(\mu_1^\nu)) \quad (5)$$

- g_1 is decreasing and g_a is decreasing.

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

$$\therefore \mathbb{E}(N_a(T)) \leq 1 + \sum_{t=K}^{T-1} \mathbb{P}(\ell_1(t) > g_1(\mu_1^\nu)) + \sum_{t=K}^{T-1} \mathbb{P}(A_{t+1} = a, \ell_a(t) \leq g_a(\mu_1^\nu)) \quad (6)$$

Let $\hat{\lambda}_{a,s}$ = empirical mean feedback from arm a after s samples.
We first upper bound the two sums

$$\sum_{t=K}^{T-1} \mathbb{P}(u_1(t) < g_1(\mu_1^\nu)) \quad \text{and} \quad \sum_{t=K}^{T-1} \mathbb{P}(\ell_1(t) > g_1(\mu_1^\nu)) \tag{7}$$

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

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

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

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

the two quantities in (7) 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

$$\sum_{t=K}^{T-1} \mathbb{P}(u_1(t) < g_1(\mu_1)) \leq 3 + 4e \log(\log T) \in o(\log T) \tag{8}$$

$$\sum_{t=K}^{T-1} \mathbb{P}(\ell_1(t) > g_1(\mu_1)) \leq 3 + 4e \log(\log T) \in o(\log T) \tag{9}$$

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

where $\mathbb{1}$ is the indicator function. So we can write, when g_a is increasing,

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

One obtains, when g_a is increasing,

$$\sum_{t=K}^{T-1} \mathbb{P}(A_{t+1}=a, u_a(t) \geq g_a(\mu_1^\nu)) \leq \sum_{s=1}^{T-1} \mathbb{P}\left(s\cdot d^+(\hat{\lambda}_{a,s}, g_a(\mu_1^\nu)) \leq f(T)\right). \tag{10}$$

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

$$\sum_{t=K}^{T-1} \mathbb{P}(A_{t+1}=a, \ell_a(t) \leq g_a(\mu_1^\nu)) \leq \sum_{s=1}^{T-1} \mathbb{P}\left(s\cdot d^-(\hat{\lambda}_{a,s}, g_a(\mu_1^\nu)) \leq f(T)\right). \tag{11}$$

The quantity in the right-hand side of (10) is upper bounded in Appendix A.2. of Cappé et al. (2013) by

$$\frac{f(T)}{d(g_a(\mu_a), g_a(\mu_1)} + \sqrt{2\pi}\sqrt{\frac{(d'(g_a(\mu_a), g_a(\mu_1)))^2}{(d(g_a(\mu_a), g_a(\mu_1)))^3}}\sqrt{f(T)} + 2\left(\frac{d'(g_a(\mu_a), g_a(\mu_1))}{d(g_a(\mu_a), g_a(\mu_1))}\right)^2 + 1. \tag{12}$$

Following the same approach as Cappé et al. (2013), we now prove that the right-hand side of (11) is upper bounded by the same quantity.

Letting $s_0 = \left\lceil \frac{f(T)}{d(g_a(\mu_a), g_a(\mu_1))} \right\rceil$,

$$\sum_{s=1}^{T-1} \mathbb{P}(s\cdot d^-(\hat{\lambda}_{a,s}, g_a(\mu_1)) \leq f(T)) \quad \leq \quad s_0 + \sum_{s=s_0+1}^{\infty} \mathbb{P}(s\cdot d^-(\hat{\lambda}_{a,s}, g_a(\mu_1)) \leq f(T))$$

The mapping $x \mapsto d^-(x, g_a(\mu_1))$ is increasing and for every $\gamma \in]0, d(g_a(\mu_a), g_a(\mu_1))[$, there exists a unique $x_\gamma^* \in]g_a(\mu_1), g_a(\mu_a)[$ such that

$$d^-(x_\gamma^*, g_a(\mu_1)) = \gamma.$$

For $s \geq s_0 + 1$, $f(T)/s \leq d(g_a(\mu_a), d(g_a(\mu_1))$ hence $x^*_{f(T)/s} \leq g_a(\mu_a)$ and using Chernoff inequality, one gets

$$\begin{aligned}
\sum_{s=1}^{T-1} \mathbb{P}(s \cdot d^-(\hat{\lambda}_{a,s}, g_a(\mu_1)) &\leq s_0 + \sum_{s=s_0+1} \mathbb{P}\left(\hat{\lambda}_{a,s} \leq x^*_{f(T)/s}\right) \\
&\leq s_0 + \sum_{s=s_0+1} \exp\left(-s \cdot d\left(x^*_{f(T)/s}, g_a(\mu_a)\right)\right) \\
&\leq s_0 + \sum_{s=s_0+1} \exp\left(-f(T)\phi\left(\frac{s}{f(T)}\right)\right),
\end{aligned}$$

where we introduce the mapping $\phi : [d(g_a(\mu_a), g_a(\mu_1))^{-1}, +\infty[$ defined

$$\phi(u) = u \cdot d(x^*_{1/u}, g_a(\mu_a)).$$

As the two mappings,

$$\gamma \longmapsto x^*_\gamma \longmapsto d(x^*_\gamma, g_a(\mu_a))$$

where $\gamma \in [0, d(g_a(\mu_a), g_a(\mu_1)]$ and $x^*_\gamma \in [g_a(\mu_1), g_a(\mu_a)]$ are respectively increasing and decreasing, their composition is a decreasing application. Thus the mapping ϕ is increasing. Moreover, one can prove the lower bound

$$\phi(u) \geq 2u\left(\frac{d(g_a(\mu_a), g_a(\mu_1)) - 1/u}{d'(g_a(\mu_a), g_a(\mu_1))}\right)^2. \tag{13}$$

Indeed, as $x \mapsto d(x, g_a(\mu_1))$ is convex, its curve is above its tangent in $g_a(\mu_a)$

$$d(x, g_a(\mu_1)) \geq d(g_a(\mu_a), g_a(\mu_1)) + d'(g_a(\mu_a), g_a(\mu_1))(x - g_a(\mu_a)),$$

which yields that x^*_γ is upper bounded by the solution in x to

$$d(g_a(\mu_a), g_a(\mu_1)) + d'(g_a(\mu_a), g_a(\mu_1))(x - g_a(\mu_a)) = \gamma,$$

and finally $x^*_\gamma \leq g_a(\mu_a) - \frac{d(g_a(\mu_a), g_a(\mu_1)) - \gamma}{d'(g_a(\mu_a), g_a(\mu_1))}$. As $x \mapsto d(x, g_a(\mu_a))$ is decreasing on $[0, g_a(\mu_a)]$,

$$d(x^*_\gamma, g_a(\mu_a)) \geq d\left(g_a(\mu_a) - \frac{d(g_a(\mu_a), g_a(\mu_1)) - \gamma}{d'(g_a(\mu_a), g_a(\mu_1))}, g_a(\mu_a)\right) \geq 2\left(\frac{d(g_a(\mu_a), g_a(\mu_1)) - \gamma}{d'(g_a(\mu_a), g_a(\mu_1))}\right)^2,$$

by Pinsker's inequality[2], which yields (13).

In the sequel, we introduce the shorthand $D = d(g_a(\mu_a), g_a(\mu_1))$ and $D' = d'(g_a(\mu_a), g_a(\mu_1))$. Using the monotonicity of ϕ and the lower bound (13) yield

$$\begin{aligned}
\sum_{s=1}^{T-1} \mathbb{P}(s \cdot d^-(\hat{\lambda}_{a,s}, g_a(\mu_1)) &\leq s_0 + \int_{s_0}^{\infty} \exp\left(-f(T) \cdot \phi\left(\frac{s}{f(T)}\right)\right) ds, \\
&\leq s_0 + f(T) \int_{\frac{1}{D}}^{\infty} \exp\left(-f(T) \cdot \phi\left(u\right)\right) du, \\
&\leq s_0 + f(T) \underbrace{\int_{\frac{1}{D}}^{\infty} \exp\left(-2f(T) \cdot u \cdot \left(\frac{D - 1/u}{D'}\right)^2\right) du}_{I}
\end{aligned}$$

2. Pinsker's inequality states that if P and Q are two probability distributions on a measurable space (X, Σ), then $d(P, Q) \leq \sqrt{KL(P, Q)/2}$

To conclude, we upper bound the integral I in the right-hand side by

$$\begin{aligned}
&\int_{\frac{1}{D}}^{\frac{2}{D}} \exp\left(-2f(T)\cdot\frac{(D-1/u)^2}{DD'^2}\right)du + \int_{\frac{2}{D}}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{D}{D'}\right)^2 f(T)\cdot u\right)du \\
&\leq \frac{4}{D^2}\int_0^{D/2}\exp\left(-\frac{2f(T)}{DD'^2}v^2\right)dv + 2\left(\frac{D'}{D}\right)^2\frac{1}{f(T)} \\
&\leq \sqrt{2\pi}\sqrt{\frac{D'^2}{D^3}}\frac{1}{\sqrt{f(T)}} + 2\left(\frac{D'}{D}\right)^2\frac{1}{f(T)}.
\end{aligned}$$

Putting things together, we have

$$\sum_{s=1}^{T-1}\mathbb{P}(sd^-(\hat{\lambda}_{a,s}, g_a(\mu_1)) \leq \frac{f(T)}{D} + \sqrt{2\pi}\sqrt{\frac{D'^2}{D^3}}\sqrt{f(T)} + 2\left(\frac{D'}{D}\right)^2 + 1, \tag{14}$$

which is identical to (12).

Combining inequalities (8), (9), (10) and (12) ,(11) and (14) with the initial decomposition of $\mathbb{E}[N_a(T)]$ yield in all cases

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

This yields that the regret of KL-UCB-CF is upper bounded by

$$\sum_{i\neq i^*}\Delta_i\Big[\frac{\log(T)}{D_i} + \sqrt{2\pi}\sqrt{\frac{(D_i')^2}{D_i^3}}\sqrt{\log(T) + 3\log\log(T)} + \left(4e + \frac{3}{D_i}\right)\log\log(T) + 2\left(\frac{D_i'}{D_i}\right)^2 + 4\Big]$$

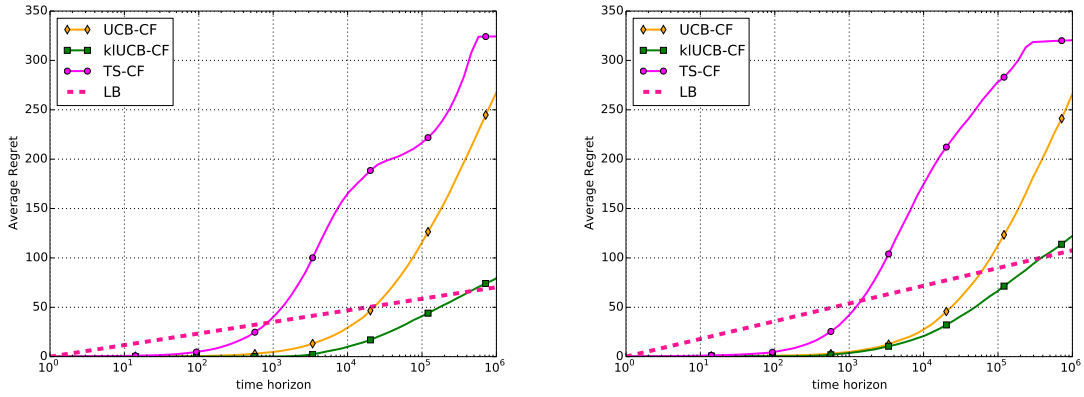
where $D_i = d(g_a(\mu_a), g_a(\mu^*))$ and $D_i' = d'(g_a(\mu_a), g_a(\mu^*))$, which concludes the proof.

Appendix C. Additional experimental plots

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. Each each experiment was repeated 400 times. In Table 1, we provide the mean rewards and the mean feedbacks for the arms of the respective scenarios.

Table 1: Mean reward and feedback for experiments

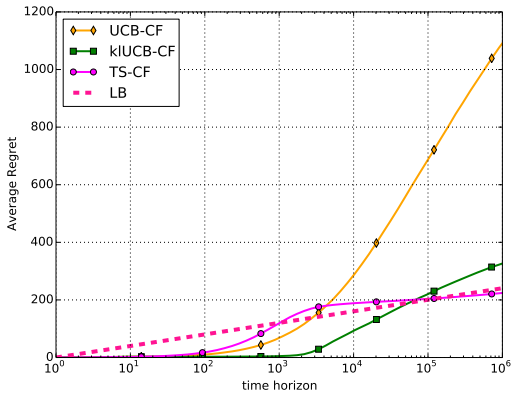
Scenario	Mean	Arms									
		1	2	3	4	5	6	7	8	9	10
2	Reward	0.9	0.8								
	Feedback	0.58	0.74								
3	Reward	0.55	0.45								
	Feedback	0.51	0.46								
4	Reward	0.9	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
	Feedback	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
5	Reward	0.9	0.8	0.8	0.8	0.7	0.7	0.7	0.6	0.6	0.6
	Feedback	0.58	0.74	0.74	0.74	0.66	0.66	0.66	0.58	0.58	0.58
6	Reward	0.9	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
	Feedback	0.58	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74
7	Reward	0.55	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45	0.45
	Feedback	0.51	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.46



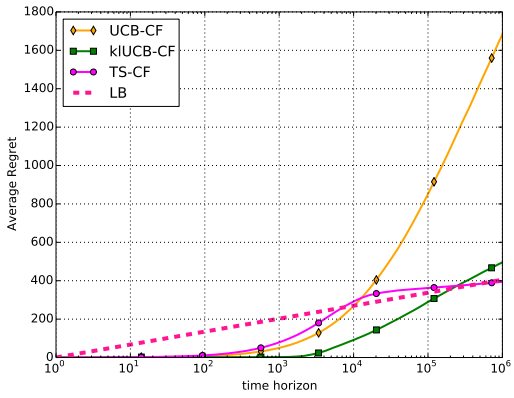
(a) For scenario 2

(b) For scenario 3

Figure 2: Regret plots

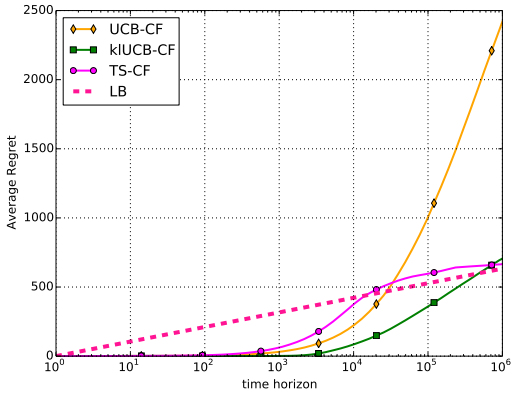


(a) For scenario 4

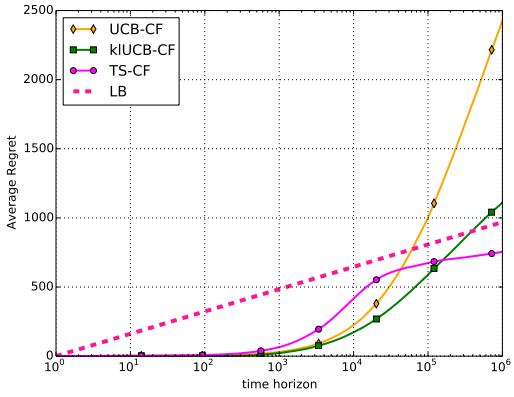


(b) For scenario 5

Figure 3: Regret plots



(a) For scenario 6



(b) For scenario 7

Figure 4: Regret plots