Corrupt Bandits

Pratik Gajane[†⋆], Tanguy Urvoy[†], Emilie Kaufmann[⋆]
pratik.gajane@inria.fr, tanguy.urvoy@orange.com, emilie.kaufmann@univ-lille1.fr
Orange labs[†], CNRS & CRIStAL (SequeL)[⋆]

EWRL 2016 Barcelona

Motivation

- preserve user privacy in online recommender systems.
- conceal individual choices about sensitive behaviors and beliefs.

  Example: *Randomized response method (RR)* [Warner (1965)]

➜ We introduce generalized *corruption functions*.

Problem setting

- K arms with means $\mu_1, \ldots, \mu_K$ w.l.o.g. $\mu_1 > \mu_{2,\cdots,K}$
- Learner pulls an arm A_t at time $t = 1, \ldots, T$
  - receives **reward** $\sim$ Bernoulli distribution with mean μ_{A_t}
  - observes **feedback** $\sim$ Bernoulli distribution with mean λ_{A_t}
- A known corruption function $g_a : \mu_a \mapsto \lambda_a$
- Assumption: g_a is monotonic and continuous.

➜ Goal: Minimize $\text{Regret}_T = \sum_{a=2}^{K} \Delta_a \mathbb{E}[N_a(T)]$ where $N_a(T) = \sum_{t=1}^{T} \mathbb{1}_{(A_t = a)}$ and $\Delta_a = \mu_1 - \mu_a$

Randomized response

- Corruption function $g_a : \lambda_a = p_{10}(a) + (p_{11}(a) - p_{10}(a))\mu_a$
- $\mathbb{P}(\text{feedback} = x \mid \text{reward} = y) = \mathbb{M}_a(x, y)$

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

Lower bound on regret

Definition 1. *An uniformly efficient algorithm for the corrupt bandit problem is an algorithm which, for any bandit model, has* $\text{Regret}_T = o(T^\alpha)$ *for all* $\alpha \in]0, 1[$.

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

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

KLUCB-CF

1: **Input:** A bandit model having K arms
2: **Parameters:** $\{g\}_{a=1}^{K}$, a non-decreasing (exploration) function $f : \mathbb{N} \to \mathbb{R}$, $d(x, y) = KL(\mathcal{B}(x), \mathcal{B}(y))$.
3: **Initialization:** Pull each arm once.
4: At time $t \geqslant K + 1$, do
5: Compute for each arm a, one of the following quantities:

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

where

$$\begin{aligned} \ell_a(t) &= \min\{q : N_a(t) \cdot d(\hat{\lambda}_a(t), q) \leqslant f(t)\} \\ u_a(t) &= \max\{q : N_a(t) \cdot d(\hat{\lambda}_a(t), q) \leqslant f(t)\} \end{aligned}$$

6: Pull arm $A_{t+1} = \arg\max_a \ \text{Index}_a(t)$.
7: Observe feedback F_{t+1}.

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*

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

UCB-CF

Modification of UCB1 [Auer et al. (2002)] with changed index given below:

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

Theorem 3. *The expected regret of UCB-CF using* $f(t) = \log(t) + 3\log(\log(t))$, $\text{Regret}_T \in \mathcal{O}\left(\sum_{a=2}^{K} \frac{\Delta_a \log(T)}{(g_a(\mu_a) - g_a(\mu^*))^2}\right)$

Thompson Sampling-CF

1: Keep a Beta posterior distribution on the mean feedback of each arm.
2: At time t, for each arm a, draw a sample $\theta_a(t)$ from the posterior distribution on λ_a^{ν}.
3: Pull the arm for which $g_a^{-1}(\theta_a(t))$ is largest.

Corrupted feedback to enforce differential Privacy

Definition 2. *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* $\mathcal{S} \subseteq Range(\tilde{g})$

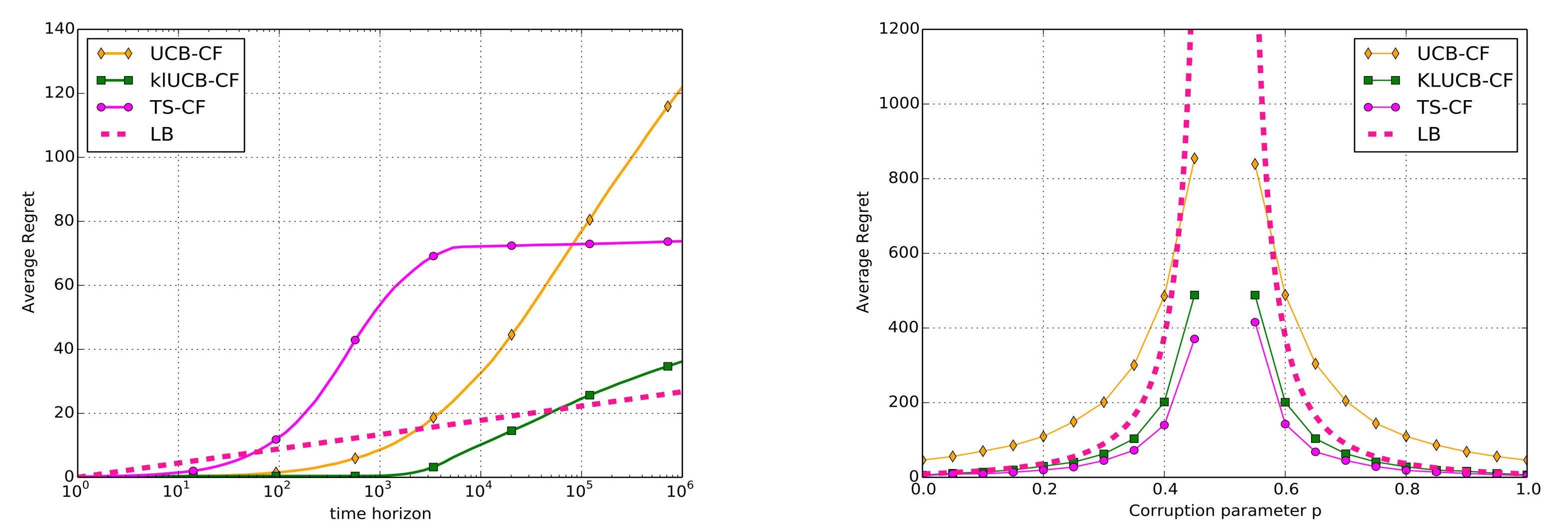
$$\mathbb{P}[\tilde{g}(R_{t1}, \ldots, R_{t2}) \in \mathcal{S}] \leqslant e^{\epsilon} \cdot \mathbb{P}[\tilde{g}(R'_{t1}, \ldots, R'_{t2}) \in \mathcal{S}] + \delta$$

- Privacy preserving input
- Differential privacy requires that $\max_{a\in K}\left(\frac{p_{00}(a)}{p_{11}(a)}, \frac{p_{11}(a)}{p_{10}(a)}\right) \leqslant e^{\epsilon} + \delta$
- To achieve (ϵ, δ)-differential privacy with randomized response,

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

Experiments

- Randomized response as corruption function.
- Scenario 1: Two arms with mean rewards 0.9 and 0.6
- Figure 1(a) shows average regret for $p_{00}(1) = p_{11}(1) = 0.6$ and $p_{00}(2) = p_{11}(2) = 0.9$
- Figure 1(b) shows the performance for varying values of $p = p_{00}(1) = p_{11}(1) = p_{00}(2) = p_{11}(2)$ with $T = 10^4$



1(a) 1(b)

Conclusion

- UCB-CF, KLUCB-CF, and Thompson Sampling-CF provide suitable solutions. KLUCB-CF is the best solution as it is asymptotically optimal and outperforms others in experiments.
- We provide appropriate corruption matrices that achieve a desired level of differential privacy.

Key references

[Warner Stanley (1965)] Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias.
[Auer Peter, Cesa-Bianchi Nicolò, and Fischer Paul (2002)] Finite-time Analysis of the Multiarmed Bandit Problem.