Corrupt Bandits for Preserving Local Privacy

Pratik Gajane [1] Tanguy Urvoy [2] Emilie Kaufmann [3]

7$^{\mathrm{th}}$ April 2018

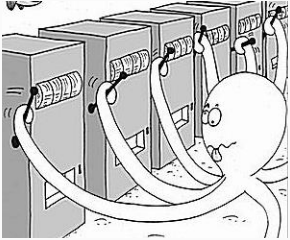
[1]Montanuniversität Leoben

[2]Orange labs

[3]CNRS & Univ. Lille & Inria-SequeL

Motivation and Formalization

Classical Stochastic Bandits



- K arms/actions
- Unknown reward distributions with mean μ_a for arm a
- Learner pulls arm a
  - receives reward $\sim$ distribution for a
  - feedback = received reward (**Absolute feedback**)

- Regret = best possible reward - reward of pulled arm
- Learner's goal = minimize cumulative regret

Motivation for Corrupt Bandits: Privacy

Motivation for Corrupt Bandits: Privacy

"If you're doing something that you don't want other people to know, maybe you shouldn't be doing it in first place"

"Privacy is no longer a social norm!"

Local Differential Privacy

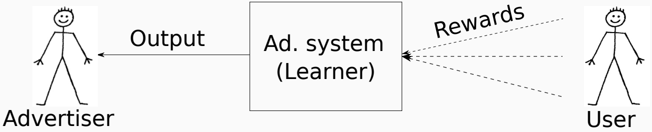


Figure 1: Ad system using bandits

- Ad application as bandit problem.
- Feedback from users on ads (arms).
- Information about user tastes as output to advertisers.

Local Differential Privacy

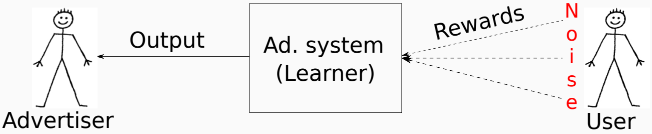


Figure 1: Ad system using bandits

- Ad application as bandit problem.
- Feedback from users on ads (arms).
- Information about user tastes as output to advertisers.
- Local differential privacy (DP), by Duchi et al.(2014) [2].
- Classical bandits unable to deal with noisy feedback.

Questions???

- Bandit setting to deal with Corrupted/Noisy Feedback?
- Regret Lower Bound for such Bandit setting?
- Algorithms to solve this Bandit setting?

Corrupt Bandits: Formalization

- Formally characterized by
    - K arms
    - unknown reward distribution with mean μ_a for each a
    - unknown feedback distribution with mean λ_a for each a
    - known mean corruption function g_a for each a
- $g_a(\mu_a) = \lambda_a$
- Learner's goal: minimize cumulative regret

Lower Bound on Regret

Lower Bound

Theorem (Thm. 1, PG, Urvoy & Kaufmann(2018) [4])

Any algorithm for a Bernoulli corrupt bandit problem satisfies,

$$\liminf_{T\to\infty} \frac{\text{Regret}_T}{\log(T)} \geq \sum_{a=2}^{K} \frac{\Delta_a}{d\left(\lambda_a, g_a(\mu_1)\right)}.$$

$$\textit{where } d(x,y) := \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y))$$

- Δ_a = optimal mean reward - mean reward of a (μ_a)
- 1 is assumed to be the optimal arm w.l.o.g.
- $\lambda_a = g_a(\mu_a)$. Behaviour of g_a on μ_a and μ_1 affects lower bound.

Proposed Algorithms

Proposed algorithm: kl-UCB-CF

Algorithm: kl-UCB-CF

Pull at time t an arm maximizing

$$\mathrm{Index}_a(t) := \max\{q : N_a(t) \cdot d(\hat{\lambda}_a(t), g_a(q)) \le f(t)\}$$

- Similar to kl-UCB by Cappé et al. (2013) [1] for classical bandits.
- $\mathrm{Index}_a(t) =$ UCB on μ_a from confidence interval on λ_a and using exploration function f
- $\hat{\lambda}_a(t) =$ emp. mean of feedback of a until time t
- UCB1 (Auer et al. (2002)) can be updated to UCB-CF.

Upper Bound for kl-UCB-CF

Theorem (Thm. 2, PG, Urvoy & Kaufmann(2018) [4])

Regret of kl*-UCB-CF* $\leq \sum_{a=2}^{K} \frac{\Delta_a \log(T)}{d(\lambda_a, g_a(\mu_1))} + O(\sqrt{\log(T)})$.

- Recall that 1 is assumed to be the optimal arm.
- More explicit bound can be provided.
- Optimal as upper bound matches lower bound.

Proposed Algorithm: TS-CF

Algorithm: TS-CF

1. Sample $\theta_a(t)$ *from Beta posterior distribution on mean feedback of arm a.*
2. Pull arm $\hat{a}_{t+1} = \arg\max_a g_a^{-1}(\theta_a(t))$.

- Similar to Thompson sampling by Thompson (1933) [5] for classical bandits.
- Probability (a is played) = posterior probability (a is optimal).

Upper Bound for TS-CF

Theorem (Thm. 3, PG, Urvoy & Kaufmann(2018) [4])

Regret of TS-CF $\leq \sum_{a=2}^{K} \frac{2\Delta_a \log(T)}{d(\lambda_a, g_a(\mu_1))} + O(\sqrt{\log(T)})$

- Recall that 1 is assumed the be the optimal arm.
- A tighter bound can be provided.
- Optimal as upper bound matches lower bound.

Experiments

Experiments with varying time

- Bernoulli corrupt bandit: $\mu_1 = 0.9 \qquad \mu_2 = \cdots = \mu_{10} = 0.6$
- Comparison over a period of time for fixed corruption

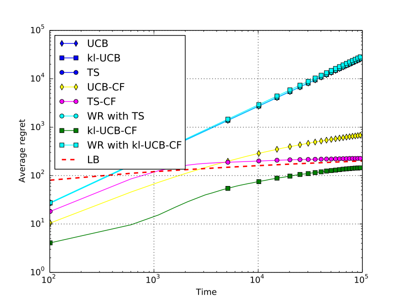


Figure 2: Regret plots with varying T up to 10^5

Experiments with varying Local DP

- Bernoulli corrupt bandit: $\mu_1 = 0.9 \qquad \mu_2 = \cdots = \mu_{10} = 0.6$
- Comparison with varying level of Local DP; ϵ from $\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$

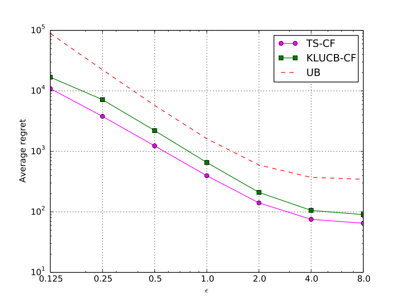


Figure 3: Regret with varying level of Local DP

Final Remarks

Final Remarks

Covered in this talk:

- Introduced Corrupt Bandits to provide privacy.
- Proved the lower bound. Provided optimal algorithms matching the lower bound.

Final Remarks

Covered in this talk:

- Introduced Corrupt Bandits to provide privacy.
- Proved the lower bound. Provided optimal algorithms matching the lower bound.

Not covered in this talk:

- Provided optimal mechanism for achieving local DP.

Final Remarks

Covered in this talk:

- Introduced Corrupt Bandits to provide privacy.
- Proved the lower bound. Provided optimal algorithms matching the lower bound.

Not covered in this talk:

- Provided optimal mechanism for achieving local DP.
- Proved regret guarantees for achieving required level of local DP (Trade-off between utility and privacy).

Final Remarks

Covered in this talk:

- Introduced Corrupt Bandits to provide privacy.
- Proved the lower bound. Provided optimal algorithms matching the lower bound.

Not covered in this talk:

- Provided optimal mechanism for achieving local DP.
- Proved regret guarantees for achieving required level of local DP (Trade-off between utility and privacy).

Future work:

- Contextual corruption?

Final Remarks

Covered in this talk:

- Introduced Corrupt Bandits to provide privacy.
- Proved the lower bound. Provided optimal algorithms matching the lower bound.

Not covered in this talk:

- Provided optimal mechanism for achieving local DP.
- Proved regret guarantees for achieving required level of local DP (Trade-off between utility and privacy).

Future work:

- Contextual corruption?
- Corrupted feedback in RL? (a recent publication by Everitt et al. (2017) [3]).

Thank you all.

References

[1] Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3):1516–1541, 2013.

[2] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Privacy aware learning. *J. ACM*, 61(6):38:1–38:57, December 2014.

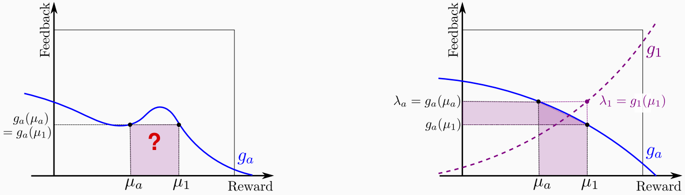
[3] Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. Reinforcement learning with a corrupted reward channel. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4705–4713, 2017.

[4] Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt bandits for preserving local privacy. In *Proceedings of the 29th International Conference on Algorithmic Learning Theory (ALT)*, 2018.

[5] 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.

Interpretation of Lower Bound for Corrupt Bandits

- Divergence between λ_a and $g_a(\mu_1)$ plays a crucial role in distinguishing arm a from the optimal arm.



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

Figure 4: On the left, g_a is such that $\lambda_a = g_a(\mu_1)$. On the right, a steep monotonic g_a leads $\Delta_a = \mu_1 - \mu_a$ into a clear gap between λ_a and $g_a(\mu_1)$.

- If the g_a function is non-monotonic, it might be impossible to distinguish between arm a and the optimal arm.
- Assumption: Corruption functions strictly monotonic.

Optimal mechanism for local DP and regret

- Corruption matrix

$$\mathbb{M}_a = \begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \left[\begin{array}{cc} \frac{e^\epsilon}{1+e^\epsilon} & \frac{1}{1+e^\epsilon} \\ \frac{1}{1+e^\epsilon} & \frac{e^\epsilon}{1+e^\epsilon} \end{array} \right] \end{array}.$$

Corollary

The regret of kl*-UCB-CF or TS-CF at time T with* ϵ*-locally differentially private bandit feedback corruption scheme is*

$$\mathrm{Regret}_T \leq \sum_{a=2}^{K} \frac{2\log(T)}{\Delta_a \left(\frac{e^\epsilon - 1}{e^\epsilon + 1}\right)^2} + O(\sqrt{\log(T)}).$$

Local DP vs global DP

- For low values of ϵ, $\left(\frac{e^{\epsilon}-1}{e^{\epsilon}+1}\right) \approx \epsilon/2$.
- In-line with global DP algorithms with a multiplicative factor of $O(\epsilon^{-1})$ or $O(\epsilon^{-2})$.
- One global DP algorithm with additive factor of $O(\epsilon^{-1})$. Our lower bound shows that's not possible for local DP.