Corrupt Bandits

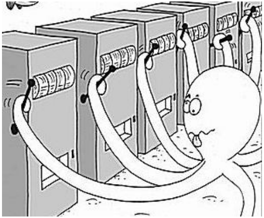
Pratik Gajane

INRIA SequeL, Université Lille 3 & Orange labs

Jan 15, 2017

Joint work with Tanguy Urvoy and Emilie Kaufmann

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

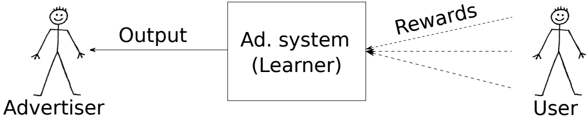


Figure 1: Ad system using bandits

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

Local Differential Privacy

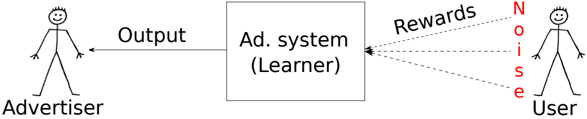


Figure 1: Ad system using bandits

- 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) [3].
- 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

Theorem (Thm. 1, PG, Urvoy & Kaufmann(2016) [1])

Any algorithm for a Bernoulli corrupt bandit problem satisfies,

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

$$d(x,y) := \mathrm{KL}(\mathcal{B}(x), \mathcal{B}(y)) = x \cdot \log\left(\frac{x}{y}\right) + (1-x)\cdot\log\left(\frac{1-x}{1-y}\right)$$

- 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 algorithm: kl-UCB-CF

Algorithm: kl-UCB-CF

Pull at time t an arm maximizing

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

- Similar to kl-UCB by Cappé et al. (2013) [2] 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(2016) [1])

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

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

Proof outline for kl-UCB-CF regret

- $\mathrm{Index}_a(t) := \max\left\{q : \ N_a(t) \cdot d(\hat{\lambda}_a(t), g_a(q)) \leq f(t)\right\}$
  or on $g_a(\mu_a)$ if g_a is decreasing
  Upper bound $u_a(t)$ on $g_a(\mu_a)$ if g_a is increasing

Proof outline for kl-UCB-CF regret

- $\mathrm{Index}_a(t) := \max\left\{q : \; N_a(t) \cdot d(\hat{\lambda}_a(t), g_a(q)) \leq f(t)\right\}$
  or on $g_a(\mu_a)$ if g_a is decreasing
  Upper bound $u_a(t)$ on $g_a(\mu_a)$ if g_a is increasing
- a is pulled at time $t+1$ by kl-UCB-CF $\Longrightarrow$
  - $g_1(\mu_1) < \ell_1(t)$ or $g_1(\mu_1) > u_1(t)$. **Unlikely event**.
  - $g_1(\mu_1)$ is inside its confidence interval. **Likely event**.

Proof outline for kl-UCB-CF regret

- $\mathrm{Index}_a(t) := \max \left\{ q : \ N_a(t) \cdot d(\hat{\lambda}_a(t), g_a(q)) \leq f(t) \right\}$
  or on $g_a(\mu_a)$ if g_a is decreasing
  Upper bound $u_a(t)$ on $g_a(\mu_a)$ if g_a is increasing

- a is pulled at time $t+1$ by kl-UCB-CF $\Longrightarrow$
  - $g_1(\mu_1) < \ell_1(t)$ or $g_1(\mu_1) > u_1(t)$. **Unlikely event**.
  - $g_1(\mu_1)$ is inside its confidence interval. **Likely event**.

- Probability of **unlikely event** $= o(\log T)$.

Proof outline for kl-UCB-CF regret

- $\mathrm{Index}_a(t) := \max\left\{q : \; N_a(t) \cdot d(\hat{\lambda}_a(t), g_a(q)) \leq f(t)\right\}$
  or on $g_a(\mu_a)$ if g_a is decreasing
  Upper bound $u_a(t)$ on $g_a(\mu_a)$ if g_a is increasing
- a is pulled at time $t+1$ by kl-UCB-CF $\Longrightarrow$
  - $g_1(\mu_1) < \ell_1(t)$ or $g_1(\mu_1) > u_1(t)$. **Unlikely event**.
  - $g_1(\mu_1)$ is inside its confidence interval. **Likely event**.
- Probability of **unlikely event** $= o(\log T)$.
- Probability of **likely event** $= \frac{\log T}{d(\lambda_a, g_a(\mu_1))} + \cdots$

Proof outline for kl-UCB-CF regret

- $\mathrm{Index}_a(t) := \max\left\{q : \; N_a(t) \cdot d(\hat{\lambda}_a(t), g_a(q)) \leq f(t)\right\}$
  or on $g_a(\mu_a)$ if g_a is decreasing
  Upper bound $u_a(t)$ on $g_a(\mu_a)$ if g_a is increasing

- a is pulled at time $t+1$ by kl-UCB-CF $\implies$
  - $g_1(\mu_1) < \ell_1(t)$ or $g_1(\mu_1) > u_1(t)$. **Unlikely event**.
  - $g_1(\mu_1)$ is inside its confidence interval. **Likely event**.

- Probability of **unlikely event** $= o(\log T)$.

- Probability of **likely event** $= \frac{\log T}{d(\lambda_a, g_a(\mu_1))} + \cdots$

- Above leads to upper bound on $\mathbb{E}[N_a(T)]$ and
  $\text{Regret}_T = \sum_{a=2}^{K} \Delta_a \cdot \mathbb{E}[N_a(T)]$.

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

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

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

Proof outline for $\mathrm{TS\text{-}CF}$ regret

- Two thresholds u_a and w_a
  $\lambda_a < u_a < w_a < g_a(\mu_1)$ if g_a is increasing and,
  $\lambda_a > u_a > w_a > g_a(\mu_1)$ if g_a is decreasing.

Proof outline for $\mathrm{TS\text{-}CF}$ regret

- Two thresholds u_a and w_a
  $\lambda_a < u_a < w_a < g_a(\mu_1)$ if g_a is increasing and,
  $\lambda_a > u_a > w_a > g_a(\mu_1)$ if g_a is decreasing.
- Event $E_a^{\lambda}(t) = \{g_a^{-1}(\hat{\lambda}_a(t)) \leq g_a^{-1}(u_a)\}$
  Event $E_a^{\theta}(t) = \{g_a^{-1}(\theta_a(t)) \leq g_a^{-1}(w_a)\}$

Proof outline for TS-CF regret

- Two thresholds u_a and w_a
  $\lambda_a < u_a < w_a < g_a(\mu_1)$ if g_a is increasing and,
  $\lambda_a > u_a > w_a > g_a(\mu_1)$ if g_a is decreasing.
- Event $E_a^\lambda(t) = \{g_a^{-1}(\hat{\lambda}_a(t)) \leq g_a^{-1}(u_a)\}$
  Event $E_a^\theta(t) = \{g_a^{-1}(\theta_a(t)) \leq g_a^{-1}(w_a)\}$
- $$\begin{aligned}\mathbb{E}[N_a(T)] \leq & \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), \overline{E_a^\theta(t)}) \\ & + \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), E_a^\theta(t)) \\ & + \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \overline{E_a^\lambda(t)}).\end{aligned}$$

Proof outline for TS-CF regret

- Two thresholds u_a and w_a
  $\lambda_a < u_a < w_a < g_a(\mu_1)$ if g_a is increasing and,
  $\lambda_a > u_a > w_a > g_a(\mu_1)$ if g_a is decreasing.
- Event $E_a^\lambda(t) = \{g_a^{-1}(\hat{\lambda}_a(t)) \leq g_a^{-1}(u_a)\}$
  Event $E_a^\theta(t) = \{g_a^{-1}(\theta_a(t)) \leq g_a^{-1}(w_a)\}$
- $$\begin{aligned}\mathbb{E}[N_a(T)] \leq & \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), \overline{E_a^\theta(t)}) \\ & + \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \overline{E_a^\lambda(t)}, E_a^\theta(t)) \\ & + \sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \overline{E_a^\lambda(t)}).\end{aligned}$$
- Last two terms are $o(\log(T))$.

Proof outline for TS-CF regret

- Two thresholds u_a and w_a
  $\lambda_a < u_a < w_a < g_a(\mu_1)$ if g_a is increasing and,
  $\lambda_a > u_a > w_a > g_a(\mu_1)$ if g_a is decreasing.
- Event $E_a^{\lambda}(t) = \{g_a^{-1}(\hat{\lambda}_a(t)) \leq g_a^{-1}(u_a)\}$
  Event $E_a^{\theta}(t) = \{g_a^{-1}(\theta_a(t)) \leq g_a^{-1}(w_a)\}$
- $$\begin{aligned}\mathbb{E}[N_a(T)] \leq & \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^{\lambda}(t), \overline{E_a^{\theta}(t)}) \\ & + \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^{\lambda}(t), E_a^{\theta}(t)) \\ & + \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \overline{E_a^{\lambda}(t)}).\end{aligned}$$
- Last two terms are $o(\log(T))$.
- First term is $\leq \frac{\log(T)}{d(u'_a, w_a)} + 1$ for large T and suitable u'_a.

Proof outline for TS-CF regret

- Two thresholds u_a and w_a
  $\lambda_a < u_a < w_a < g_a(\mu_1)$ if g_a is increasing and,
  $\lambda_a > u_a > w_a > g_a(\mu_1)$ if g_a is decreasing.
- Event $E_a^\lambda(t) = \{g_a^{-1}(\hat{\lambda}_a(t)) \leq g_a^{-1}(u_a)\}$
  Event $E_a^\theta(t) = \{g_a^{-1}(\theta_a(t)) \leq g_a^{-1}(w_a)\}$
- $$\begin{aligned}\mathbb{E}[N_a(T)] \leq & \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), \overline{E_a^\theta(t)}) \\ & + \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, E_a^\lambda(t), E_a^\theta(t)) \\ & + \textstyle\sum_{t=0}^{T-1} \mathbb{P}(\hat{a}_{t+1} = a, \overline{E_a^\lambda(t)}).\end{aligned}$$
- Last two terms are $o(\log(T))$.
- First term is $\leq \frac{\log(T)}{d(u_a', w_a)} + 1$ for large T and suitable u_a'.
- Binding above leads to upper bound on $\mathbb{E}[N_a(T)]$ and $\text{Regret}_T = \sum_{a=2}^{K} \Delta_a \cdot \mathbb{E}[N_a(T)]$.

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

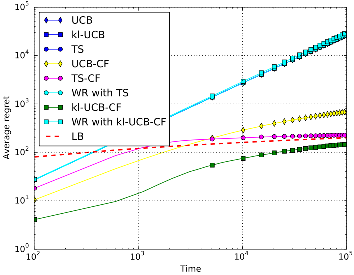


Figure 2: Regret plots with varying T up to 10^5

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 $\{1/8, 1/4, 1/2, 1, 2, 4, 8\}$

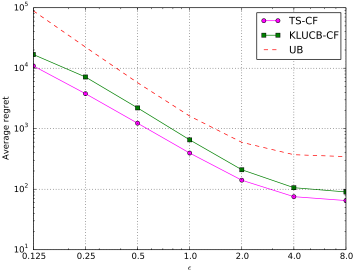


Figure 3: Regret with varying level of Local DP

Final Remarks

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).
- Provided lower bound on sample complexity for best arm identification and two corresponding algorithms.

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).
- Provided lower bound on sample complexity for best arm identification and two corresponding algorithms.

Future work:

- Contextual corruption?
- Corrupted feedback in RL? (a very recent arXiv article by Everitt et al. (2017) [4]).

Thank you all.

References I

[1] Corrupt bandits. *The European Workshop in Reinforcement Learning (EWRL)*, 2016.

[2] 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.

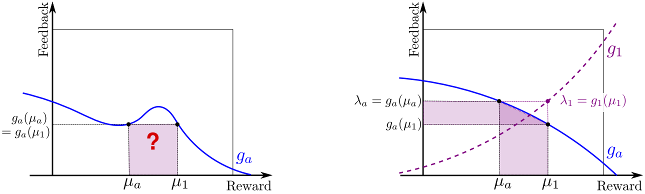
[3] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Privacy aware learning. *J. ACM*, 61(6):38:1–38:57, December 2014.

[4] Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel. *CoRR*, abs/1705.08417, 2017.

[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.