Multi-agent Perspective of Fake Feedback Attacks on Stochastic Multi-armed Bandits

Charles A. N. Costa, Célia Ghedini Ralha

Computer Science Department
University of Brasília

Agenda

- Introduction.
- Contributions.
- Adversarial vs Fake feedback.
- Agent roles.
- Fake feedback Attacks.
- Experiments.
- Conclusions.

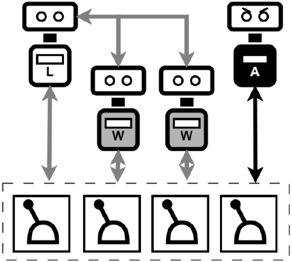
Introduction

- Multi-armed bandits (MAB) – To balance exploration and exploitation.
- Well-known stochastic MAB: ϵ-Greedy and UCB1.
- Stochastic MAB are vulnerable to data poisoning attacks.
- Many studies focus only on adversarial attacks when an attacker controls the reward delivery mechanism (generality).
- Just a few approaches to this problem as a Multi-agent problem, although roles, goals, intentions, behavior, and capacities emerge from definitions.

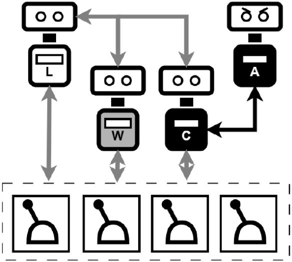
Contributions

- **Main contribution:** Our main contribution is the analysis of attacks on stochastic MAB from a multi-agent perspective.
- Secondary contributions:
  - Describe four fake-feedback attacks using our framework.
  - Present data from synthetic experiments.

Adversarial vs Fake feedback



(a) Adversarial attack.



(b) Fake feedback attack.

Agent roles

- Learner
  - Goal: Collect reward from arms (options).
  - Applies a MAB policy $\pi(B_l, t)$.
- Attacker
  - Goal: Manipulate the learner to increase target arm pulls.
  - Applies attack policy $\rho(B_A, k_T, t)$.
- Witnesses
  - Goal: Collect reward from arms (options).
- Co-opted witnesses
  - Goal: Help the attacker manipulate the learner.
  - Follow attacker instructions to corrupt reward reports.

Attacks – Constant and Adaptive Attack

- Constant Attack
  - Idea: Attack every arm but the target arm with a constant C.
  - Advantages
    - Straightforward.
    - Fixed cost.
  - Disadvantages: Need to fix C in advance.
- Adaptive Attack
  - Idea: Adjust corruption level to keep target arm pulls between a range.
  - Advantages
    - Simple.
    - Tends to be less costly than the Constant attack.
  - Disadvantage: More parameters than the Constant attack.

Jun's Adversarial relaxed attacks

- Jun et al. (2018). Adversarial attacks on stochastic bandits. In *Advances in Neural Information Processing Systems*.
- Original idea: Carefully craft the corruption level to minimize cost and maximize manipulation.
- Not agnostic. Defined to ϵ-Greedy and UCB1.
- Need to know in advance: next pulled arm, next pulled arm reward value, the learner's policy.
- Relaxations
  - Use an unbiased estimator, like sample mean, instead of the next reward value.
  - Attack all arms but the target arm!
- Incurs a higher cost than the no-relaxed version (weaker!).

Experiments - Set up

- MAB: UCB1 and ϵ-Greedy.
- Attacks: Constant, two set-ups adaptive, Jun's adversarials relaxed.
  - Constant: $C = 1$.
  - Adaptive 1, ranging $(0.4, 0.6)$.
  - Adaptive 2, ranging $(0.8, 0.9)$.
  - Two Jun's attacks.
- Baseline: No attack.
- Execution:
  - 5 arms from three reward classes:
    - A1: $\mathcal{N}(0.9, 0.1)$.
    - B1 and B2: $\mathcal{N}(0.85, 0.30)$.
    - C1 and C2: $\mathcal{N}(0.75, 0.50)$.
  - Target arm: C2.
  - Witnesses: 9 (10 players counting the learner).
  - Co-opted witnesses: 5.
  - 2000 rounds.
  - 30 repetition.
- Source code: github.com/charlesANC/BanditsExperiment

Experiments - Performance measures

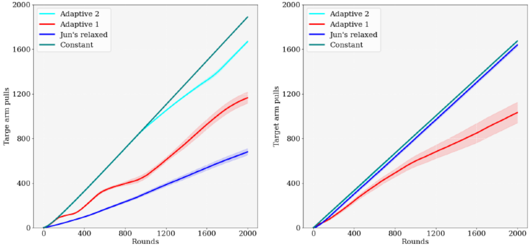
1. Regret – $R_L(T)$
   - Estimate the maximum reward the Learner could achieve and subtract the actual accumulated reward.
2. Total corruption level – $C(T)$
   - Sum all the corruption in co-opted witnesses' reports.
3. Achieved Pulls – $AP(T)$
   - Increase in the target arm pulls when compared to a zero-corruption scenario.
4. Cost per Achieved pull – $CP(T)$.
   - Divide Total corruption level by Achieved pulls.

Experiments - Outcomes

Table 2. Resumed measures over MAB algorithms and attacks. The values represent the mean with the standard variation in parentheses.

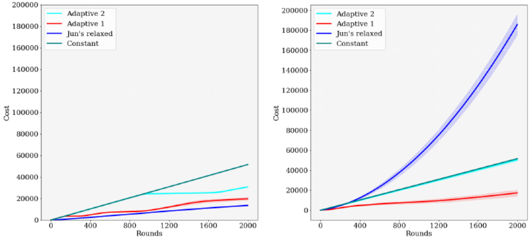
MAB	Attack	$R_L(T)$	$N(k_t, T, C(T))$	$AP(T)$	$C(T)$	$CP(T)$
UCB1	-	70.93 (8.77)	131.50 (9.70)	-	-	-
UCB1	Constant	297.20 (21.42)	1,889.07 (1.69)	1,757.57 (10.03)	51,549.73 (68.40)	29.33 (0.18)
UCB1	Adaptive 1	216.34 (24.91)	1,164.87 (134.40)	1,033.37 (135.04)	19,769.67 (3,620.93)	19.11 (2.51)
UCB1	Adaptive 2	275.42 (24.39)	1,668.43 (24.99)	1,536.93 (27.41)	30,803.83 (1,565.36)	20.04 (0.89)
UCB1	Jun's relaxed	181.52 (19.12)	679.83 (77.59)	548.33 (78.79)	13,601.23 (1,271.82)	25.04 (2.18)
ϵ-Greedy	-	31.45 (8.96)	79.67 (8.83)	-	-	-
ϵ-Greedy	Constant	274.45 (23.26)	1,673.80 (13.08)	1,594.13 (15.41)	51,581.90 (68.72)	33.45 (0.30)
ϵ-Greedy	Adaptive 1	180.72 (38.44)	1,032.57 (246.09)	952.90 (244.51)	17,421.66 (8,252.80)	20.12 (9.58)
ϵ-Greedy	Adaptive 2	269.08 (21.88)	1,594.90 (18.22)	1,674.57 (18.74)	50,468.31 (3,910.64)	31.65 (2.48)
ϵ-Greedy	Jun's relaxed	268.87 (20.86)	1,638.10 (37.38)	1,594.90 (38.61)	185,662.13 (26,423.06)	123.55 (19.79)

Experiments - Outcomes



(a) UCB1 - Target arm pulls. (b) ϵ-Greedy - Target arm pulls.

Figure 2. Target arm pulls over MAB algorithms and attacks.



(a) UCB1 - Total Cost. (b) ϵ-Greedy - Total Cost.

Figure 3. Cost of corruption over MAB algorithms and attacks.

Conclusions

- This paper emphasized understanding the problem of fake feedback attacks on stochastic MAB within a MAS framework.
- Our findings suggest that agnostic attacks could be effective against UCB1 and e-Greedy, even compared to policy-based attacks.
- Future work should focus on developing effective defenses against fake feedback attacks that consider the MAS perspective.

Thank you!

charles.costa@aluno.unb.br
ghedini@unb.br